Journal of Data Science,17(1). P. 1-36 ,2019
DOI:10.6339/JDS.201901_17(1).0001

A COMPARISON OF REGULARIZED LINEAR DISCRIMINANT FUNCTIONS FOR POORLY-POSED CLASSIFICATION PROBLEMS

L. A. Thompson[*], Wade Davis[2] , Phil D. Young[3],

Dean M. Young[4], Jeannie S. Hill[4]

[*]*Division of Biostatistics, Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, MD 20993*

[2]*Department of Health Management and Information, Department of Statistics, University of Missouri, 187 Galena Hall, DC018.0, Columbia, MO 65211*

[3]*Paul L. Foster Campus for Business and Innovation, 1621 S. 3rd St., Baylor University, Waco, TX 76706*

[4]*Department of Statistical Science, Baylor University, Waco, TX 76798-7140*

[*]Corresponding author: lthompson10@yahoo.com

Abstract

For statistical classification problems where the total sample size is slightly greater than the feature dimension, regularized statistical discriminant rules may reduce classification error rates. We review ten dispersion-matrix regularization approaches, four for the pooled sample covariance matrix, four for the inverse pooled sample covariance matrix, and two for a diagonal covariance matrix, for use in Anderson's (1951) linear discriminant function (*LDF*). We compare these regularized classifiers against the traditional *LDF* for a variety of parameter configurations, and use the estimated expected error rate (*EER*) to assess performance. We also apply the regularized *LDF*s to a well-known real-data example on colon cancer. We found that no regularized classifier uniformly outperformed the others. However, we found that the more contemporary classifiers (e.g., Thomaz and Gillies, 2005; Tong et al., 2012; and Xu et al., 2009) tended to outperform the older classifiers, and that certain simple methods (e.g., Pang et al., 2009; Thomaz and Gillies, 2005; and Tong et al., 2012) performed very well, questioning the need for involved cross-validation in estimating regularization parameters. Nonetheless, an older regularized classifier proposed by Smidt and McDonald (1976) yielded consistently low misclassification rates across all scenarios, despite the shape of the true covariance matrix. Finally, our simulations showed that regularized classifiers that relied primarily on asymptotic approximations with respect to the training sample size rarely outperformed the traditional *LDF*, and are thus not recommended. We discuss our results as they pertain to the effect of high dimension, and offer general guidelines for choosing a regularization method for poorly-posed problems.

Keywords: Poorly-posed classification problems, Shrinkage estimator, Eigenvalue adjustment, Expected error rate

1 Introduction

For classification problems with high dimensional sets of features, feature selection methods are often used to locate the features most important for separating classes. Once these methods are used, one hopes that the resulting classification problem can proceed using conventional methods such as linear or quadratic discriminant analysis. For example, Soukup and Lee (2004) construct a classifier to classify malignant and benign tumors using 57 tissue samples (36 tumor; 21 normal) and 2,000 genes. The number of genes was reduced to the two most-relevant genes using a stepwise feature selection method. The resulting classifier correctly classified all 19 test samples. Feature selection methods can be very successful in reducing the classification problem to a size where conventional methods may be used. Indeed, a two-stage combination of principal components analysis followed by *LDA* has been used in the facial recognition literature for years (Belhumeur, et al., 1997). However, with some problems, the number of important features does not reduce to a small set of genes, and might remain with more genes than the number of training samples. For example, for the same data set, Klaus (2013) found between 13 and 168 genes depending on the variable selection method used. Therefore, it can be important to use a regularization method after feature selection when building classifiers for high dimensional data sets.

Some regularization methods include a variable selection step. For example, in Guo et al.'s (2007) "Shrunken Centroids" regularized classifier, the class means of each gene are shrunk toward their grand means, based on a threshold, to the extent that the class-specific means may be determined not to differ across classes. This procedure selects the most important genes for classification. More recently, Ramey et al. (2017) included variable selection for their regularization of the quadratic discriminant function, and Klaus (2013) introduced a variable selection method called "misclassification rate based variable selection" that chooses features based on their effect sizes. However, most regularization methods do not also introduce a new variable selection method. Therefore, we consider a situation where feature selection has been completed (perhaps qualitatively), leaving the number of features just under the training sample size, and in need of further dimension reduction. When the number of features is less than, but close to the total training size, the classification problem may be called *poorly-posed*. Despite the total training sample size being larger than the number of features, poorly-posed problems can still lead to instability in estimating the classifier (Seber, 1984; Bai and Saranadasa, 1996).

In this paper we compare ten regularization methods plus *LDA* for the linear discriminant function (*LDF*) using four configurations of true means and covariances taken from Friedman (1989). The methods were selected so that they apply to the *LDF*, and do not include a feature selection step. However, we also evaluate Guo et al.'s (2007) and Klaus' (2013) methods when we re-analyze the colon cancer data set above. In addition, we evaluate a contemporary regularization method from the facial recognition literature (Yang and Wu, 2014) on this data set. We focus on linear discriminant analysis because of its parsimony in terms of covariance elements, and its general good performance in discrimination tasks even under evidence of different class covariance matrices (e.g., Dudoit, et al., 2002).

We have organized the remainder of the paper as follows. In Section 2, we introduce notation regarding the *LDF* for classification. In Section 3, we briefly discuss poorly-posed *LDF*s and several previously proposed solutions, while in Sections 4, 5, and 6, we review the different regularization approaches. In Section 7 we describe the simulation configurations and the subsequent simulation results. We then apply the competing linear discriminant regularized classifiers and the *LDF* to the colon cancer data set from Soukup and Lee (2004) in Section 8. We give a discussion in Section 9, providing guidelines based on simulation results, and mention further results in an appendix.

2 Notational Background

Consider *L* distinct populations or classes that occur with prior probabilities $\pi_k = 1, \dots, L$. For a statistical classification problem, we wish to construct a classification rule that classifies unlabeled objects into one of the *L* distinct groups based on a vector of measurements or features. For multivariate normal population problems with known means and equal known covariance matrices, the Bayes classifier is equivalent to classifying an unlabeled real observation vector x into the *k*th class using the following rule:

$$Assign\ x\ to\ C_k \Leftrightarrow d_k(x) = \min_{k=1,\dots,L} (x-\mu_k)' \sum^{-1} (x-\mu_k) - 2(ln\,\pi_k), \qquad (1)$$

where $\mu_k \in \mathbb{R}_{p\times 1}$ is the k^{th} population mean, $\sum \in \mathbb{R}_p^>$ is the common covariance matrix, $\mathbb{R}_{p\times q}$ represents the real space for a $p \times q$ matrix, $\mathbb{R}_p^>$ represents a positive definite $p \times p$ matrix, and π_k is the *a priori* probability that an unlabeled observation x belongs to population $C_k, k = 1, \dots, L$. Because the population parameters μ_k and Σ are unknown in practice, one must estimate them from training samples whose class origins are known. Here, we consider only the case of two Gaussian populations C_1 and C_2.

Assuming equal *a priori* probabilities π_k, $\mathrm{k} = 1,\ 2,$ and equal covariance matrices $\sum_1 = \sum_2$, one can easily show that an estimated Bayes rule for classifying an unlabeled observation x is:

Assign x to C_1 if $W(x) \geq 0$; otherwise, assign x to C_2, where

$$W(x) \equiv \left[x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\right]' S^{-1}(\bar{x}_1 - \bar{x}_2) \qquad (2)$$

with $\bar{x}_k = (1/n_k)\sum_{i=1}^{n_k} x_{ki}$ as the k^{th} class sample mean and $S = (n_1 + n_2 - 2)^{-1}[(n_1 - 1)S_1 + (n_2 - 1)S_2]$ as the common estimated covariance matrix, where $S_k = \big(1/(n_k - 1)\big)\sum_{i=1}^{n_k}(x_{ki} - \bar{x}_k)(x_{ki} - \bar{x}_k)'$ is the estimated covariance matrix for class k, $i = 1, \dots, n_k$, and $k = 1,\ 2$. The discriminant function (2) is known as Anderson's linear discriminant

function (*LDF*) (see Anderson, 1951). One can express the expected error rate (*EER*) associated with (2) analytically as follows. Using a 0-1 loss function and assuming $\pi_1 = \pi_2$, we have that

$$EER \equiv \frac{1}{2}\sum_{k=1}^{2} P((-1)^{2-k}W(x) \geq 0 | x \in C_k). \quad (3)$$

John (1961) has derived the exact unconditional distribution of (2) under the assumption of multivariate normality, though it is cumbersome. However, one can condition on the current training samples and use (3) as a measure of the probability of misclassification for (2). Thus, conditioning on $\bar{x}_1$, $\bar{x}_2$, and S while assuming $\pi_1 = \pi_2$, we have that the *conditional error rate* (*CER*),

$$CER \equiv \frac{1}{2}\sum_{k=1}^{2} \Phi\left(\frac{(-1)^{2-k}\left(\mu_k - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\right)' S^{-1}(\bar{x}_1 - \bar{x}_2)}{[(\bar{x}_1 - \bar{x}_2)' S^{-1} \sum S^{-1}(\bar{x}_1 - \bar{x}_2)]^{\frac{1}{2}}} \right), \quad (4)$$

is an explicit function of the estimators $\bar{x}_k, S$, the location parameter $\mu_k, \mathrm{k} = 1,2$, and the common dispersion parameter Σ. Averaging (4) over the distributions of the sample moments estimates the *EER* in (3).

When the total training-sample size $N = n_1 + n_2$ is slightly greater than p, the inverse covariance matrix estimator S^{-1} exists in theory but is highly unstable. The instability has been discussed under the general category of asymptotics of multivariate statistics when $N, p \to \infty$ but $p/N \to 1$ (e.g., Yao, et al., 2015). Bai (1999) showed that the limiting spectral distribution (*LSD*) of S (the limiting empirical distribution of the sample eigenvalues) is the (generalized) Marčenko-Pastur distribution with ratio index y, where $p/N \to y \in (0, \infty)$. As a result, Yao et al. (2015) illustrate that when $\Sigma = I_p$, the sample eigenvalues of S do not converge to the population values (of 1) as $p/N \to y \in (0, \infty)$. Because the sample eigenvalues are functions of the elements of S, S is not reliably estimated. Consequently, the estimator S^{-1} produces a highly volatile and unreliable classification rule (2), notably when $y \in (0, \infty)$ (Bai and Silverstein, 1996). Thus, although additional feature information may be available to discriminate among the two classes, classification accuracy does not improve unless one obtains enough training-sample observations to reliably estimate the increased number of parameters (i.e., when $p/N \to 0$, and hence the *LSD* of S has "jumps" at the true eigenvalues). When *N* is small relative to p but $N > P$, we consider the classification problem to be *poorly-posed*; when $N < P$, we define the classification problem to be *ill-posed*. This paper focuses mainly on poorly-posed problems for Anderson's *LDF* given in (2), but addresses the case of $N = P$ for appropriate methods.

We use Monte Carlo simulations to compare different dispersion matrix regularization approaches, using the estimated *EER* as our comparison criterion. We examine the performance of these regularization techniques, along with the *LDF*, for the two-class problem using several types of mean configurations and covariance matrices, for both small and moderate training-sample sizes relative to p.

3 A Poorly-posed Linear Discriminant Function

One can readily see the effect of an unstable dispersion estimator on (2) when expressing S^{-1} in terms of the spectral decomposition as $S^{-1} = \sum_{j=1}^{p} v_j v_j' / e_j$, where e_j is the jth largest eigenvalue of S and v_j is the associated eigenvector, $j = 1, \ldots, p$. Using this notation, one can express the estimated discriminant score for (1) as

$$\hat{d}_k(x) \equiv \sum_{j=1}^{p} \frac{\left[v_j'(x - \bar{x}_k)\right]^2}{e_j} \tag{5}$$

Clearly, the smallest eigenvalues and the directions associated with their eigenvectors highly influence (5). The eigenvalues of S^{-1} are well known to be biased such that the smallest eigenvalues are underestimated (see, for example, Seber, 1984, and Ledoit and Wolf, 2004). This bias increases as the training-sample size decreases relative to p. Consequently, with relatively small total training-sample size N, (2) yields an inaccurate estimator of the population discriminant rule.

A more formal demonstration (and proof) of the unreliability of S^{-1} when $N \sim p$ is provided in Bai and Saranadasa (1996), and in comprehensive detail in Yao et al. (2015), and references therein. Briefly, with *iid* random *p*-vectors $x_1, \ldots, x_N$ with mean 0 and covariance matrix I_p, when $p/N \to y \in (0,1)$, the ratio of the largest to smallest eigenvalues of S tends to $\left(1 + \sqrt{y}\right)^2 / \left(1 - \sqrt{y}\right)^2$. Hence, when *y* is close to 1 (so, for example, p is close to, but less than, N), S^{-1} will have some huge eigenvalues. Bai and Saranadasa (1996) illustrate that this problem can occur in practice even if p and N are not especially large, which is the case for our simulations presented later.

Conventional statistical methods of dealing with poorly-posed discriminant analysis problems include variable selection or dimension reduction procedures, such as principal components analysis, canonical analysis, and other dimension-reduction methods (see Cook and Yin, 2001 or Tibshirani, et al., 2003). Although feature selection is popular and effective for ill-posed scenarios where $N \ll p$, such as in the genomics literature, we do not consider it formally here because our focus is mainly on regularization as a method of dimension reduction, potentially after feature selection has been done. Nonetheless, for the case $N = P$, we consider methods that incorporate a feature selection property in their estimation.

Regularizing the sample covariance matrix prior to inverting S offers one solution for reducing the variability in (5), and researchers have proposed many other covariance matrix regularization algorithms to stabilize (2) (see Mkhadri, Celeux, and Nasroallah (1997), Greene and Rayens (1989), Rayens and Greene (1991), or Peck and Van Ness (1982)). In the next three sections, we describe both recent and older methods from the literature.

4 Covariance Matrix Estimator Regularization

Numerous methods of regularization in the linear discriminant case shrink the extreme sample eigenvalues toward more moderate and, thus, more stable values. Known as *eigenvalue adjustment*, this technique effectively decreases larger eigenvalues while increasing smaller eigenvalues to counter their bias (Koolaard, Ganesalingam, and Lawoko, 1998). One method for performing eigenvalue adjustment consists of augmenting the pooled sample covariance matrix with a matrix that is proportional to the identity matrix. Early attempts by DiPillo (1976), Smidt and McDonald (1976a), and Smidt and McDonald (1976b) have substituted the ridge-like estimator $\left(S+\gamma I_p\right)^{-1}$, $0 \leq \gamma \leq 1$, in place of S^{-1} in (2). DiPillo (1976) and DiPillo (1979) have given results for the optimal value of γ for various parameter configurations, while Smidt and McDonald (1976b) have shown that for certain values of γ, $\left(S+\gamma I_p\right)^{-1}$ has smaller mean squared error than S^{-1}.

4.1. Smidt and McDonald's (1976) Shrunken Sample Covariance Matrix Estimators

Although Smidt and McDonald (1976a) have proposed several estimators of γ, we consider two of their covariance matrix estimators of the form

$$\hat{\Sigma}_{SM}^{-1} \equiv \left[S+\left(r\hat{\lambda}_p\right)I\right]^{-1},$$

where $\hat{\lambda}_p$ is the smallest eigenvalue of S and *r* is a positive constant determined from the data. Smidt and McDonald (1976b) have shown that when γ is proportional to $\hat{\lambda}_p$, the bias in the estimated eigenvalues is corrected. The difference between the two estimators that we use is the method applied for choosing the shrinkage estimator r.

The first regularized covariance estimator determines a scalar r that minimizes the leave-one-out cross-validation error rate. We refer to the regularized linear classifier using this shrunken covariance estimator as *SM1*. In the second shrunken covariance estimator, we attempt to find a choice of *r* using the discriminant vector

$$\left[S+\left(r\hat{\lambda}_p\right)I\right]^{-1}(\bar{x}_1-\bar{x}_2) \tag{6}$$

Consider a set of m non-negative numbers $R=(r_1,\ldots,r_m)$, and let $v_{r,i}$ denote the i^{th} element of the vector in (6). We seek to find the value of $r \in R$, say $\mathrm{r_j}$, such that $\left|v_{r_j,i}\right| / \left|v_{r_{j+1},i}\right| \in (0.95,1.05)$ for all $i=1,\ldots,p, j=1,\ldots,m-1$. We use the interval (0.95,1.05) in order for the ratio of successive values of (6) to be relatively stabilized around 1.0. The authors do not recommend a particular method for choosing *m* or the values of $r \in R$. We have found that a grid of values from 0 to 10 by increments of 0.1 has worked well. We refer to this classifier as *SM2*.

4.2. Raudys, Skurikhina, Cibas, and Gallinari's (1994, 1995) Shrunken Covariance Matrix Estimator

Raudys and Skurihina (1994) and Raudys, Skurihina, Cibas, and Gallinari (1995) have derived a regularized inverse covariance matrix estimator that minimizes Raudys' approximation to the EER of the corresponding shrunken classification rule. The regularized covariance matrix inverse estimator of Raudys et al. (1995) is

$$\hat{\Sigma}_{RSCG}^{-1} \equiv \left(S + \hat{\lambda}_{RSCG} I\right)^{-1}.$$

Moreover, Raudys et al. (1995) have shown that the shrinkage estimator that minimizes an asymptotic approximation to the *EER* is

$$\lambda_{RSCG} \equiv C^{-1} - 2\mathrm{B}^{-1} \tag{7}$$

where $\mathrm{B} \equiv \frac{2}{1-y}\beta_2 + \frac{tr D^{-1}}{m}$, $\mathrm{C} \equiv \frac{1}{1-y}\beta_1$, $\beta_k \equiv \frac{M' D^{-1} M}{\delta^2}\alpha_k + \frac{tr D^{-1}}{2m-p}$, for $\mathrm{K} = 1,2\ k = 1, 2$. And, $y \equiv \frac{p}{2m}$, $\mathrm{m} = \mathrm{n}_1 = n_2, \alpha_1 = 1, \alpha\_2 \equiv \left(1 + \frac{2 tr D^{-1}}{m M' D^{-1}}\right)$, $\delta^2 \equiv M'M = (\mu_1 - \mu_2)' \sum^{-1} (\mu_1 - \mu_2)$, and D is a diagonal matrix composed of the eigenvalues of Σ. Clearly, the parameters in (7) must be estimated with sample moments from the training data. We also remark that as the common training sample size *m* increases (with fixed *p*), $\hat{\lambda}_{RSCG} \to 0$, and the resulting classification rule approaches Anderson's *LDF* given in (2). We denote this regularized classification rule by *RSCG*.

4.3. Thomaz and Gillies' (2005) Shrunken Covariance Matrix Estimator

Thomaz and Gillies' (2005) attempt at eigenvalue adjustment adjusts only the smallest eigenvalues of the sample covariance matrix toward the mean of the sample eigenvalues. Each of the sample eigenvalues is replaced by the sample mean of the eigenvalues if the mean is larger than the sample eigenvalue itself. Then, a *new* pooled sample covariance matrix is constructed with the adjusted eigenvalues. The method is intended to stabilize the smallest sample eigenvalues, thereby regularizing the covariance matrix. We denote this regularized classifier by (*NLDA*) to stand for *new* LDA.

5 Shrinkage Estimator Regularization

5.1. Loh's (1995) Regularized Shrunken Inverse Covariance Matrix Estimator

Another approach to regularization involves shrinkage estimates resembling Stein-like biased estimators that are convex combinations of S^{-1} and the identity matrix I_P. Alternatively, Loh (1995) uses a convex combination of S and I_P.

$$\hat{\Sigma}^{-1} = [(1-\gamma)S + (\gamma trS)I_P]^{-1}. \tag{8}$$

One obvious advantage to regularizing S rather than S^{-1} is that no computational stability issues occur as with computing S^{-1}.

Loh (1995) has derived an expression for γ that minimizes an asymptotic expansion of the difference in *EER* between the regularized linear classification rule and (2). The shrinkage-constant estimator given in Loh (1995) is

$$\gamma^*_{ASYMP} \equiv \frac{\frac{p}{trS}\left\{\left(\frac{n_1+n_2}{n_1 n_2}\right)[(trS^{-1}) - pD_1^{-2}D_2^2] + \frac{1}{n_1+n_2-2}\,[D_1^2(trS^{-1}) - D_2^2]\right\}}{(D_3^2 - D_1^{-2}D_4^2)},$$

where $D_j^2 \equiv (\bar{x}_1 - \bar{x}_2)'S^{-j}(\bar{x}_1 - \bar{x}_2),\ j = 1,2,3,4$. If $\gamma^*_{ASYMP} \notin (0,1)$, one sets γ^*_{ASYMP} to 0 or 1, appropriately. The inverse of the shrunken covariance matrix estimator is then

$$\hat{\Sigma}^{-1}_{LOH} \equiv \left[(1-\gamma^*_{ASYMP})S + \frac{\gamma^*_{ASYMP}}{p}(trS)I_p\right]^{-1} \tag{9}$$

Loh (1995) has compared the performance of the corresponding regularized discriminant function using (9) with the *LDF* and ridge classification rules where $\hat{\gamma}$ minimizes either the cross-validated *CER* or a bootstrap error rate. We remark that these criteria for selecting γ are relatively more computationally demanding. We denote the classification rule corresponding to this estimator by *LOH*.

5.2. Mkhadri's (1995) Regularized Inverse Covariance Matrix Estimators

Mkhadri (1995) has proposed two inverse covariance matrix estimators that are convex combinations of S^{-1} and I_p, where the regularization parameter γ is chosen to minimize either a cross-validated estimated *EER* or an estimated generalized distance measure. We refer to the two corresponding regularized classification rules as *MCV* and *MGD*, respectively. The first regularized inverse covariance estimator uses the shrinkage parameter that minimizes the cross-validated risk, where for $i = 1, \ldots, n_1 + n_2$,

$$\gamma_{opi}^{CV} = \begin{cases} 0, & if\ only\ W^{(i)}\ (x_i) is\ used \\ 1, & if\ only\ E^{(i)}\ (x_i) is\ used \\ \left[aW^{(i)}(x_i)\right]/\left[aW^{(i)}(x_i) - \left(b/trS_{/i}\right)E^{i}(x_i)\right], & otherwise \end{cases} \tag{10}$$

$S_{/i}$ is the sample covariance matrix with the ith observation removed. Here, $W^{(i)}(x_i)$ is (2) with the ith observation removed, $E^{(i)}\ (x_i)$ is the Euclidean classifier, which is $W^{(i)}(x_i)$ with the matrix I in place of $S_{/i}^{-1}$, and a and b are constants defined later. As alluded to in Mkhadri (1995), a maximum of $N + 2$ candidate regularization parameters exist for any training set (the additional 2 candidates account for values of 0 and 1). For each $\gamma_{op_i}^{CV}$, we calculate the regularized leave-one-out classifier, leaving out each observation in turn, except the ith used to calculate $\gamma_{op_i}^{CV}$. The value of $\gamma_{op_i}^{CV}$ with the lowest misclassification rate among the $N - 1$ rates is chosen as the optimal regularization parameter for that training sample.

The second regularized inverse covariance matrix estimator uses the shrinkage constant that minimizes the distance between the training-set observations and their actual class mean vectors, yielding the criterion expression

$$\widehat{D}^{*}(\gamma) \equiv \sum_{k=1}^{2}\sum_{i\in C_k}\left(x_{ki} - \bar{x}_{k/i}\right)' S_{/i}^{*-1}(\gamma)\left(x_{ki} - \bar{x}_{k/i}\right) + \ln\left|S_{/i}^{*-1}(\gamma)\right| \ , \tag{11}$$

where $S_{/i}^{*-1}(\gamma)$ is the regularized estimator of $S_{/i}^{-1}$,and $\bar{x}_{k/i}$ is the sample mean vector for the kth group with the ith observation omitted. The optimal shrinkage constant is

$$\gamma_{op}^{G} \equiv \frac{aD_1 - bD_2 + \sum_k\sum_i\sum_{j=1}^{p}\left(\beta_k^j - 1\right)}{\sum_k\sum_i\sum_{j=1}^{p}\left(\beta_k^j - 1\right)^2}, \tag{12}$$

$$\mathrm{D}_1 \equiv \sum_k\sum_i \left(x_{ki} - \bar{x}_{k/i}\right)' S_{/i}^{-1}\left(x_{ki} - \bar{x}_{k/i}\right),$$

$$\mathrm{D}_2 \equiv \sum_k\sum_i \left(x_{ki} - \bar{x}_{k/i}\right)' \left(x_{ki} - \bar{x}_{k/i}\right)/\left(\frac{n}{n-1}trS - r_i' r_i\right)$$

,

$\beta_{ki}^{j} \equiv \frac{b(\lambda_j - r_{ki}' r_{ki})}{a\left(\frac{n}{n-1}\right)trS - r_{ki}' r_{ki}}$,$\mathrm{r_{ki}} \equiv \frac{(x_{ki} - \bar{x}_k)}{\sqrt{(n-1)(n_k-1)/n_k}}$, $n \equiv n_1 + n_2 - 2$, $a \equiv (n - p - 3)/n$, $b \equiv p/n$, λ_j is the jth eigenvalue of S, and $j = 1, \ldots, p$. The resulting two regularized inverse covariance-matrix estimators are

$$\hat{\Sigma}_{MH}^{-1} \equiv \left(1 - \gamma_{op}^{H}\right) aS^{-1} + \left(b\gamma_{op}^{H}/trS\right) I_p, \text{ where } H = CV, GD. \tag{13}$$

Our formulation of (12) differs slightly from that of Mkhadri (1995). We have provided the modification to reconcile Mkhadri's (1995) regularized classifier with that of the regularized classifier of Friedman (1989). Furthermore, if $n \leq p + 3$, we set $a = 1/n$ to ensure that a is always positive.

In a series of simulations using different Mahalanobis distances and intraclass correlation matrices, Mkhadri (1995) has shown that the *MCV* and *MGD* linear classifiers perform as well as or better than (2) for the configurations he considered. He has also noted that the shrinkage parameter values for (12) tend to be smaller when one is minimizing the generalized distance (11), as opposed to the cross-validated *EER*.

5.3. Xu, Brock, and Parrish's (2009) Modified Regularized Inverse Covariance Matrix Estimator

Xu et al. (2009) introduced a *modified* regularizer that uses Ledoit and Wolf's (2004) "well-conditioned estimator" for large-dimensional covariance matrices. Their goal is to find the linear combination $\Sigma^* = \rho_1 I + \rho_2 S$ whose expected quadratic loss $\mathrm{E}(\|\Sigma^* - \Sigma\|^2)$ is minimized, where $\|A\| = \sqrt{tr(AA'/p)}$. Ledoit and Wolf (2004) show that the solution is $\Sigma^* = \frac{\beta^2}{\delta^2}\mu I + \frac{\alpha^2}{\delta^2}S$, where $\mu = tr(\Sigma)/p$, which can be estimated using the sample covariance matrix. The other parameters can also be estimated consistently using sample values. See Ledoit and Wolf (2004) for the expressions. The authors note that μI can be interpreted as a shrinkage target, with its corresponding weight $\frac{\beta^2}{\delta^2}$ the normalized shrinkage intensity. As the shrinkage intensity increases, there is more shrinkage toward the target than toward the sample covariance matrix. The optimally shrunken estimator Σ^* corrects for the bias in the sample eigenvalues by shrinking them toward their grand mean. The estimated Σ^* is then used in (2). Xu et al. (2009) showed that their *Modified* regularizer outperformed several other regularized estimators, including Thomaz and Gilles' (2005) estimator mentioned above, as well as Guo et al.'s (2007) "Shrunken Centroids" regularizer (referred to as *SCRDA*). However, their problems were ill-posed, where the number of features was much larger than the number of observations. We evaluated Xu et al.'s (2009) modified regularizer in poorly-posed problems, and found (consistent with the authors' Figure 3 and Ledoit and Wolf's simulations) that the advantages of their classifier show up when *p* approaches *N*, but that it is also competitive in our poorly-posed scenarios. We refer to the classification rule (2) that uses Xu et al.'s (2009) modified regularized covariance matrix as *MLDA* to be consistent with their paper.

6 Diagonal Estimators of the Covariance Matrix

Another category of regularization focuses on diagonal covariance matrices. Although it may rarely be true that features are truly uncorrelated, Dudoit et al. (2002) showed good performance for diagonal discriminant analysis even with correlation among features. We consider two related estimators of this type. The first adjusts the diagonal elements of the sample covariance u.sing a shrinkage parameter estimated by minimizing the average risk using a James-Stein loss function. The second uses the sample diagonal common covariance matrix, and adds a shrinkage adjustment to the sample class mean vectors so that the average risk is minimized under quadratic loss.

6.1. Pang et al.'s (2009) Shrinkage-based Diagonal Discriminant Analysis

Pang et al. (2009) use a diagonal linear discriminant rule, but incorporate regularization of the estimated diagonal covariance matrix in order to improve reliability when p is large. They start with estimates of the feature-specific precisions on the diagonal of the inverse of the common sample covariance matrix, then adjust the precisions using a shrinkage parameter, α, so that for feature j, the precision estimate is $\tilde{\sigma}_j^{-2}(\alpha)$, where $0 \leq \alpha \leq 1$. Pang et al. (2009) use Tong and Wang's (2007) family of shrinkage estimators

$$\tilde{\sigma}_j^{-2}(\alpha) = \left(h_p \hat{\sigma}_{pool}^{-2}\right)^{\alpha} \left(h_1 \hat{\sigma}_j^{-2}\right)^{1-\alpha}$$

where $h_p = \left(\frac{2}{N-2}\right)\left(\frac{\Gamma\left(\frac{N-2}{2}\right)}{\Gamma((N-2)/2-1/p)}\right)^p$, $\hat{\sigma}_{pool}^{-2} = \prod_{j=1}^{p}\left(\hat{\sigma}_j^{-2/p}\right)$, $\hat{\sigma}_j^{-2}$is the jth feature's estimated precision, p is the number of features, and N is the combined sample size of the two classes. The shrinkage parameter α is chosen as the minimizer of the average risk when the loss function is the Stein loss function $L_{Stein}(\sigma^2, \tilde{\sigma}^2) = \tilde{\sigma}^2/\sigma^2 - ln(\tilde{\sigma}^2/\sigma^2) - 1$ (see Pang et al., 2009, for details). The resulting regularized inverse covariance matrix is

$$\hat{\Sigma}_{SDLDA}^{-1} = Diag\left(\tilde{\sigma}_1^{-2}(\alpha^*), \ldots, \tilde{\sigma}_p^{-2}(\alpha^*)\right) \tag{14}$$

where α^* is the minimizer mentioned above.

After shrinking the inverse covariance matrix, Pang et al. (2009) discuss regularizing between the individual class-based covariance estimates and the pooled covariance estimate. However, because our work deals only with *linear* discriminant analysis, we do not study this regularization. We refer to Pang et al.'s estimator of the *LDF* that uses (14) as *SDLDA*.

6.2. Tong et al.'s (2012) Mean-adjusted Diagonal Discriminant Analysis

Tong et al. (2012) propose a shrinkage-based diagonal discriminant rule that replaces the sample class means by estimators that shrink them toward their grand mean. When the number of training class samples is small, in particular less than the number of features, the ordinary sample class means can be unreliable. The authors' proposed class mean estimator is based on the posterior mean under a hierarchical Bayesian model with conjugate priors, as well as the James-Stein type estimator, incorporating a shrinkage parameter, *r*:

$$\hat{\mu}_k(r) = \left(1 - \frac{r}{\|\bar{X}_k\|_S^2}\right)\bar{X}_k \tag{15}$$

where S estimates the common covariance matrix (the authors use a diagonal matrix with sample feature variances on the diagonal), and $\bar{X}_k$ is the sample class mean. The authors note that the estimator in (15) dominates the sample mean under a quadratic loss function when $0 < \mathrm{r} < 2(p-2)/(N+1)$.

Tong et al. (2012) provide the expression for the optimal shrinkage parameter minimizing risk under a quadratic loss function, then show that for high dimensional data (large *p*), it can be approximated by $\hat{r}_{opt} = [(N-1)(p-2)]/[N(N-3)]$. The optimal shrinkage estimator for the sample mean then substitutes $\hat{r}_{opt}$ for *r* in equation (15). They then substitute these class specific shrinkage mean estimators into the *LDF*, along with the diagonal sample common covariance matrix estimate to arrive at their shrinkage-based diagonal discriminant rule. In our simulations, we use the shrinkage estimator applied to the deviations of the class means from the grand mean (see equation (6) in Tong et al. (2012)). We refer to the resulting classification rule as *SmDLDA*. We also tried using the diagonals of Pang et al.'s (2009) regularized covariance matrix in Section 6.1, but the results were almost identical to using *SDLDA*. This occurrence indicates that both adjustments may not be beneficial to the scenarios we examined, answering a question posed by Tong et al. (2012) in their discussion.

7 Simulation Description and Results

We conducted simulations to evaluate the classification performance of the classifiers described in Sections 4 through 6, in addition to the *LDF*. For each simulation, we generated 100,000 training samples (using a high performance computing cluster running 1,000 nodes simultaneously) from specified Gaussian distributions with dimension $p = 50$ and equal training sample sizes of either $n_1 = n_2 = 26$, $n_1 = n_2 = 35$, or $n_1 = n_2 = 50$. For each regularized inverse covariance estimator $S_{R_j}^{-1}$ and training set, and assuming $\pi_1 = \pi_2$, the *CER* is

$$CER_{R_j} \equiv \frac{1}{2}\sum_{k=1}^{2} \Phi\left(\frac{(-1)^{2-k}\left(\mu_k - \frac{1}{2}(\bar{X}_1 + \bar{X}_2)\right)' S_{R_j}^{-1}(\bar{X}_1 - \bar{X}_2)}{\left[(\bar{X}_1 - \bar{X}_2)' S_{R_j}^{-1} \sum S_{R_j}^{-1}(\bar{X}_1 - \bar{X}_2)\right]^{1/2}}\right)$$

where $S_{R_j}^{-1}$, j = *SM1*, *SM2*, *RSCG*, *LOH*, *MCV*, *MGD*, *MLDA*, *NLDA*, *SDLDA*, *SmDLDA*, *LDF*, are the competing inverse sample covariance matrices defined in Sections 4 through 6, and μ_k, k = 1,2, and Σ are the multivariate Gaussian location and dispersion parameters, respectively. For each classifier and for (2), we reported the estimated *EER* of the corresponding classifier and its standard deviation by averaging over the 100,000 training sets. We also reported the average of each regularization parameter and its associated standard deviation.

To accommodate ties in the choice of parameters for any of the configurations, we chose the estimate of the parameter(s) that yielded the smallest level of regularization. Some evidence exists that the manner in which ties are broken does not affect *CER* estimates (Koolaard et al. 1998). However, for the estimators that used the tie-breaking scheme (i.e., *MCV* and *SM1*), the tie-breaking scheme appeared to moderately affect the *EER* (In the Appendix, we present the results after breaking ties in the direction of more regularization). Descriptions of each of the four simulation configurations (A-D) appear in Table 1. The configurations are similar to those used by Friedman (1989).

For each configuration, we have used $\Sigma = diag(e_j)$, $j = 1, \ldots, p$, for the common population covariance matrices, and we have normed the mean differences to obtain a Mahalanobis distance of 3.76, which corresponds to a Bayes classification error of 0.03.

Table 1: Simulation Configurations for $p = 50$ and $\mu_1 = O_p$

	Description	μ_2	Eigenvalues of Σ
A	Means differ in one orthogonal direction; spherical covariance	$\begin{bmatrix}\sqrt{3.76}\\ 0_{p-1}\end{bmatrix}$	$e_j = 1, 1 \le j \le p$
B	Means differ in first $p-1$ features; differ in low-variance subspace	$\mu_{2j} = 2.5\sqrt{e_j/p}\,\frac{p-j}{p/2-1}$ $1 \le j \le p$	$e_j = [9(j-1)/(p-1)+1]^2$
C	Means differ in last $p-1$ features; differ in high-variance subspace	$\mu_{2j} = 2.5\sqrt{e_j/p}\,\frac{j-1}{p/2-1}$ $1 \le j \le p$	$e_j = [9(j-1)/(p-1)+1]^2$
D	Means differ in low-variance subspace and are informative in first $[p/2]$ directions	$\mu_{2j} = \begin{cases} e_j, 1 \le j \le \lfloor p/2 \rfloor \\ 0, j > \lfloor p/2 \rfloor \end{cases}$	$e_j = [9(j-1)/(p-1)+1]^2$

7.1. Configuration A: Means differ in one orthogonal direction; equal eigenvalues

In the first parameter configuration, the eigenvalues of the common population covariance matrix are all equal to one. Therefore, we expect greater shrinkage toward the identity matrix for the regularized covariance matrix estimators. As mentioned in Section 3, the sample eigenvalues may be much different from 1, ranging from 0.004 to 3.93. So, regularization that achieves eigenvalue adjustment, especially of the smallest sample eigenvalues, may perform well in classification. Table 2 shows the estimated *EER*s and their corresponding average regularization parameters, along with standard deviations.

Table 2: Configuration A: Estimated EERs, Average Regularization Parameter Estimates, and Corresponding Standard Deviations

	Estimated *EER*s and SDs (·)				Regularization Parameter Estimates and SDs (·)		
Discriminant Function	$n_1 = n_2 = 26$	$n_1 = n_2 = 35$	$n_1 = n_2 = 50$		$n_1 = n_2 = 26$	$n_1 = n_2 = 35$	$n_1 = n_2 = 50$
LDF	0.453 (0.045)	0.350 (0.035)	0.290 (0.026)		*None*	*None*	*None*
SM 1	0.374 (0.085)	0.316 (0.051)	0.274 (0.032)	$\bar{r}SM$ 1	0.357 (1.150)	0.489 (1.374)	0.374 (1.210)
SM 2	0.253 (0.021)	0.235 (0.016)	0.218 (0.012)	$\bar{r}SM$ 2	7.129 (3.279)	8.814 (1.868)	8.958 (1.732)
RSCG	0.453 (0.045)	0.348 (0.035)	0.285 (0.024)	$\bar{\lambda}$	1e-8 (1e-8)	0.001 (0.0003)	0.017 (0.004)
LOH	0.451 (0.046)	0.335 (0.039)	0.268 (0.027)	$\bar{\Upsilon}LOH$	0.0001 (0.001)	0.021 (0.023)	0.105 (0.073)
MCV	0.288 (0.083)	0.346 (0.037)	0.290 (0.025)	$\bar{\Upsilon}MC$	0.813 (0.390)	0.289 (0.328)	0.216 (0.280)
MGD	0.370 (0.106)	0.350 (0.035)	0.290 (0.026)	$\bar{\Upsilon}MG$	0.408 (0.492)	0.0003 (0.003)	0.0001 (0.001)
MLDA	0.250 (0.020)	0.234 (0.016)	0.218 (0.012)	$\bar{\rho}_1$	0.991 (0.032)	0.995 (0.027)	0.997 (0.022)
				$\bar{\rho}_2$	0.009 (0.018)	0.005 (0.013)	0.003 (0.009)
NLDA	0.259 (0.021)	0.242 (0.017)	0.225 (0.013)		*None*	*None*	*None*
SDLDA	0.220 (0.022)	0.207 (0.016)	0.195 (0.011)	$\bar{\alpha}$	0.456 (0.049)	0.438 (0.048)	0.412 (0.046)
SmDLDA	0.331 (0.018)	0.312 (0.017)	0.288 (0.016)		*None*	*None*	*None*

The regularized diagonal rule estimator *SDLDA* gives the lowest error rates for all three total sample sizes. Its average shrinkage parameter is relatively large compared to its average from the other configurations B through D, as will be seen. This is reasonable because the feature variances are actually the same, and so a common pooled value on the diagonal of S should be preferred compared to other forms of S. *NLDA* and *MLDA* also did well due to their equalizing of the sample eigenvalues. In general, estimators that chose greater amounts of regularization on average (in particular, *MLDA* and *SM2*) obtained good classification results, whereas *SM1* and *MGD* chose relatively smaller parameter values, and therefore did not perform as well. One reason why *SM2* performed so much better than *SM1* may be due *SM2*'s attempt to stabilize the sample discriminants, whereas *SM1* does not have such a goal. Mkhadri (1995) found that *MCV* tended to produce larger regularization parameters than *MGD*, which leads to *MCV* performing better than *MGD* in this configuration. The *MGD* classifier had a relatively high estimated *EER* for all sample sizes, as it did not regularize much, therefore performing not much better than the *LDF*.

SM2, MLDA, *NLDA*, and *SDLDA* performed well for all sample sizes while *LOH* and *SM1* did relatively better with the larger training-sample sizes. In both cases, *MLDA* correctly chose a higher shrinkage toward the target μ (which was reasonably estimated as about 1.00 on average). Not surprisingly, the asymptotic-based classifiers *LOH* and *RSCG* performed relatively poorly for the smaller total training-sample size $N = 52$. Based on their average regularization parameter estimates, *LOH* and *RSCG* chose estimates that were too small

compared to the other methods. Finally, the *SmDLDA* diagonal classifier did not perform well in this configuration, at least compared to the other classifiers. One reason may be due to its direct use of the diagonals of **S** instead of any eigenvalue equalization.

Any differences in the estimated *EER*s between *LDF* and the regularized *LDF*s diminished as *N* increased, and *LDF* became suitable.

7.2. Configuration B: Means differ in last $p-1$ features; elliptical covariance matrix

In this configuration, we speculated that regularized classifiers might not perform well in discrimination because dispersion regularizing matrices tend to equalize the unequal eigenvalues, thus making the regularized covariance matrix more spherical. If the covariance matrix becomes more spherical and the dispersion potentially increases in the low-variance subspace, the detection of the mean differences in the low-variance subspace may diminish.

Thus, we thought that smaller estimated regularization parameters would be preferred to improve classification. However, we were incorrect. The true covariance matrix has large magnitude diagonals, covering a wide range. Any given sample covariance matrix may be much different from the true covariance matrix, requiring regularization to stabilize eigenvalues for better class separation. We provide the estimated *EER*s, the corresponding average regularization parameters, and their respective standard deviations in Table 3.

Table 3: Configuration B: Estimated *EER*s, Average Regularization Parameter Estimates, and Corresponding Standard Deviations

	Estimated *EER*s and SDs (·)				Regularization Parameter Estimates and SDs (·)		
Discriminant Function	$n_1 = n_2 = 26$	$n_1 = n_2 = 35$	$n_1 = n_2 = 50$		$n_1 = n_2 = 26$	$n_1 = n_2 = 35$	$n_1 = n_2 = 50$
LDF	0.387 (0.072)	0.185 (0.041)	0.290 (0.026)		*None*	*None*	*None*
SM 1	0.335 (0.093)	0.171 (0.041)	0.274 (0.032)	$\bar{r}SM\ 1$	0.102 (0.492)	0.720 (1.905)	2.086 (3.098)
SM 2	0.171 (0.033)	0.136 (0.024)	0.218 (0.012)	$\bar{r}SM\ 2$	7.079 (3.282)	8.267 (2.208)	7.713 (2.424)
RSCG	0.387 (0.072)	0.184 (0.041)	0.285 (0.024)	$\bar{\lambda}$	4e-8 (6e-8)	0.002 (0.001)	0.020 (0.016)
LOH	0.382 (0.074)	0.174 (0.040)	0.268 (0.027)	$\bar{\gamma}LOH$	3e-5 (0.0002	0.003 (0.004)	0.006 (0.006)
MCV	0.247 (0.074)	0.184 (0.041)	0.290 (0.025)	$\bar{\gamma}MC$	0.840 (0.366)	0.219 (0.309)	0.162 (0.274)
MGD	0.311 (0.101)	0.186 (0.041)	0.290 (0.026)	$\bar{\gamma}MG$	0.457 (0.498)	0.0003 (0.003)	0.0001 (0.001)
MLDA	0.182 (0.030)	0.156 (0.025)	0.218 (0.012)	$\bar{\rho}_1$	24.077 (1.351)	21.441 (1.069)	18.046 (0.788)
				$\bar{\rho}_2$	0.354 (0.037)	0.424 (0.030)	0.516 (0.022)
NLDA	0.189 (0.031)	0.169 (0.026)	0.225 (0.013)		*None*	*None*	*None*
SDLDA	0.341 (0.034)	0.325 (0.033)	0.195 (0.011)	$\bar{\alpha}$	0.119 (0.009)	0.095 (0.008)	0.071 (0.006)
SmDLDA	0.088 (0.028)	0.078 (0.025)	0.288 (0.016)		*None*	*None*	*None*

Classifiers with higher average regularization parameters performed better than those with less regularization. This applied to both *SM2* and *MLDA*, at least for the two smaller total sample sizes. For all sample sizes, *MLDA* had a relatively high average shrinkage toward the identity matrix. For the largest sample size, this could have resulted in an estimated covariance matrix that was too spherical, leading to a higher error rate. The *SM1* classifier performed relatively poorly for the $N = 52$ case likely due to its low average estimated regularization parameter. *SM2* did better than *SM1*, perhaps due, again, to its process of stabilizing the sample discriminants. *MGD* performed relatively poorly, as it had a smaller average regularization parameter. Again, it chose a smaller regularization parameter on average than did *MCV*. The

RCSG and *LOH* classifiers performed about as well as the *LDF* throughout. The average regularization parameter remained very small for both classifiers.

As well, *SDLDA* reduced its average regularization for the larger sample sizes, and suffered in performance. On the other hand, the other diagonal classifier *SmDLDA* handily beat every other classifier across all total sample sizes. This may have been due to the true mean difference $\mu_2 - \mu_1$ being concentrated in the lower variance subspace. As with Tibshirani et al.'s (2003) Shrunken Centroids Classifier, the less important feature elements for class separation are shrunk toward the overall centroid. This would apply to the *p*th feature (with zero difference between class means), reducing its sample influence on the classification rule, ultimately helping classification.

With the exceptions of *SmDLDA*, *SDLDA*, *NLDA*, and *MDLDA*, as the total sample size rose, the classifiers approached one another in performance. The benefits seen for *MLDA* and *NLDA* for the small training size did not carry over to the large size. As *N* increases, S may better estimate Σ, which does not have equal eigenvalues. Therefore, *NLDA* may err too much by substituting all eigenvalues less than their mean with the mean, especially when the true mean differences are in the low variance subspace (smaller eigenvalues). This may also explain why *MLDA* loses classification ability as *N* increases. If we shrink the sample eigenvalues toward their grand mean, the smaller eigenvalues (with greater class mean difference) will increase, reducing detection of differences. In Configuration C in subsection 7.3, where the class mean differences are in the high variance subspace, *MLDA* does not lose classification ability as *N* increases because detection of differences increases when the larger eigenvalues are shrunk downward. As well, in Configuration D, where the class means differ again in the low-variance subspace, *MLDA* does not lose classification ability as *N* increases because in that configuration, the class mean differences increase with the eigenvalues of Σ. So, shrinking sample eigenvalues may help detect differences.

SmDLDA continued to outperform the other classifiers as the total sample size rose. The diagonal classifier, *SDLDA*, however did not perform well in this scenario, as the sample size increased. *SDLDA* may not perform well when class means differ in the lower variance subspace because the feature variances are shrunk toward a pooled value (potentially increasing them, and thereby diluting some differences). Indeed in Configuration D in subsection 7.4, *SDLDA* also performs relatively poorly.

7.3. Configuration C: Means differ in last $p - 1$ features; elliptical covariance matrix

For the next configuration, the means differed in the high-variance subspace, and therefore, regularization should improve classification performance because of reduced variance in those subspaces with differing means. The estimated *EER*s and average regularization parameters and associated standard deviations for this configuration appear in Table 4.

We saw large improvements over the *LDF* for the *SmDLDA*, *SDLDA*, *SM2*, *MCV*, *MLDA* and *NLDA* classifiers for the case $N = 52$. *MLDA*, *SM2*, and *MCV* all chose relatively high regularization parameter values. *SM1*'s average regularization was lower than *SM2*'s, and appeared to suffer in performance for the smaller sample sizes. The stabilization aspect of *SM2*

apparently helped it choose a better value for *r*, in particular one that is larger to offset a small $\hat{\lambda}_p$. *SM1* chooses *r* to minimize the cross-validation (CV) error rate of the training set, but the minimum CV rate may not necessarily be low.

The diagonal classifiers (*SDLDA* and *SmDLDA*) performed exceptionally well for all sample sizes. *SDLDA* performed very well despite similarly low average regularization as for Configuration B. As the mean differences were in the high variance subspace, shrinking the feature variances toward a pooled value should increase detection of class differences in higher variance subspaces. For *SmDLDA*, the shrunken centroids aspect reduces influence of features with zero differences between class means on the classification rule. This helped for Configuration B, and we will see that it also helps in Configuration D.

For $N = 70$, the regularized classifiers with higher average regularization (e.g., *SM2* and *MLDA*) performed much better than the *LDF*. The *RSCG* classifier had very small mean regularization parameters of near 0 so that it was almost identical to the *LDF* in all cases. For $N = 100$, many of the classifiers performed at the level of the *LDF*. We had expected regularizetion to be beneficial because the population means differed in a high variance subspace, and much more beneficial for the smaller training sample size where eigenvalue estimation is harder. This appeared to be true in general.

Table 4: Configuration C: Estimated *EER*s, Average Regularization Parameter Estimates, and Corresponding Standard Deviations

	Estimated *EER*s and SDs (·)				Regularization Parameter Estimates and SDs (·)		
Discriminant Function	$n_1 = n_2 = 26$	$n_1 = n_2 = 35$	$n_1 = n_2 = 50$		$n_1 = n_2 = 26$	$n_1 = n_2 = 35$	$n_1 = n_2 = 50$
LDF	0.387 (0.072)	0.184 (0.041)	0.108 (0.021)		*None*	*None*	*None*
SM 1	0.332 (0.096)	0.172 (0.042)	0.105 (0.022)	$\bar{r}SM$ 1	0.103 (0.505)	0.407 (1.381)	0.893 (2.120)
SM 2	0.122 (0.040)	0.097 (0.029)	0.079 (0.023)	$\bar{r}SM$ 2	7.028 (3.284)	7.937 (2.381)	7.728 (2.503)
RSCG	0.387 (0.072)	0.184 (0.041)	0.108 (0.021)	$\bar{\lambda}$	6e-8 (7e-8)	0.001 (0.0006)	0.010 (0.004)
LOH	0.381 (0.074)	0.170 (0.041)	0.099 (0.020)	$\bar{\Upsilon}LOH$	3e-5 (0.0002)	0.004 (0.005)	0.011 (0.012)
MCV	0.175 (0.147)	0.183 (0.041)	0.108 (0.021)	$\bar{\Upsilon}MC$	0.706 (0.455)	0.135 (0.223)	0.097 (0.191)
MGD	0.247 (0.162)	0.184 (0.041)	0.108 (0.021)	$\bar{\Upsilon}MG$	0.460 (0.498)	0.0003 (0.003)	0.0001 (0.001)
MLDA	0.082 (0.039)	0.077 (0.035)	0.071 (0.031)	$\bar{\rho}_1$	31.678 (3.523)	28.221 (3.249)	23.792 (2.607)
				$\bar{\rho}_2$	0.149 (0.097)	0.242 (0.090)	0.361 (0.072)
NLDA	0.091 (0.039)	0.085 (0.036)	0.077 (0.034)		*None*	*None*	*None*
SDLDA	0.109 (0.045)	0.104 (0.042)	0.099 (0.040)	$\bar{\alpha}$	0.106 (0.009)	0.084 (0.007)	0.063 (0.006)
SmDLDA	0.101 (0.057)	0.091 (0.052)	0.081 (0.047)		*None*	*None*	*None*

7.4. Configuration D: First $\lfloor p/2 \rfloor$ directions are informative; elliptical covariance matrix

Here, the eigenvalues of the covariance matrix were the same as in configurations B and C, but the means differed in the first $\lfloor p/2 \rfloor$features (corresponding to the lower half of the eigenvalues) so that only the first 25 directions in the measurement space were informative for classification. In addition, the mean differences increased with the eigenvalues, as they were equal to the eigenvalues. Because half of the features were essentially noise, this configuration should benefit from regularization (and likely additional feature selection). The estimated *EER*s, estimated average regularization parameter estimates, and associated standard deviations for this

configuration appear in Table 5. As with other configurations, for Configuration D, more regularization appeared to help classification for the poorly-posed scenario.

Not surprisingly, *SmDLDA* outperformed other classifiers for all sample sizes. As we have noted in other configurations where some features do not inform class separation, reducing the influence of these noise features helped classification accuracy. In this configuration with half of the features as noise, it helps substantially. *SmDLDA* readily beat *SDLDA*, which is a diagonal classifier with regularization. As in Configuration B, when class means differ in the lower variance subspace, shrinking feature variances toward a pooled value (*SDLDA*) may increase the variance associated with the differences, reducing the "signal-to-noise" ratio.

Table 5: Configuration D: Estimated *EER*s, Average Regularization Parameter Estimates, and Corresponding Standard Deviations

	Estimated *EER*s and SDs (·)				Regularization Parameter Estimates and SDs (·)		
Discriminant Function	$n_1 = n_2 = 26$	$n_1 = n_2 = 35$	$n_1 = n_2 = 50$		$n_1 = n_2 = 26$	$n_1 = n_2 = 35$	$n_1 = n_2 = 50$
LDF	0.387 (0.072)	0.184 (0.041)	0.108 (0.021)		*None*	*None*	*None*
SM 1	0.332 (0.096)	0.172 (0.042)	0.103 (0.021)	$\bar{r}SM\ 1$	0.105 (0.517)	0.319 (1.185)	0.696 (1.834)
SM 2	0.131 (0.033)	0.101 (0.022)	0.080 (0.016)	$\bar{r}SM\ 2$	7.052 (3.287)	8.018 (2.345)	7.582 (2.534)
RSCG	0.387 (0.072)	0.184 (0.041)	0.107 (0.021)	$\bar{\lambda}$	4e-8 (6e-8)	0.002 (0.001)	0.022 (0.019)
LOH	0.382 (0.074)	0.169 (0.041)	0.099 (0.020)	$\bar{\Upsilon}LOH$	3e-5 (0.0002)	0.004 (0.005)	0.012 (0.012)
MCV	0.197 (0.099)	0.183 (0.041)	0.108 (0.021)	$\bar{\Upsilon}MC$	0.820 (0.384)	0.196 (0.290)	0.140 (0.251)
MGD	0.280 (0.129)	0.184 (0.041)	0.108 (0.021)	$\bar{\Upsilon}MG$	0.461 (0.498)	0.0003 (0.003)	0.0001 (0.001)
MLDA	0.122 (0.028)	0.100 (0.024)	0.081 (0.019)	$\bar{\rho}_1$	24.84 (1.383)	22.127 (1.110)	18.611 (0.825)
				$\bar{\rho}_2$	0.333 (0.038)	0.406 (0.031)	0.501 (0.024)
NLDA	0.124 (0.029)	0.105 (0.025)	0.089 (0.022)		*None*	*None*	*None*
SDLDA	0.294 (0.046)	0.272 (0.046)	0.245 (0.044)	$\bar{\alpha}$	0.116 (0.010)	0.092 (0.008)	0.069 (0.007)
SmDLDA	0.084 (0.029)	0.073 (0.026)	0.063 (0.022)		*None*	*None*	*None*

MLDA also performed well across all sample sizes, along with *NLDA* and *SM2*, likely due to higher regularization, similar to Configuration B. *MCV* and *MGD* performed much better than *LDF* when $N = 52$, but were equal to *LDF* when the training-sample size increased and the mean regularization parameters decreased to zero. As *N* increases, the constant *a* increases, and *b* decreases in (13), putting more weight on S^{-1} and less on I_p even before Υ_{op} is determined. As these constants do not depend on the respective optimizations in *MCV* and *MGD*, their values may not be optimal when $p/N > 1/2$, as in the second sample size case. According to Mkhadri (1995), while $a \equiv (n - p - 3)/n$ was defined to make aS^{-1} an unbiased estimate of Σ^{-1}, $b \equiv p/n$ was defined for convenience.

Incidentally, we did examine the performance of Guo et al.'s (2007) *SCRDA* classifier, which includes feature selection, in this configuration (results not shown). It performed relatively well, but not as well as *MLDA* or *NLDA* (or *SmDLDA*). The diagonal covariance matrix estimator in *SmDLDA* may have given it an advantage over *SCRDA* by sharpening weaker mean differences. See the Appendix for results using *SCRDA* when $N = p$.

8 Colon Cancer Example

We applied all regularized linear classification methods and Anderson's *LDF* to the colon cancer data from Alon et al. (1999), and analyzed in Soukup and Lee (2004) and Klaus (2013). The data set consists of 2000 genes from 62 tissue samples (40 positive samples and 22 normal samples). Soukup and Lee (2004) eliminated five supposedly contaminated samples, leaving 57 samples (37 tumor and 20 normal). We also eliminated these samples. In addition, in the absence of an explicit test set, we divided the data set into "training" and "test" sets using the same split percentages that Soukup and Lee (2004) used. Of 500 random splits of the data set, each set had 38 training samples (25 tumor samples, 13 normal) and 19 test samples (12 tumor and 7 normal). For each of the 500 training sets, prior to estimating the classifiers, we applied a variable selection procedure with the goal of choosing the most discriminating 36 genes out of the 2000 available. We chose 36 to be just shy of the total training sample size of 38, and therefore putting us in a poorly-posed situation. We used the variable selection method implemented in the CMA R package, with a "shrinkcat" criterion which is the correlationadjusted t-score from Zuber and Strimmer (2009), and the score used in Klaus' (2013) misclassification rate based variable thresholding.

Table 6 displays the misclassification proportions averaged over the 500 training/test combinations using equation (2), for each of the classification methods discussed above. We see that regularization generally improved classification performance over the non-regularized *LDF*; however, only when regularization was moderate (about midway between the identity matrix and sample covariance matrix). *MLDA* performed the best. Based on its average regularization parameters, it took a middle-of-the-road approach to regularization, with perhaps a bit more weight toward the identity matrix. Similarly, *NLDA* performed well, as it appeared to do well in simulations when the ratio p/N was closer to 1.0. However, with too much regularization toward the identity matrix, misclassification increased (e.g., *SDLDA* regularized a bit more on average than in the simulations). Despite not regularizing much on average, *MCV* and *MGD* were not as good as *MLDA* or the diagonal covariance estimators (*SDLDA* and *SmDLDA*). Also, this is the first time that the average regularization parameter for *MGD* exceeded that for *MCV*. Inconsistent with our simulations, the *SM2* classifier, using the shrinkage parameter that stabilizes successive ratios of the discriminant coefficients, performed almost as poorly as the *LDF*. However, its average regularization parameter definitely appeared more moderate than in the simulations (see the Discussion section for a possible reason why. The remaining classifiers were on par with the *LDF*. The estimated rates in Table 6 reflect the relative magnitudes of the rates in Table 4 for the smallest total training sample size. In fact, an examination of the some training sets showed that larger magnitude class mean differences were often associated with smaller feature variances. However, it was not apparent that mean differences increased with feature variances (as in Configuration D).

Table 6: Colon Cancer Data Summary of Estimated EER Rates

Discriminant Function	*LDF*	*SM1*	*SM2*	*RSCG*	*LOH*	*MCV*	*MGD*	*MLDA*	*NLDA*	*SDLDA*	*SmDLDA*	*SCRDA*	*MRTRDA*
EER	*0.305*	*0.304*	*0.26*	*0.305*	*0.3*	*0.154*	*0.217*	*0.098*	*0.098*	*0.119*	*0.118*	*0.108*	*0.111*
Average								*0.552 (0.074)*			*0.084 (0.114)*		
Regularization Parameter (SD)	*NA*	0.019 (0.088)	*3.981 (1.607)*	7e-6 (9e-6)	*0.0002 (0.0008)*	*0.455 (0.498)*	*0.814 (0.386)*	*0.360 (0.119)*	*NA*	*0.731 (0.072)*	*NA*	*0.504 (0.229)*	*0.403 (0.099)*

8.1. Guo, Hastie, and Tibshirani's (2007) SCRDA and Klaus' (2013) MRTRDA

We compared the performance of the above classifiers to Guo et al.'s (2007) Shrunken Centroids regularized classifier (*SCRDA*) and Klaus' (2013) Misclassification rate based variable thresholding (*MRTRDA*) on the colon cancer data. The *SCRDA* classifier estimates the common class covariance matrix using a convex combination of the pooled sample covariance matrix **S** and the identity matrix I_p, similar to estimators in Section 5.1. The regularization parameter α controls the extent of regularizing toward I_p. Then, to accomplish a variable selection, elements of the sample class centroids are shrunk toward their overall centroid elements to the extent that the features corresponding to the elements may become unimportant for separating classes. The extent by which features are "reduced" is controlled by a thresholding parameter, Δ. The shrunken class centroids become

$$\bar{X}_k^{*\prime} = sgn(\bar{X}_k^*)(|\bar{X}_k^*| - \Delta)_+ \tag{16}$$

Where $\bar{X}_k^* = \left(\alpha S + (1-\alpha)I_p\right)^{-1}\bar{X}_k$, $0 \le \alpha \le 1$, $\Delta > 0$, and t_+ is the positive part of t.

Guo et al. (2007) describe how to choose the optimal tuning parameters (α, Δ) by minimizing the cross-validated misclassification error rate using a "min-min" rule that selects the tuning parameters from a pre-specified grid so that the minimum cross-validated error is achieved using the smallest number of features (i.e., highest Δ value). For the colon cancer data set, ties were broken by selecting the smallest α value (this corresponds to more regularization). Choosing the smallest α value ultimately gave a lower test set error rate than choosing the largest α value.

Klaus' (2013) misclassification rate based variable thresholding selects features using their effect sizes or "feature weights", and then uses these weights within an expression for the misclassification rate. The rate depends on the number of features selected. He recommends choosing the number so that the estimated probability of error is small (around 5%) but also controlling the number of features to be as small as possible (the error rate will decrease with added features, but the reduction may be very small or zero as the effect size of the feature diminishes). For predicting the class of a test sample, Klaus (2013) uses regularized discriminant analysis via the methods in the R package sda. We follow his approach, including

his helpful flowchart for obtaining the selected features. See Klaus (2013) for details regarding the feature weights and prediction error, as well as simulation results and application to several publicly available data sets. We refer to his classifier as *MRTRDA*.

Table 6 also shows the results for both *SCRDA* and *MRTRDA*, for the same three-fold cross-validation described earlier where the training sets had 38 observations. After 500 random splits of the colon cancer data into training and test sets, the average error rates for *SCRDA* and *MRTRDA* were 0.108 and 0.111, respectively. Based on their respective average regularization parameters, a middle-of-the-road approach seemed to perform well. For *SCRDA*, the average regularization (α) was 0.084, which means that the identity matrix was preferred over the pooled sample covariance matrix, on average. In addition, an average Δ of 0.504 is not large compared to a grid of values from 0 to 3. Therefore, more features tended to be chosen. Similarly, for *MRTRDA*, more than half of the 500 sets (53%) chose all 2000 genes for the feature set based on Klaus' criterion of a 0.05 probability of misclassification. The expected error rates for these classifiers were quite acceptable compared to the other rates in Table 6. For a 10-fold cross-validation using a somewhat larger data set of 62 samples and some 2300 genes, Klaus obtained an expected error rate of 0.129. However, it is not clear whether his variable selection was performed for each training sample. If we perform variable selection once on the full data set, *MRTRDA* retains 21 features as relevant, with a misclassification error rate of 0.125 using a similar CV-fold and number of repetitions.

8.2. Yang and Wu's Regularized Complete Linear Discriminant Classifier

The rule defined in (2) can also be derived by maximizing Fisher's criterion, the ratio of the between-class to within-class (generalized) variances when the training data are transformed using a linear transformation matrix, say A, which is a vector a when there are only two classes (see, for example, Johnson and Wichern, 2002). The vector a is equal to the first eigenvector of $S^{-1}S_b$, corresponding to its only non-zero eigenvalue. Here, S_b is the between-class covariance matrix of the training data, and S is defined in Section 2. Several authors (e.g., Yang and Yang, 2003, Lu et al., 2005, and Yang et al., 2005) have pointed out that for ill-posed problems, the maximum (or supremum) of Fisher criterion is technically obtained using eigenvectors from the null space of S (that are not simultaneously in the null space of S_b) because the denominator of the ratio would be zero while the numerator would be positive. Therefore, one should use not only "regular" discriminant vectors from the range space of S for classification, but also these "irregular" discriminant vectors from the null space, as they have been found to contain important discriminatory information (Yang and Yang, 2003).

Yang and Yang (2003) proposed "complete LDA" (CLDA) to determine all discriminatory information. CLDA finds "irregular" discriminant vectors in the null space of S, and "regular" discriminant vectors in the range space of S. Yang et al. (2005) then combine ("fuse") these two kinds of vectors to make a classification decision. The "fusion coefficient" determines how much weight is placed on the regular discriminant vectors. The combination is a summed normalized distance between the test sample vector and the sample class mean vector. For two

classes, the test sample vector is classified into the class with the smaller distance. In a handwriting recognition example, Yang et al., (2005) show that for two classes, using the complete discriminant information results in better classification performance over LDA, sometimes as much as 27% better.

Yang and Wu (2014) later proposed *regularized* complete LDA (*RCLDA*). They use a regularized criterion to derive both the regular and irregular discriminant vectors. The same regularization parameter ($\sigma > 0$) is used for both criteria. In our implementation of *RCLDA* for the colon cancer data, the regularization parameter σ gives a larger weight to the *regular* discriminant vectors when it is smaller, and alternatively increases the weight put on *irregular* vectors when it increases. Similar to Yang and Wu (2014), we use the summed normalized distance between a test vector and a class mean vector to make a classification decision.

Because our paper deals with poorly-posed problems, theoretically there are only regular discriminant vectors. However, we have found that even poorly-posed problems can benefit from taking advantage of irregular discriminant vectors, where a zero-valued eigenvalue can be defined as less than a threshold such as $1.e^{-06}$. In fact, with such a threshold, the misclassification error rate for the colon cancer data was as small as that for *MLDA* and *NLDA* provided that the regularization parameter (σ) was between about 0.5 and 1.0. Putting a large amount of weight on the regular discriminant vector (by setting σ closer to zero) increased the error rate. Putting most of the weight on the ”regular” discriminant vector (e.g., setting σ equal to the threshold above) reproduced the error rate of the *LDF*, as expected.

9 Discussion

We have compared the classification performance of ten regularized *LDF*s, along with Anderson’s *LDF*. Our simulation results indicate that when regularization was expected to improve classification performance, it often did so. No single regularization classifier uniformly outperformed the others, although the classifiers developed and tested for ill-posed problems (i.e., when $p \geq N$) such as *MLDA* and *NLDA*, performed consistently well across all configurations, especially with the smaller sample sizes. We did not get an opportunity to test Yang and Wu’s (2014) *RCLDA* on simulated data. However, based on its results on the colon cancer data, we suspect it would perform well. The diagonal covariance classifiers (*SDLDA* and *SmDLDA*) both performed well in Configuration C where class means differed in the high variance subspace, and *SmDLDA* performed well in Configuration B, where means differed in the low-variance subspace. Recall that *SmDLDA* adjusted the class mean estimates using a James-Stein estimator. *SDLDA* did not adjust the class means, and did not perform well in Configurations B and D, where means differed in the low-variance subspace.

SM2 performed consistently well across configurations, despite the shape of the true covariance matrix, even beating *MLDA* and *NLDA* in Configuration B. In that Configuration, *MLDA* and *NLDA* suffered in the largest total sample size, performing worse than *LDF* (as did *SDLDA*). *SM2* did not suffer in the configurations we tested where $p < N$ *SM1*, on the other hand, rarely performed well in the smaller total sample size scenarios. Also, the

asymptotic-based regularization classifiers (*LOH* and *RSCG*) never performed much better than the *LDF*, and are thus not recommended.

In the Appendix, we test all classifiers (including *SCRDA*) using an ill-posed problem where $N = p = 50$. In these cases, *MLDA* performed consistently well, but was only the best performer in Configuration C. In the examples in Xu et al. (2009), *MLDA* often beat *NLDA* and *SCRDA*. That was true in our configurations except for Configuration A where *SCRDA* outperformed all regularizers. Also, as with the main text, when $N = p$, *SmDLDA* had the best performance by far in Configurations B and D, when class mean differences were in the lower variance subspace. However, the good performance of *SM2* waned when $N = p$. We suspect it was due to the way we assessed stability of the ratios of sample discriminants by starting at the lower end of *R* and proceeding until a stable r_j was found. This caused the chosen r to be too small.

Because of the method for breaking ties, the *SM1* and *MCV* were influenced to choose lower regularization values when the cross-validated risk was tied. This phenomenon did affect their performance in most configurations, more so for *SM1* than *MCV*. In the Appendix, we present the same results but choose the highest level of regularization to break ties. In this case, we notice improved performance for Configurations A and C for *SM1* and *MCV* mostly in the smaller sample sizes. We might prudently choose to break ties in the direction of expected benefit of more or less regularization if it can be anticipated.

We cannot ignore that perhaps one reason for the good performance of the diagonal classifiers is that the true covariance matrix was diagonal. However, the diagonal classifiers did not always outperform the other classifiers, and were among the worst performers in Configurations A (*SmDLDA*), B, and D (*SDLDA*). Also, when $N < p$ or $N \sim p$, S may not resemble Σ. We intend to examine more general covariance structures in future research.

With regard to the shrunken covariance matrix, the convex combination of **S** and the identity matrix, and the diagonal regularizers in Sections 4, 5, and 6, respectively, no consistent pattern of one group notably outperformed the other for a given configuration, though the diagonal regularizers did very well in Configuration C. In general, overall, the newer developed methods performed better than the older methods, but *SM2* (Smidt, 1976b) was quite competitive with the newer methods.

Finally, despite using 100,000 training samples, we noticed considerable variability in the choice of regularization parameters for almost every precision matrix estimator. However, for large ranges of the regularization parameter, we might expect the classification rule to remain the same so that the misclassification rate does not vary substantially.

For the colon cancer data, Soukup and Lee's classifier was a stepwise cross-validated disciminant analysis where two genes were found to be most discriminatory while also defining a parsimonious model. They correctly classified all 19 test samples using a classifier built from the remaining 38 samples. While 100% classification is quite remarkable, it occurred on a

particular data set. They did not perform simulations to see how their classifier performed under specific configurations. We also note that they allowed different covariance matrices across classes (i.e., quadratic discriminant analysis). We assumed a common covariance matrix, and our results may have suffered. However, several authors have noted the advantage of the *LDF* even without distinct covariance matrices because of the reduced number of parameters to estimate (see, for example, Dudoit, Fridyland, and Speed (2002) and Marks and Dunn (1974)).

9.1. General Guidelines

Based on the results from our simulation study, we can offer the following general guidelines for choosing a regularized classifier under the conditions we studied.

1. Configurations B, C, and D all had the same true covariance matrix;they differed only in terms of class mean differences, whether there were mean differences in the lower variance subspace or higher variance subspace. Regularization seemed to have the greatest gain compared to *LDF* when class means differed in the high variance subspace. This is not surprising because if the largest sample eigenvalues are adjusted lower, the detection of differences could increase. On the other hand, when means differ in a low variance subspace, it might be best to regularize the means and not the eigenvalues (e.g., *SmDLDA*) in order to not dilute mean differences.
2. In Configuration A, all true eigenvalues were the same. However, this configuration had only a single feature differ between classes. With only a few features differing between classes, mean adjustment plus covariance regularization (e.g., *SCRDA*) may be most appropriate in order to detect the differences. Although *SmDLDA* shrinks all feature means toward their overall means, it does not perform further regularization of the covariance matrix, whereas *SCRDA* chooses the amount of mean shrinkage based on cross-validation misclassification error, in addition to regularizing the covariance.
3. Many of the less computationally demanding methods performed very well. A diagonal regularized classifier can potentially offer a big improvement in classification accuracy without much effort. Also, merely replacing the smallest sample eigenvalues with their mean (*NLDA*) can achieve very good classification results without even thinking about estimating a regularization parameter. A cross-validation approach may not be necessary, especially if the number of training samples is large requiring more time for a leave-one-out technique.

4. Whereas in the main text, we broke ties between potential parameter values by choosing the value that offered less regularization, it is clear from the Appendix that breaking ties toward more regularization would have been better. Consequently, we recommend breaking ties toward more regularization.
5. When $p = N$, many of the methods we examined can also be applied, as many were developed for an ill-posed situation. However, some methods would have to use a pseudo-inverse estimate for S^{-1} (e.g., *MCV* and *MGD*).
6. Asymptotically derived regularization parameters may not have a placein poorly-posed problems when $p \sim N$. The regularization parameter estimates appear to have been derived under an asymptotic assumption of *N* increasing with a fixed *p*. Therefore, their usefulness in the types of poorly-posed problems we examined may be limited.

9.2. Other Types of Regularization

Almost all of the methods we have discussed still have to choose the regularization parameter or fusion coefficient, often by iteration. Sharma and Paliwal (2015) choose the regularization parameter "deterministically" by maximizing a modified Fisher's criterion. The criterion is modified by replacing S with $S + \alpha I_p$ such that $S + \alpha I_p$ is non-singular. Using Lagrange's multiplier method, and setting the Lagrange multiplier (λ) equal to the largest eigenvalue of $S^+ S_b$ where S^+ is a pseudoinverse of S, and S_b is the between-class covariance matrix, they solve for α as the largest eigenvalue of $(1/\lambda)S_b - S$. It is easy to see that when S is technically invertible, the method is equivalent to *LDF* because $(1/\lambda)S_b = S$, and so $\alpha = 0$. Therefore, we did not examine this method with poorly-posed problems.

Aerts and Wilms (2017) regularize the pooled sample covariance matrix and sample class means by using "robust" estimates of each. That is, they use robust moment estimators for the standard deviations of the features, correlation coefficients of each pair of features, and the class centers. Presumably one could use any robust estimators, but the particular ones used by the authors are common choices. These robust estimates are then used in several regularized classifiers such as the graphical lasso (Friedman, et al., 2008), the joint graphical lasso (Price, et al., 2015), and *RDA* (Friedman, 1989). In their "contaminated" schemes, where a small percentage of outlying values are added to components of a set of normally distributed training vectors, the robust versions of each method show considerably better classification performance than the non-robust versions, especially as the percentage of contamination increases. In their uncontaminated schemes, the robust versions were mostly on par with the non-robust versions. Further research may examine whether this method could work well when class means differ in a high variance subspace, such as our Configuration C.

10 Appendix

In this Appendix we explore the sensitivity of classification results under two conditions. In the first condition, we examine classification results when the method for breaking ties for choosing the regularization parameter is toward higher regularization values instead of lower values. In the second condition, we compare the classifiers under an ill-posed problem.

10.1. Breaking ties in regularization parameters using the maximum instead of the minimum

For this sensitivity analysis, we re-estimated classification performance for the two estimators that needed to break ties among regularization parameters that achieved the minimum cross-validated risk (i.e., *SM1* and *MCV*). The results will be identical to the tables in the main text, except for the *SM1* and *MCV* classifiers.

10.1.1. Configuration A: Means differ in one orthogonal direction; equal eigenvalues

For this configuration, because more regularization is preferred, *SM1* and *MCV* improved classification performance. The estimated *EER*s, average regularization parameter estimates, and associated standard deviations for this configuration appear in Table A1.

Table A1: Configuration A: Estimated EERs, Average Regularization Parameter Estimates, and Corresponding Standard Deviations

	Estimated *EER*s and SDs (·)				Regularization Parameter Estimates and SDs (·)		
Discriminant Function	$n_1 = n_2 = 26$	$n_1 = n_2 = 35$	$n_1 = n_2 = 50$		$n_1 = n_2 = 26$	$n_1 = n_2 = 35$	$n_1 = n_2 = 50$
LDF	0.453 (0.045)	0.350 (0.035)	0.290 (0.026)		*None*	*None*	*None*
SM 1	0.353 (0.091)	0.279 (0.045)	0.254 (0.031)	$\bar{r}SM$ 1	0.910 (2.082)	1.278 (2.300)	0.990 (2.041)
SM 2	0.253 (0.021)	0.235 (0.016)	0.218 (0.012)	$\bar{r}SM$ 2	7.129 (3.279)	8.814 (1.868)	8.958 (1.732)
RSCG	0.453 (0.045)	0.348 (0.035)	0.285 (0.024)	$\bar{\lambda}$	1e-8 (1e-8)	0.001 (0.0003)	0.017 (0.004)
LOH	0.451 (0.046)	0.335 (0.039)	0.268 (0.027)	$\bar{\Upsilon}LOH$	0.0001 (0.001)	0.021 (0.023)	0.105 (0.073)
MCV	0.270 (0.065)	0.344 (0.038)	0.290 (0.025)	$\bar{\Upsilon}MC$	0.901 (0.299)	0.411 (0.328)	0.332 (0.294)
MGD	0.370 (0.106)	0.350 (0.035)	0.290 (0.026)	$\bar{\Upsilon}MG$	0.408 (0.492)	0.0003 (0.003)	0.0001 (0.001)
MLDA	0.250 (0.020)	0.234 (0.016)	0.218 (0.012)	$\bar{\rho}_1$	0.991 (0.032)	0.995 (0.027)	0.997 (0.022)
				$\bar{\rho}_2$	0.009 (0.018)	0.005 (0.013)	0.003 (0.009)
NLDA	0.259 (0.021)	0.242 (0.017)	0.225 (0.013)		*None*	*None*	*None*
SDLDA	0.220 (0.022)	0.207 (0.016)	0.195 (0.011)	$\bar{\alpha}$	0.456 (0.049)	0.438 (0.048)	0.412 (0.046)
SmDLDA	0.331 (0.018)	0.312 (0.017)	0.288 (0.016)		*None*	*None*	*None*

10.1.2. Configuration B: Means differ in last $p-1$ features; elliptical covariance matrix

Configuration B also tended to favor regularization in the main text. Breaking ties based on larger values of the regularization parameter estimates helped the classification performance of *SM1* and *MCV*. The estimated *EER*s, estimated average regularization parameter estimates, and associated standard deviations for this configuration appear in Table A2.

Table A2: Configuration B: Estimated EERs, Average Regularization Parameter Estimates, and Corresponding Standard Deviations

	Estimated *EER*s and SDs (·)				Regularization Parameter Estimates and SDs (·)		
Discriminant Function	$n_1=n_2=26$	$n_1=n_2=35$	$n_1=n_2=50$		$n_1=n_2=26$	$n_1=n_2=35$	$n_1=n_2=50$
LDF	0.387 (0.072)	0.185 (0.041)	0.108 (0.021)		*None*	*None*	*None*
SM 1	0.318 (0.101)	0.155 (0.036)	0.105 (0.019)	$\bar{r}SM$ 1	0.285 (1.020)	1.968 (3.190)	4.454 (4.231)
SM 2	0.171 (0.033)	0.136 (0.024)	0.106 (0.018)	$\bar{r}SM$ 2	7.079 (3.282)	8.267 (2.208)	7.713 (2.424)
RSCG	0.387 (0.072)	0.184 (0.041)	0.108 (0.021)	$\bar{\lambda}$	4e-8 (6e-8)	0.002 (0.001)	0.020 (0.016)
LOH	0.382 (0.074)	0.174 (0.040)	0.104 (0.020)	$\bar{\Upsilon}LOH$	3e-5 (0.0002)	0.003 (0.004)	0.006 (0.006)
MCV	0.235 (0.059)	0.183 (0.041)	0.108 (0.021)	$\bar{\Upsilon}MC$	0.918 (0.274)	0.372 (0.332)	0.315 (0.317)
MGD	0.311 (0.101)	0.186 (0.041)	0.108 (0.021)	$\bar{\Upsilon}MG$	0.457 (0.498)	0.0003 (0.003)	0.0001 (0.001)
MLDA	0.182 (0.030)	0.156 (0.025)	0.127 (0.021)	$\bar{\rho}_1$	24.077 (1.351)	21.441 (1.069)	18.046 (0.788)
				$\bar{\rho}_2$	0.354 (0.037)	0.424 (0.030)	0.516 (0.022)
NLDA	0.189 (0.031)	0.169 (0.026)	0.148 (0.023)		*None*	*None*	*None*
SDLDA	0.341 (0.034)	0.325 (0.033)	0.306 (0.031)	$\bar{\alpha}$	0.119 (0.009)	0.095 (0.008)	0.071 (0.006)
SmDLDA	0.088 (0.028)	0.078 (0.025)	0.069 (0.021)		*None*	*None*	*None*

10.1.3. Configuration C: Means differ in last $p-1$ features; elliptical covariance matrix

For Configuration C, regularization was expected to help classification. When we broke ties using the maximum regularization, *SM1* and *MCV* improved, but for *MCV*, improvement was only for the smallest sample size. The estimated *EER*s, estimated average regularization parameter estimates, and associated standard deviations for this configuration appear in Table A3.

Table A3: Configuration C: Estimated EERs, Average Regularization Parameter Estimates, and Corresponding Standard Deviations

	Estimated *EER*s and SDs (·)				Regularization Parameter Estimates and SDs (·)		
Discriminant Function	$n_1=n_2=26$	$n_1=n_2=35$	$n_1=n_2=50$		$n_1=n_2=26$	$n_1=n_2=35$	$n_1=n_2=50$
LDF	*0.387 (0.072)*	*0.184 (0.041)*	0.108 (0.021)		*None*	*None*	*None*
SM 1	*0.312 (0.107)*	*0.150 (0.041)*	0.096 (0.022)	$\bar{r}SM$ 1	0.281 (0.997)	1.283 (2.498)	2.413 (3.426)
SM 2	*0.122 (0.040)*	*0.097 (0.029)*	0.079 (0.023)	$\bar{r}SM$ 2	7.028 (3.284)	7.937 (2.381)	7.728 (2.503)
RSCG	*0.387 (0.072)*	*0.184 (0.041)*	0.108 (0.021)	$\bar{\lambda}$	6e-8 (7e-8)	0.001 (0.0006)	0.010 (0.004)
LOH	*0.381 (0.074)*	*0.170 (0.041)*	0.099 (0.020)	$\bar{\Upsilon}LOH$	3e-5 (0.0002)	0.004 (0.005)	0.011 (0.012)
MCV	*0.141 (0.126)*	*0.182 (0.041)*	0.108 (0.021)	$\bar{\Upsilon}MC$	0.820 (0.384)	0.235 (0.263)	0.191 (0.240)
MGD	*0.247 (0.162)*	*0.184 (0.041)*	0.108 (0.021)	$\bar{\Upsilon}MG$	0.460 (0.498)	0.0003 (0.003)	0.0001 (0.001)
MLDA	*0.082 (0.039)*	*0.077 (0.035)*	0.071 (0.031)	$\bar{\rho}_1$	31.678 (3.523)	28.221 (3.249)	23.792 (2.607)
				$\bar{\rho}_2$	0.149 (0.097)	0.242 (0.090)	0.361 (0.072)
NLDA	*0.091 (0.039)*	*0.085 (0.036)*	0.077 (0.034)		*None*	*None*	*None*
SDLDA	*0.109 (0.045)*	*0.104 (0.042)*	0.099 (0.040)	$\bar{\alpha}$	0.106 (0.009)	0.084 (0.007)	0.063 (0.006)
SmDLDA	*0.101 (0.057)*	*0.091 (0.052)*	0.081 (0.047)		*None*	*None*	*None*

10.1.4. Configuration D: First $[p/2]$ directions are informative; elliptical covariance matrix

For Configuration D, regularization generally helped classification. For *SM1*, more regularization helped all training-sample sizes, but as with Configuration C, for *MCV*, it only helped the smallest sample size. The estimated *EER*s, estimated average regularization parameter estimates, and associated standard deviations for this configuration appear in Table A4.

Table A4: Configuration D: Estimated EERs, Average Regularization Parameter Estimates, and Corresponding Standard Deviations

	Estimated *EER*s and SDs (·)				Regularization Parameter Estimates and SDs (·)		
Discriminant Function	$n_1 = n_2 = 26$	$n_1 = n_2 = 35$	$n_1 = n_2 = 50$		$n_1 = n_2 = 26$	$n_1 = n_2 = 35$	$n_1 = n_2 = 50$
LDF	0.387 (0.072)	0.184 (0.041)	0.108 (0.021)		*None*	*None*	*None*
SM 1	0.312 (0.107)	0.149 (0.040)	0.094 (0.020)	$\bar{r}SM\ 1$	0.286 (1.013)	1.162 (2.315)	2.256 (3.296)
SM 2	0.131 (0.033)	0.101 (0.022)	0.080 (0.016)	$\bar{r}SM\ 2$	7.052 (3.287)	8.018 (2.345)	7.582 (2.534)
RSCG	0.387 (0.072)	0.184 (0.041)	0.107 (0.021)	$\bar{\lambda}$	4e-8 (6e-8)	0.002 (0.001)	0.022 (0.019)
LOH	0.382 (0.074)	0.169 (0.041)	0.099 (0.020)	$\bar{\gamma}LOH$	3e-5 (0.0002)	0.004 (0.005)	0.012 (0.012)
MCV	0.177 (0.077)	0.182 (0.041)	0.108 (0.021)	$\bar{\gamma}MC$	0.911 (0.284)	0.343 (0.317)	0.287 (0.302)
MGD	0.280 (0.129)	0.184 (0.041)	0.108 (0.021)	$\bar{\gamma}MG$	0.461 (0.498)	0.0003 (0.003)	0.0001 (0.001)
MLDA	0.122 (0.028)	0.100 (0.024)	0.081 (0.019)	$\bar{\rho}_1$	24.84 (1.383)	22.127 (1.110)	18.611 (0.825)
				$\bar{\rho}_2$	0.333 (0.038)	0.406 (0.031)	0.501 (0.024)
NLDA	0.124 (0.029)	0.105 (0.025)	0.089 (0.022)		*None*	*None*	*None*
SDLDA	0.294 (0.046)	0.272 (0.046)	0.245 (0.044)	$\bar{\alpha}$	0.116 (0.010)	0.092 (0.008)	0.069 (0.007)
SmDLDA	0.084 (0.029)	0.073 (0.026)	0.063 (0.022)		*None*	*None*	*None*

10.2. Ill-posed Problem where $N = p$

For this sensitivity analysis, we estimated classification performance for all classifiers, including those specifically developed and tested for ill-posed problems (i.e., *MLDA*, *NLDA* and *SCRDA*), for the case with $N = 50$ and $p = 50$. The results are shown in the following tables. For classifiers that needed to invert the sample covariance matrix (including *LDF*), we substituted the Moore-Penrose pseudo-inverse. For *SCRDA* we used the default options in the authors' R function mentioned in their paper. That is, we used the "min-min" rule that selects the (α, Δ) pair with the smallest number of features, when there were ties for the minimum cross-validation error. We again used a grid of values between 0 and 3 for candidate Δ values.

10.2.1. Configuration A: Means differ in one orthogonal direction; equal eigenvalues

The estimated *EER*s, average regularization parameter estimates, and associated standard deviations for this configuration appear in Table B1. *SCRDA* outperforms the other classifiers. The average α is not especially small, suggesting that despite the true covariance matrix being the identity, the observed covariance matrix may be much different, reducing regularization toward I_p. Based on its average Δ value, *SCRDA* retained fewer features on average in Configuration A than B. Note that for Configuration A, all but one feature is noise. Table B1 shows that estimators that chose relatively larger amounts of regularization on average, specifically *SM1max*, *MCVmax* and *MLDA*, obtained good classification results. The two

estimators that chose regularization parameters using asymptotic considerations (*LOH* and *RSCG*) did not perform well, choosing little regularization.

Many results in Table B1 are similar to those in Table 2 for the $N = 52$ case. However, *SM2* performed much worse due to its small amount of regularization on average. The algorithm for choosing *r* in *SM2* works through the grid *R*, starting with the smallest value of 0, and stopping when it finds an *r* that stabilizes the successive ratios of (6). With an ill-posed problem, the smallest eigenvalue of S is essentially zero. Therefore, stabilization could have occurred more quickly, leading to smaller values of *r* on average. In fact, all four configurations result in smaller average *r* values for *SM2* in the ill-posed problem than in the poorly-posed problems.

Table B1: Configuration A: Estimated EERs, Average Regularization Parameter Estimates, and Corresponding Standard Deviations

Discriminant Function	Estimated *EER*s and SDs (·) $n_1 = n_2 = 25$		Regularization Parameter Estimates and SDs (·) $n_1 = n_2 = 25$
LDF	*0.432 (0.044)*		*None*
SM 1*min*	*0.305 (0.092)*	$\bar{r}SM\ 1$	*2.475 (2.655)*
SM 1*max*	*0.265 (0.031)*	$\bar{r}SM\ 1$	*4.014 (3.405)*
SM 2	*0.494 (0.065)*	$\bar{r}SM\ 2$	*0.024 (0.304)*
RSCG	*0.432 (0.044)*	$\bar{\lambda}$	*0.000 (0.000)*
LOH	*0.440 (0.046)*	$\bar{\Upsilon}LOH$	*0.001 (0.003)*
MCVmin	*0.286 (0.074)*	$\bar{\Upsilon}MC$	*0.812 (0.391)*
MCVmax	*0.270 (0.058)*	$\bar{\Upsilon}MC$	*0.901 (0.298)*
MGD	*0.348 (0.096)*	$\bar{\Upsilon}MG$	*0.467 (0.499)*
MLDA	*0.252 (0.020)*	$\bar{\rho}_1$	*0.990 (0.033)*
		$\bar{\rho}_2$	*0.010 (0.019)*
NLDA	*0.262 (0.022)*		*None*
SCRDA	*0.200 (0.087)*	$\bar{\alpha}$	*0.218 (0.302)*
		Δ	*0.605 (0.446)*
SDLDA	*0.222 (0.023)*	$\bar{\alpha}$	*0.458 (0.050)*
SmDLDA	*0.334 (0.018)*		*None*

10.2.2. Configuration B: Means differ in last $p - 1$ features; elliptical covariance matrix

We did not expect this configuration to favor regularization because dispersion regularizing matrices tend to equalize the unequal eigenvalues, making the regularized covariance matrix more spherical. However, as in the poorly-posed problems, when $p = N$, some classifiers with high amounts of regularization performed quite well. We provide the estimated *EER*s, the corresponding average regularization parameters, and their respective standard deviations in Table B2. *SmDLDA* performed the best by far using the mean-adjusted diagonal covariance estimator, as it did in Table 3 in the main text. *NLDA* and *MLDA* also performed well, as might be expected due to the ill-posed problem. *SM2* improved over Configuration A, with a higher amount of regularization, though still lower on average than in Table 3, in the main text. *SM1* improved over its performance in Table 3 for the poorly-posed problems. With an ill-posed situation, where the smallest sample eigenvalue is essentially zero, the value of *r* that minimizes the cross-validation error rate may be much higher than in a poorly-posed situation. This applies to Configurations C and D, as well.

SCRDA did not do as well (relatively) in this configuration. It chose to retain a larger number of features on average than in other configurations. Configurations B and C have the fewest features that are pure noise with no discriminatory ability. However, in Configuration C, the average Δ is much higher at 0.667 (relatively smaller number of features). In Configuration B, class mean differences are in the lower variance subspace. Therefore, less shrinkage of features toward their overall mean may be needed, leading to a lower average value of Δ.

Table B2: Configuration B: Estimated EERs, Average Regularization Parameter Estimates, and Corresponding Standard Deviations

Discriminant Function	Estimated *EER*s and SDs (·) $n_1 = n_2 = 25$		Regularization Parameter Estimates and SDs (·) $n_1 = n_2 = 25$
LDF	0.348 (0.067)		*None*
$SM\ 1min$	0.210 (0.077)	$\bar{r}SM\ 1$	2.665 (2.597)
$SM\ 1max$	0.198 (0.062)	$\bar{r}SM\ 1$	3.707 (3.245)
$SM\ 2$	0.254 (0.145)	$\bar{r}SM\ 2$	5.112 (4.305)
RSCG	0.348 (0.067)	$\bar{\lambda}$	0.000 (0.000)
LOH	0.358 (0.070)	$\bar{\gamma}LOH$	0.0002(0.001)
MCVmin	0.244 (0.061)	$\bar{\gamma}MC$	0.834 (0.368)
MCVmax	0.235 (0.050)	$\bar{\gamma}MC$	0.919 (0.273)
MGD	0.288 (0.081)	$\bar{\gamma}MG$	0.486 (0.500)
MLDA	0.187 (0.030)	$\bar{\rho}_1$	24.372(1.377)
		$\bar{\rho}_2$	0.346 (0.038)
NLDA	0.193 (0.031)		*None*
SCRDA	0.230 (0.083)	$\bar{\alpha}$	0.228 (0.220)
		Δ	0.382 (0.412)
SDLDA	0.342 (0.034)	$\bar{\alpha}$	0.122 (0.010)
SmDLDA	0.089 (0.029)		*None*

10.2.3. Configuration C: Means differ in last p−1 features; elliptical covariance matrix

With this configuration, high regularization from *MLDA* provided the lowest *EER* of 0.084, similar to Table 4 in the main text. Table B3 gives the results from this configuration. Classifiers that used higher amounts of regularization performed better than those that did not. *SCRDA* regularized quite a bit toward the identity covariance matrix (smaller $\bar{\alpha}$), and chose a lower number of features on average than in Configuration B (higher $\bar{\Delta}$). *LOH* suffered from a very small average amount of regularization, performing worse than the *LDF* unlike in the poorly-posed problems in the main text. The diagonal classifiers performed well, similar to Table 4 in the main text.

Table B3: Configuration C: Estimated *EER*s, Average Regularization Parameter Estimates, and Corresponding Standard Deviations

Discriminant Function	Estimated *EER*s and SDs (·) $n_1 = n_2 = 25$		Regularization Parameter Estimates and SDs (·) $n_1 = n_2 = 25$
LDF	0.340 (0.069)		*None*
$SM\ 1min$	0.179 (0.089)	$\bar{r}SM\ 1$	2.652 (2.594)
$SM\ 1max$	0.162 (0.075)	$\bar{r}SM\ 1$	3.707 (3.245)
$SM\ 2$	0.218 (0.167)	$\bar{r}SM\ 2$	5.099 (4.295)
RSCG	0.340 (0.067)	$\bar{\lambda}$	0.000 (0.000)
LOH	0.355 (0.071)	$\bar{\gamma}LOH$	0.0002 (0.001)
MCVmin	0.161 (0.126)	$\bar{\gamma}MC$	0.708 (0.455)
MCVmax	0.132 (0.108)	$\bar{\gamma}MC$	0.821 (0.383)
MGD	0.216 (0.139)	$\bar{\gamma}MG$	0.488 (0.500)
MLDA	0.084 (0.040)	$\bar{\rho}_1$	32.069 (3.563)
		$\bar{\rho}_2$	0.139 (0.097)
NLDA	0.091 (0.040)		*None*
SCRDA	0.125 (0.085)	$\bar{\alpha}$	0.053 (0.136)
		Δ	0.667 (0.645)
SDLDA	0.107 (0.042)	$\bar{\alpha}$	0.109 (0.009)
SmDLDA	0.100 (0.054)		*None*

10.2.4. Configuration D: First [p/2] directions are informative; elliptical covariance matrix

In this configuration, higher regularization tended to perform better than lower regularization. As with Table 5, *SmDLDA* performed the best, with *MLDA* and *NLDA* also performing well. The average regularization parameter for *SCRDA* was small in magnitude, regularizing more toward the identity matrix. Its average shrinkage threshold was somewhat higher than in the other configurations (0.760), leading to fewer features retained for classification. This makes sense as only 25 of the 50 features had non-zero true mean differences between classes. These choices resulted in more regularization, and thus a lower *EER* than some of the other classifiers. Table B4 gives the results from this configuration.

Table B3: Configuration C: Estimated *EERs*, Average Regularization Parameter Estimates, and Corresponding Standard Deviations

Discriminant Function	Estimated *EER*s and SDs (·) $n_1 = n_2 = 25$		Regularization Parameter Estimates and SDs (·) $n_1 = n_2 = 25$
LDF	*0.340 (0.069)*		*None*
$SM\ 1min$	*0.182 (0.085)*	$\bar{r}SM\ 1$	*2.658 (2.585)*
$SM\ 1max$	*0.167 (0.070)*	$\bar{r}SM\ 1$	*3.700 (3.231)*
$SM\ 2$	*0.224 (0.161)*	$\bar{r}SM\ 2$	*5.114 (4.293)*
RSCG	*0.340 (0.068)*	$\bar{\lambda}$	*0.000 (0.000)*
LOH	*0.355 (0.072)*	$\bar{\gamma}LOH$	*0.0002 (0.001)*
MCVmin	*0.192 (0.082)*	$\bar{\gamma}MC$	*0.822 (0.382)*
MCVmax	*0.175 (0.065)*	$\bar{\gamma}MC$	*0.911 (0.285)*
MGD	*0.252 (0.106)*	$\bar{\gamma}MG$	*0.487 (0.500)*
MLDA	*0.126 (0.030)*	$\bar{\rho}_1$	*25.161 (1.421)*
		$\bar{\rho}_2$	*0.325 (0.039)*
NLDA	*0.128 (0.031)*		*None*
SCRDA	*0.178 (0.063)*	$\bar{\alpha}$	*0.122 (0.149)*
		Δ	*0.760 (0.608)*
SDLDA	*0.297 (0.044)*	$\bar{\alpha}$	*0.119 (0.010)*
SmDLDA	*0.085 (0.029)*		*None*

References

[1] Aerts, S., and Wilms, I. (2017). Cellwise robust regularized discriminant analysis. Statistical Analysis and Data Mining: The ASA Data Science Journal 10:436–447.

[2] Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D. and Levine, A. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS 96:6745–6750.

[3] Anderson, T. W. (1951). Classification by multivariate analysis. Psychometrika 16:31–50.

[4] Bai, Z. (1999). Methodologies in spectral analysis of large dimensional random matrices. Statistica Sinica 9:611–677.

[5] Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. Statistica Sinica 6:311–329.

[6] Bai, Z. and Silverstein, J. (2004). CLT for Linear Spectral Statistics of Large-Dimensional Sample Covariance Matrices. The Annals of Probability 32:553–605.

[7] Belhumeur, P., Hespanha, J., and Kriengman, D. (1997). Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19:711–720.

[8] Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). Australian and New England Journal of Statistics 43:147–199.

[9] DiPillo, P. J. (1976). The application of bias to discriminant analysis. Communications in Statistics: Theory and Models 5:843–854.

[10] DiPillo, P. J. (1979). Biased discriminant analysis: evaluation of the optimum probability of misclassification. Communications in Statistics: Theory and Models 8:1447–1457.

[11] Dudoit, P. J., Fridyland, J., and Speed, T (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Assocation 97:77–87.

[12] Friedman, J. H. (1989). Regularized discriminant analysis. Journal of the American Statistical Association 84:165–175.

[13] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9:432–441.

[14] Greene, T. and Rayens, W. S. (1989). Partially pooled covariance matrix estimation in discriminant analysis. Communications in Statistics: Theory and Models 18:3679–3702.

[15] Guo, Y., Hastie, T., and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. Biostatistics 8:86–100.

[16] Hastie, T. and Tibshirani, R. (2001). Elements of Statistical Learning. Springer, New York, NY.

[17] John, S. (1961). Errors in discrimination. Annals of Mathematical Statistics 32:1125–1144.

[18] Johnson, R. and Wichern, D. (2002). Applied Multivariate Statistical Analysis. Prentice Hall, New Jersey.

[19] Klaus, B. (2013). Effect size estimation and misclassification rate based variable selection in linear discriminant analysis. Journal of Data Science 11:537–558.

[20] Koolaard, P., Ganesalingam, S., and Lawoko, C. (1998). Comparison of regularized discriminant analysis with standard discrimination methods. Computational Statistics 13:495–509.

[21] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large dimensional covariance matrices. Journal of Multivariate Analysis 88:365– 411.

[22] Loh, W. (1995). On linear discriminant analysis with adaptive ridge classification rules. Journal of Multivariate Analysis 53:264–278.

[23] Lu, J., Plataniotis, A., and Venetsanopoulos (2005). Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. Pattern Recognition Letters 26:181–191.

[24] Marks, S. and Dunn, O. (1974). Discriminant functions when covariance matrices are unequal. Journal of the American Statistical Association 69:555–559.

[25] Mkhadri, A. (1995). Shrinkage parameter for the modified linear discriminant analysis. Pattern Recognition Letters 16:267–275.

[26] Pang, H., Tong, T., and Zhao, H. (2009). Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. Biometrics 65:1021–1029.

[27] Peck, R. and Van Ness, J. (1982). The use of shrinkage estimators in linear discriminant analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 4:530–536.

[28] Price, B., Geyer, C., and Rothman, A. (2015). Ridge fusion in statistical learning. Journal of Computational and Graphical Statistics 24:439–454.

[29] Ramey, J., Stein, C., Young, P., and Young, D. (2017). High Dimensional Regularized Discriminant Analysis. Available at https://arxiv.org/abs/1602.01182.

[30] Raudys, S. and Skurihina, S. M. (1994). Small sample properties of ridge estimates of the covariance matrix in statistical and neural net classification. In: Tiit, E. M., Kollo, T., Niemi, H. (Eds)., New Trends in Probability and Statistics: Multivariate Statistics and Matrices in Statistics. Vol. 3 TEV, Vilnius, & VSP, Utrecht, pp. 237–245.

[31] Raudys, S., Skurihina, M., Cibas, T., and Gallinari, P. (1995). Optimal regularization of neural networks and ridge estimates of the covariance matrix in statistical classification. Pattern Recognition and Image Analysis 5:633–650.

[32] Rayens, W. S. and Greene, T. (1991). Covariance pooling and stabilization for classification. Computational Statistics and Data Analysis 11:17–42.

[33] Seber, G. F. (1984). Multivariate Observations. Wiley, New York.

[34] Sharma, A. and Paliwal, K. (2015). A deterministic approach to regularized linear discriminant analysis. Neurocomputing 151:207–214.

[35] Smidt, R. K. and McDonald, L. L. (1976a). Ridge discriminant analysis. Tech. Rep. 108, University of Wyoming Department of Statistics.

[36] Smidt, R. K. and McDonald, L. L. (1976b). Ridge estimation of the inverse of the inverse of a covariance matrix, Tech. Rep. 106, University of Wyoming Department of Statistics.

[37] Soukup, M., and Lee, J. (2004). Developing optimal prediction models for cancer classification using gene expression data. Journal of Bioinformatics and Computational Biology 1:681–694.

[38] Thomaz, C. and Gillies, D. (2005). A maximum uncertainty lda-based approach for limited sample size problems with application to face recognition. In: 18th Brazilian Symposium on Computer Graphics and Image Processing SIBGRAPH 2005 89–96.

[39] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statistical Science 18:104–117.

[40] Tong, T. and Wang, Y. (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis. Journal of the American Statistical Association 102:113–122.

[41] Tong, T., Chen, L., and Zhao, H. (2012). Improved mean estimation and its application to diagonal discriminant analysis. Bioinformatics 28:531–537.

[42] Wolberg, W. H. and Mangasarian, O. L. (1990). Multi-surface method of pattern separation for medical diagnosis applied to breast cancer. Proceedings of the National Academy of Sciences 87:9193–9196.

[43] Xu. P., Brock, G., and Parrish, R. (2009). Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. Computational Statistics and Data Analysis 53:1674–1687.

[44] Yang, W. and Wu, H. (2014). Regularized complete linear discriminant analysis. Neurocomputing 137:185–191.

[45] Yang, J. and Yang, J. (2003). Why can LDA be performed in PCA transformed space. Pattern Recognition 36:563–566.

[46] Yang, J. Frangi, A., Yang, J., Zhang, D., and Jin, Z. (2005). KPCA plus LDA: A complete kernal Fisher discriminant framework for feature extraction and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 27:230–243.

[47] Yao, J., Zheng, S., and Bai, Z. (2015). Sample covariance matrices and high dimensional data analysis. Cambridge University Press: London.