SLAC – PUB – 4389(REV)
July 1988
(M)

Regularized Discriminant Analysis*

JEROME H. FRIEDMAN
Department of Statistics
and
Stanford Linear Accelerator Center
Stanford University, Stanford CA 94309

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. These studies indicate that in many circumstances dramatic gains in classification accuracy can be achieved.

Submitted to *Journal of the American Statistical Association*

---

*Work supported by the Department of Energy, contract DE–AC03–76SF00515.

1.0 Classification

The formal purpose of classification or discriminant analysis is to assign objects to one of several (K) groups or classes based on a set of measurements $\underline{X} = (X_1, X_2, \ldots, X_p)$ obtained from each object or observation. Classification techniques are also used informally to study the separability of labeled groups of observations in the measurement space. In the formal setting, 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

$$L(k, \hat{k}) \qquad 1 \leq k, \hat{k} \leq K \quad , \tag{1}$$

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 k (population) are seldom identical but comprise a distribution of values characterized by a probability density $f_k(\underline{X})$. The usual goal is 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 densities $f_k(\underline{X})$ 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 $\underline{X}$ as $\hat{k}$ is

$$R(\hat{k}|\underline{X}) \;=\; \frac{\sum_{k=1}^{K} L(k, \hat{k})\; f_k(\underline{X})\; \pi_k}{\sum_{k=1}^{K} f_k(\underline{X})\; \pi_k} \quad , \tag{2}$$

where π_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. (2). For the special but commonly occurring case

$$L(k, \hat{k}) \;=\; 1 - \delta(k, \hat{k}) \quad , \tag{3}$$

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

$$f_{\hat{k}}(\underline{X})\; \pi_{\hat{k}} \;= \max_{1 \leq k \leq K} f_k(\underline{X})\; \pi_k \quad . \tag{4}$$

The loss matrix [Eq. (3)] 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 rule resulting from choosing $\hat{k}$ to minimize $R(\hat{k}|\underline{X})$ [Eq. (2) or (4)] is known as the Bayes rule and it achieves minimal misclassification risk among all possible rules.

The class conditional densities $f_k(\underline{X})$ 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 by obtaining suitable estimates of the $f_k(\underline{X})$. Since these estimates generally deviate from the true population densities, 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

$$\hat{\pi}_k = \frac{W_k}{W} \quad , \tag{5}$$

with

$$W_k = \sum_{c(v)=k} w_v \quad , \tag{6a}$$

and

$$W = \sum_{k=1}^{K} W_k \quad . \tag{6b}$$

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.

2.0 Linear and Quadratic Discriminant Analysis

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

$$f_k(\underline{X}) = (2\pi)^{-p/2} \ |\Sigma_k|^{-1/2} \ e^{-1/2(\underline{X}-\underline{\mu}_k)^T \Sigma_k^{-1} (\underline{X}-\underline{\mu}_k)} \quad , \tag{7}$$

where $\underline{\mu}_k$ and Σ_k are the class k $(1 \leq k \leq K)$ population mean vector and covariance matrix. Assuming the simple loss structure [Eq. (3)] and substituting Eq. (7) into Eq. (4) leads to the classification rule

$$d_{\hat{k}}(\underline{X}) = \min_{1 \leq k \leq K} d_k(\underline{X}) \quad , \tag{8}$$

with

$$d_k(\underline{X}) = (\underline{X} - \underline{\mu}_k)^T \Sigma_k^{-1} (\underline{X} - \underline{\mu}_k) + \ell n \, |\Sigma_k| - 2\ell n \, \pi_k \quad . \tag{9}$$

This quantity is often called the discriminant score for the k^{th} class, whereas $d_k(\underline{X}) + 2\ell n \, \pi_k$ is referred to as the discriminant function. The first term on the right-hand side of Eq. (9) is the well-known Mahalonobis distance between $\underline{X}$ and $\underline{\mu}_k$.

Using the classification rule [Eqs. (8) and (9)] 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

$$\Sigma_k = \Sigma \qquad 1 \leq k \leq K \quad . \tag{10}$$

This is referred to as linear discriminant analysis (LDA) because the quadratic terms associated with Eqs. (8) and (9) 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 $\underline{\mu}_k$ and covariance matrices Σ_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)]. Classification rules based on QDA are known to require generally larger samples than those based on LDA [Wahl and Kronmal (1977)] and seem to be more sensitive to violations of the basic assumptions.

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

$$\hat{\underline{\mu}}_k = \overline{\underline{X}}_k = \frac{1}{W_k} \sum_{c(v)=k} w_v \, \underline{X}_v \quad , \tag{11}$$

and

$$\hat{\Sigma}_k = \frac{S_k}{W_k} = \frac{1}{W_k} \sum_{c(v)=k} w_v (\underline{X}_v - \overline{\underline{X}}_k)(\underline{X}_v - \overline{\underline{X}}_k)^T \quad , \tag{12}$$

with W_k given by Eq. (6a). These so-called *"plug-in"* estimates are straightforward to compute and represent the corresponding maximum likelihood estimates. (Often the covariance matrix estimates are scaled by a factor to remove bias.) Although seemingly reasonable, this approach can be justified only on intuitive grounds, and it enjoys no optimality properties (except asymptotically) even when the population distributions are normal [Anderson (1958)]. Also, any sensible Bayesian rule will not lead to this approach, except either asymptotically or under very restrictive conditions [Enis and Geisser (1974)].

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 decompositions

$$\Sigma_k = \sum_{i=1}^{p} e_{ik} \, \underline{v}_{ik} \, \underline{v}_{ik}^T \quad ,$$

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

$$\Sigma_k^{-1} = \sum_{i=1}^{p} \frac{\underline{v}_{ik}\, \underline{v}_{ik}^T}{e_{ik}} \quad ,$$

and the discriminant score [Eq. (9)] becomes

$$\begin{aligned} d_k(\underline{X}) &= \sum_{i=1}^{p} \frac{[\underline{v}_{ik}^T(\underline{X} - \underline{\mu}_k)]^2}{e_{ik}} \\ &+ \sum_{i=1}^{p} \ell n\, e_{ik} - 2\ell n\, \pi_k \quad . \end{aligned} \tag{13}$$

The discriminant score [Eq. (13)] 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 $\widehat{\Sigma}_k$ [Eq. (12)].

It is well known that the estimates based on Eq. (12) 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 perhaps to orthogonality constraints.

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. Therefore, most of the variance incurred in estimating the discriminant scores [Eqs. (9) and (13)] is associated with directions of low sample variance in the measurement space.

3.0 Regularization and Shrinkage

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 Srmivasan (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 $\widehat{\Sigma}_k$ 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." [Cornfield (1967) suggested applying James–Stein shrinkage to the individual class location estimates.] Regularization reduces the variance associated with the sample based estimate 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.

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

$$\widehat{\Sigma}_k = \widehat{\Sigma} = \frac{S}{W} \quad , \tag{14}$$

where

$$S = \sum_{k=1}^{K} S_k \quad , \tag{15}$$

with W given by Eq. (6b) and S_k by Eq. (12). 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

$$\widehat{\Sigma}_k(\lambda) = \frac{S_k(\lambda)}{W_k(\lambda)} \quad , \tag{16a}$$

where

$$S_k(\lambda) = (1-\lambda)\, S_k + \lambda S \quad , \tag{16b}$$

and

$$W_k(\lambda) = (1-\lambda)\, W_k + \lambda W \quad , \tag{16c}$$

with S_k given by Eq. (12), S by Eq. (15), and W_k and W by Eq. (6). The regularization parameter λ 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 λ (than $\lambda = 1$) have the potential of superior performance when the population class covariance matrices substantially differ.

The regularization provided by Eqs. (16) 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

$$N = \sum_{k=1}^{K} N_k \tag{17}$$

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 [trace $(\widehat{\Sigma}_k)/p$] 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. (16) through

$$\widehat{\Sigma}_k(\lambda,\gamma) \;=\; (1-\gamma)\;\widehat{\Sigma}_k(\lambda) + \frac{\gamma}{p}\;\mathrm{trace}\left[\widehat{\Sigma}_k(\lambda)\right]\;I \quad , \tag{18}$$

with $\widehat{\Sigma}_k(\lambda)$ given by Eqs. (16) and I being the identity matrix. For a given value of λ, the additional regularization parameter γ, $0 \leq \gamma \leq 1$, controls shrinkage toward a multiple of the identity matrix. The multiplier is just the average eigenvalue of $\widehat{\Sigma}_k(\lambda)$. 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 (16) and (18) represent a two-parameter family of regularized sample class covariance matrix estimators, to be used with the class discriminant scores

$$\begin{aligned} d_k(\underline{X}) &= (\underline{X} - \overline{\underline{X}}_k)^T \, \hat{\Sigma}_k^{-1}(\lambda,\gamma)(\underline{X} - \overline{\underline{X}}_k) \\ &+ \ell n \left| \hat{\Sigma}_k(\lambda,\gamma) \right| - 2\ell n \, \pi_k \quad , \end{aligned} \tag{19}$$

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 4.0). We refer to this approach as *"regularized discriminant analysis"* (RDA).

Regularized discriminant analysis provides for a fairly rich class of regularization alternatives. The four corners defining the extremes of the λ, γ 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 γ fixed at zero and varying λ produces models in between QDA and LDA. Holding λ fixed at zero and increasing γ attempts to unbias the sample based eigenvalue estimates. Holding λ fixed at one and increasing γ gives rise to a ridge regression analogue for LDA.

4.0 Model Selection

A good pair of values for λ and γ 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. For classification, two sample resuse methods, cross-validation [Lachenbruch (1975) and Geisser (1977)] and Bootstrapping [Efron (1983)] have been suggested. The computational advantages associated with the cross-validation approach in this particular application (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 $\underline{X}_v$ by removing it from the model building process. That is, the classification rule is developed on the $N-1$ training observations exclusive of $\underline{X}_v$ and then it is used to classify $\underline{X}_v$. 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.

Our approach to model selection is to choose values of the covariance matrix mixing parameter λ, and the eigenvalue shrinkage parameter γ, 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 λ, γ 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, $\widehat{\lambda}$ and $\widehat{\gamma}$. 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. (17)] sets of discriminant scores [Eq. (19)] 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. (19)] with the observation to be classified (say $\underline{X}_v$) left out

$$\begin{aligned} d_{k\backslash v}(\underline{X}_v) &= (\underline{X}_v - \overline{\underline{X}}_{k\backslash v})^T \, \widehat{\Sigma}^{-1}_{k\backslash v}(\lambda,\gamma)(\underline{X}_v - \overline{\underline{X}}_{k\backslash v}) \\ &+ \ell n \left| \widehat{\Sigma}_{k\backslash v}(\lambda,\gamma) \right| - 2\ell n\, \pi_k \quad . \end{aligned} \tag{20}$$

Here the notation $\backslash v$ 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 $\underline{X}_v$. 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 $d_{k\backslash v}$ [Eq. (20)] from d_k [Eq. (19)] [see Golub and Van Loan (1983)].

Unfortunately, removing an observation does not result in a rank-one downdate of $\widehat{\Sigma}_k(\lambda,\gamma)$ [Eqs. (16),(18)]. It can be shown that

$$W_{k\backslash v}(\lambda)\ \widehat{\Sigma}_{k\backslash v}(\lambda,\gamma) \;=\; W_k(\lambda)\ \widehat{\Sigma}_k(\lambda,\gamma) - (1-\gamma)\ \underline{Z}_v\underline{Z}_v^T - \frac{\gamma}{p}\ |\underline{Z}_v|^2\ I \quad , \tag{21a}$$

with $W_k(\lambda)$ given by [Eq. (16c)], and

$$W_{k\backslash v}(\lambda) \;=\; W_k(\lambda) - s_k(v)\ w_v \quad , \tag{21b}$$

$$s_k(v) \;=\; \begin{cases} 1 & \text{if } c(v)=k \\ \lambda & \text{otherwise,} \end{cases} \tag{21c}$$

$$\underline{Z}_v \;=\; \sqrt{b_k\,(v)}\ \left(\underline{X}_v - \overline{\underline{X}}_{c(v)}\right) \quad , \tag{21d}$$

and

$$b_k(v) \;=\; \frac{s_k(v)\,W_{c(v)}\ w_v}{W_{c(v)} - w_v} \quad . \tag{21e}$$

Thus, removing an observation is equivalent to downdating $\widehat{\Sigma}_k(\lambda,\gamma)$ 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:

$$W_k^{-1}(\lambda)\ \widehat{\Sigma}_k^{-1}(\lambda,\gamma) \;=\; \sum_{i=1}^{p} \frac{\underline{v}_i\,\underline{v}_i^T}{e_i} \quad . \tag{22a}$$

Then

$$\left[W_k(\lambda)\ \widehat{\Sigma}_k(\lambda,\gamma) - aI\right]^{-1} \;=\; \sum_{i=1}^{p} \frac{\underline{v}_i\,\underline{v}_i^T}{(e_i-a)} \quad , \tag{22b}$$

where e_i is the i^{th} eigenvalue of $W_k(\lambda)\ \widehat{\Sigma}_k(\lambda,\gamma)$, $\underline{v}_i$ 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)]:

$$\left(A - \underline{r}\underline{r}^T\right)^{-1} = A^{-1} + \frac{A^{-1}\,\underline{r}\underline{r}^T\,A^{-1}}{1 - \underline{r}^T\,A^{-1}\,\underline{r}} \quad , \tag{23}$$

where A is a nonsingular matrix and $\underline{r}$ is a vector. In our case A^{-1} is given by [Eq. (22)] with

$$a = \frac{\gamma\,|\underline{Z}_v|^2}{p} \quad , \tag{24a}$$

and

$$\underline{r} = \sqrt{1-\gamma}\;\underline{Z}_v \quad , \tag{24b}$$

with $\underline{Z}_v$ given by Eq. (21).

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. (20)]. It is easily verified that

$$\overline{\underline{X}}_{k\backslash v} = \begin{cases} \overline{\underline{X}}_k & \text{if } c(v) \neq k \\ \dfrac{W_k\overline{\underline{X}}_k - w_v\underline{X}_v}{W_k - w_v} & \text{otherwise} \end{cases} \quad , \tag{25}$$

and

$$\ell n\left|W_{k\backslash v}(\lambda)\;\widehat{\Sigma}_{k\backslash v}(\lambda,\gamma)\right| = \sum_{i=1}^{p}\ell n(e_i - a) + \ell n\left[1 - \sum_{i=1}^{p}\frac{r_i^2}{e_i - a}\right] \tag{26}$$

with e_i given by Eq. (22) and a and $\underline{r}$ given by Eq. (24).

These quantities [Eqs. (21)–(26)] can be substituted into Eq. (20) 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 λ and γ.

A substantial amount of additional computation can be saved by taking advantage of the fact that for a fixed value of λ the eigenvectors $\underline{v}_i$ [Eq. (22a)] are independent of γ. Changing γ is equivalent to an update by a multiple of the identity matrix. Thus, the K spectral decompositions and the corresponding rotations $\underline{v}_{ik}^T(\underline{X}_v - \overline{\underline{X}}_k)$ $(1 \leq i \leq p,\ 1 \leq k \leq K,\ 1 \leq v \leq N)$ need only be recalculated when the value of λ changes. For each distinct value of λ on the optimization grid, the set of points corresponding to different values of γ can each be cross-validated in time proportional to pN. Therefore, the grid points should be visited in an order that causes λ to change as few times as possible.

5.0 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 γ. 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 counterproductive. 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 regulariztion 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.

6.0 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. (7)] and the total sample size was $N = 40$ [Eq. (17)]. 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 and 40. In all cases there were $K = 3$ groups or classes. The optimization grid of (λ, γ) 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 $N = 40$ class identity labels were randomly drawn. Then, conditioned on each label, measurement vectors were drawn from the appropriate class distribution. 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. (3).

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.

6.1 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 over the 100 replications. The standard deviations of the corresponding averages are one tenth these amounts.

As suspected, RDA gives uniformally 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 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 λ and γ on average.

Table 1. Equal spherical covariance matrices.

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

6.2 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. As conjectured, RDA strongly dominates with smaller risk at all dimensionalities, 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 λ, and very large values for the eigenvalue shrinkage parameter γ.

Table 2. Unequal spherical covariance matrices.

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

6.3 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 subspace, whereas in the second they

are concentrated in the high variance subspace. The eigenvalues of the common population covariance matrices are given by

$$e_i = \left[\frac{9(i-1)}{p-1} + 1\right]^2 \quad , \qquad 1 \leq i \leq p \quad , \tag{27}$$

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

$$\mu_{2i} = 2.5\sqrt{\frac{e_i}{p}}\,\frac{p-i}{\frac{p}{2}-1} \quad ,$$

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

with e_i given by Eq. (27). 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 λ, γ 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. (27)] is used for each class population, but the mean differences are concentrated in the high variance subspace. This provides the shrinkage strategy with at least a chance at accomplishing some improvement. For this case the class two and class three means are given by

Table 3. Equal, highly ellipsodial covariance matrices — mean differences in low-variance subspace.

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

$$\mu_{2i} = 2.5\sqrt{\frac{e_i}{p}}\,\frac{i-1}{\frac{p}{2}-1} \quad ,$$

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

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 ($\simeq 30\%$).

Table 4. Equal, highly eillipsodial covariance matrices — mean differences in high variance subspace.

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

6.4 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. (27). Those for class two are given by

$$e_{i2} = \left[\frac{9(p-i)}{p-1} + 1\right]^2 \quad , \qquad 1 \leq i \leq p \quad ,$$

while those for class three are

$$e_{i3} = \left\{\frac{9[i - \frac{(p-1)}{2}]}{(p-1)}\right\}^2 \quad , \qquad 1 \leq i \leq p \quad .$$

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

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

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

Table 5. Unequal, highly ellipsoidal covariance matrices — zero mean differences.

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

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.

6.5 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 λ, and eigenvalue shrinkage parameter γ, 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 seems to loose so little in situations unfavorable to it 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 ($\simeq$ 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.

Table 6. Unequal, highly ellipsodial covariance matrices — nonzero mean differences.

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

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.

7.0 Wine Tasting Data

This data set 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.

For this example, the prior probabilities were taken to be equal, $\pi_k = 1/3$, for all classes. The optimization grid point values for λ were the same as for the simulation examples. The values for γ 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 to study the effect of regularization on misclassification risk, and not to present a complete or definitive analysis of these data.

Two studies 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 λ over the 100 bootstrap replications had a mean of $\overline{\lambda} = 0.49$ and a standard deviation of $\sigma(\lambda) = 0.37$. The corresponding quantities for the distribution of γ were $\overline{\gamma} = 0.40$ and $\sigma(\gamma) = 0.31$.

The results for RDA averaged over the two half sample runs ($N = 19$) gave $\overline{\lambda} = 0.56$, $\overline{\gamma} = 0.48$, 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.

8.0 Invariance Properties

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 γ. 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.

Changing the scales of the measurement variables or their linear combinations is equivalent to changing the regularization matrix for $\widehat{\Sigma}_k(\lambda, \gamma)$ in Eq. (18). There $\widehat{\Sigma}_k(\lambda)$ [Eq. (16)] 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

$$\widehat{\Sigma}_k(\lambda,\gamma) \;=\; (1-\gamma)\;\widehat{\Sigma}_k(\lambda) + \gamma\, t M \quad , \tag{28}$$

with M a prespecified positive definite symmetric matrix and

$$t \;=\; \frac{trace[\widehat{\Sigma}_k(\lambda)]}{trace(M)} \quad . \tag{29}$$

One can implement this generalized approach, using the techniques outlined in Sections 3 and 4, by first applying a transformation to the data (rotation and scaling) that takes M to the identity matrix. Let

$$M \;=\; L\,L^T$$

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

$$\underline{Y}_v \;=\; L^{-1}\,\underline{X}_v \quad , \qquad 1 \le v \le N \quad , \tag{30}$$

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. (30)] is equivalent to specifying M as the regularizing matrix [Eqs. (28) and (29)] in the original coordinate system.

A common procedure is to standardize or "auto-scale" the data so that all variables have the same variance. This is equivalent to using the diagonal matrix

$$M \;=\; diag\left(\widehat{\sigma}_1^2,\, \widehat{\sigma}_2^2,\, \ldots,\, \widehat{\sigma}_p^2\right) \quad ,$$

for regularization, where $\widehat{\sigma}_i$ is the sample standard deviation of the i^{th} measurement variable. A more natural choice might be to auto-scale the data using the global within class standard deviations. Another approach would be to shrink in

a way that preserves large correlations at the expense of the smaller ones [Devlin et al. (1975)]. Note that if the regularizing matrix M [Eq. (28)] depends on the data, then the matrix updating formulae for cross-validation derived in Section 4.0 are only approximate in that they do not account for the sampling variability associated with the estimate of M. 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 one might experiment with several choices using the minimized cross-validated risk estimate as a guide.

There can be situations, however, where particular regularizations are suggested. When the data measurement vectors $\underline{X}_v$ 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

$$M = H^T H \quad , \tag{31a}$$

with H being the matrix representation of some smoothing kernel

$$H_{ij} = h\left(\frac{d_{ij}}{s}\right) \quad . \tag{31b}$$

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

$$\sum_{j=1}^{p} H_{ij} = 1 \quad , \tag{31c}$$

d_{ij} is a distance between the indices i and j, and s is the bandwidth parameter for the smoothing kernal. In the case of a signal this will produce a banded regularizing matrix with large values only near the diagonal. Using M [Eq. (31)] 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.

9.0 Variable Subset Selection

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. (Note that, unlike squared-error loss, there is no fast branch-and-bound algorithm for all subset selection using misclassification risk.) 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. If this is true and if one can reliably 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)].

10.0 Concluding Remarks

The simulation studies and the data example indicate that the method of regularization applied here has the potential to (sometimes dramatically) increase 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 situations unfavorable to it, and often substantial gains in favorable circumstances. Of course, one does not generally know the type of situation in advance when confronted with a particular data set.

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 nonparametric classification techniques are generally 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

Anderson, T. W. (1958). *An Introduction to Multivariate Analysis.* Wiley, New York.

Andrews, D. F. and Hertzberg, A. M. (1985). *Data. A collection of problems from many fields for the student and research worker.* Springer–Verlag, New York.

Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. J. (1983). *Classification and Regression Trees.* Monograph, Wadsworth, Belmont, California.

Copas, J. B. (1983). *Regression, Prediction and Shrinkage (with Discussion),* J. Royal Statist. Soc. B, **45**, 311–354.

Cornfield, J. (1967). *Discriminant functions. Review of the International Statistical Institute,* **35**, 142–153.

Crawford, S. (1986). Resampling strategies for recursive partitioning classification with the CART® algorithm. Ph.D. dissertation, Department of Education, Stanford University.

Devlin, S. J., Gnanadesikan, R., Kettenring, J. R. (1975). *Robust estimation and outlier detection with correlation coefficients,* Biometrika **62**, 531–545.

Dey, D. K. and Srmivasan, C. (1985). *Estimation of a Covariance Matrix under Stein's Loss,* Ann. Statis., **13**, 1581–1591.

Efron, B. (1983). *Estimating the Error Rate of a Prediction Rule: Improvement on Cross–Validation,* J. Amer. Statis. Assoc., **78**, 316–331.

Efron, B. and Morris, C. (1976). *Multivariate Empirical Bayes and Estimation of Covariance Matrices,* Ann. Statis., **4**, 22–32.

Enis, P. and Geisser, S. (1974). *Optimal predictive linear discriminants.* Annals of Statist., **2**, 403–410.

Geisser, S. (1977). Discrimination, allocatory and separatory, linear aspects. *Classification and Clustering,* ed. J. Van Ryzin. Academic Press.

Golub, G. H. and Van Loan, C. F. (1983). *Matrix Computations.* The Johns Hopkins University Press, Baltimore.

Gong, G. (1982). Cross–Validation, the Jackknife, and the Bootstrap. Excess Error Estimation in Forward Logistic Regression. Ph.D. dissertation, Department of Statistics, Stanford University Tech. Report No. 80.

Haff, L. R. (1980). *Empirical Bayes Estimation of the Multivariate Normal Covariance Matrix,* Ann. Statist., **8**, 586–597.

James, M. (1985). *Classification Algorithms.* William Collins Sons & Co., Ltd. London.

James, W. and Stein, C. (1961). *Estimation with Quadratic Loss,* Proc. Fourth Berkeley Symp. Math. Stat. Prob., **1**, 361–379, University of California Press.

Kwan, W. and Kowalski, B. R. (1980). *Data Analysis of Sensory Scores. Evaluations of panelists and wine score cards,* J. Food Sci., **45**, 213–216.

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

Lin, S. P. and Perlman, M. D. (1984). *A Monte Carlo Comparison of Four Estimators for a Covariance Matrix,* Multivariate Analy., **6**, 41–429., ed., P. R. Krishnaiah, North Holland, Amsterdam.

Olkin, I., and Sellian, J. B. (1977). Estimating Covariances in a Multivariate Normal Distribution. *Statistical Decision Theory and Related Topics II,* 313–326, ed., S. S. Gupta and D. Moore, Academic Press, New York.

O'Sullivan, F. (1986). *A Statistical Perspective on Ill–Posed Inverse Problems,* Statistical Science, **1**, 502–527.

Stein, C. (1973). *Estimation of the Mean of a Multivariate Normal Distribution,* Proc. Prague Symp. Asymptotic Statist., 345–381.

Stein, C. (1975). Reitz Lecture, 38th Annual Meeting IMS., Atlanta, Georgia.

Stein, C., Efron, B. and Morris, C. (1972). *Improving the Usual Estimator of a Normal Covariance Matrix,* Department of Statistics, Stanford University Report No. 37.

Takemura, A. (1984). An Orthogonally Invariant Minimax Estimator of the Covariance Matrix of a Multivariate Normal Population, Tsukuba J. Math., **8**, 367–376.

Titterington, D. M. (1985). *Common Structure of Smoothing Techniques in Statistics,* Int. Statist. Review, **53**, 141–170.

Wald, P. W. and Kronmal, R. A. (1977). Discriminant functions when covariances are unequal and sample sizes are moderate. *Biometrics,* **33**, 479–484.