Sparse Linear Discriminant Analysis with Applications to High Dimensional Low Sample Size Data

Zhihua Qiao,* Lan Zhou† and Jianhua Z. Huang‡

Abstract— **This paper develops a method for automatically incorporating variable selection in Fisher's linear discriminant analysis (LDA). Utilizing the connection of Fisher's LDA and a generalized eigenvalue problem, our approach applies the method of regularization to obtain sparse linear discriminant vectors, where "sparse" means that the discriminant vectors have only a small number of nonzero components. Our sparse LDA procedure is especially effective in the so-called high dimensional, low sample size (HDLSS) settings, where LDA possesses the "data piling" property, that is, it maps all points from the same class in the training data to a common point, and so when viewed along the LDA projection directions, the data are piled up. Data piling indicates overfitting and usually results in poor out-of-sample classification. By incorporating variable selection, the sparse LDA overcomes the data piling problem. The underlying assumption is that, among the large number of variables there are many irrelevant or redundant variables for the purpose of classification. By using only important or significant variables we essentially deal with a lower dimensional problem. Both synthetic and real data sets are used to illustrate the proposed method.**

Keywords: Classification, linear discriminant analysis, variable selection, regularization, sparse LDA

1 Introduction

Fisher's linear discriminant analysis (LDA) is typically used as a feature extraction or dimension reduction step before classification. The most popular tool for dimensionality reduction is principal components analysis (PCA, Pearson 1901, Hotelling 1933). PCA searches for a few directions to project the data such that the projected data explain most of the variability in the original data. In this way, one obtains a low dimensional representation of the data without losing much information. Such attempt to reduce dimensionality can be described as "parsimonious summarization" of the data. However, since PCA targets for the unsupervised problem, it would not be suitable for classification problems.

For a classification problem, how does one utilize the class information in finding informative projections of the data? Fisher (1936) proposed a classic approach: Find the projection direction such that for the projected data, the between-class variance is maximized relative to the within-class variance. Additional projection directions with decreasing importance in discrimination can be defined in sequence. The total number of projection directions one can define is one less than the number of classes. Once the projection directions are identified, the data can be projected to these directions to obtain the reduced data, which are usually called discriminant variables. For the discriminant variables, any classification method can be carried out, such as nearest centroid, k-nearest neighborhood, and support vector machines. A particular advantage of Fisher's idea is that one can exploit the graphical tools. For example, with two projection directions one can view the data in a two-dimensional plot, color-coding the classes.

An important query in application of Fisher's LDA is whether all the variables on which measurements are obtained contain useful information or only some of them may suffice for the purpose of classification. Since the variables are likely to be correlated, it is possible that a subset of these variables can be chosen such that the others may not contain substantial additional information and may be deemed redundant in the presence of this subset of variables. A case for variable selection in Fisher's LDA can be made further by pointing out that by increasing the number of variables we do not necessarily ensure an increase in the discriminatory power. This is a form of overfitting. One explanation is that when the number of variables is large, the within-class covariance matrix is hard to be reliably estimated. In additional to avoiding overfitting, interpretation can be facilitated if

*MIT Sloan School of Management, 50 Memorial Drive, E52-456, Cambridge, MA 02142, USA. Email: zqiao@MIT.EDU

†Department of Statistics, Texas A&M University, 447 Blocker Building, College Station, TX 77843-3143, USA. Email: lzhou@stat.tamu.edu.

‡*Corresponding Author.* Department of Statistics, Texas A&M University, 447 Blocker Building, College Station, TX 77843-3143, USA. Email:jianhua@stat.tamu.edu.

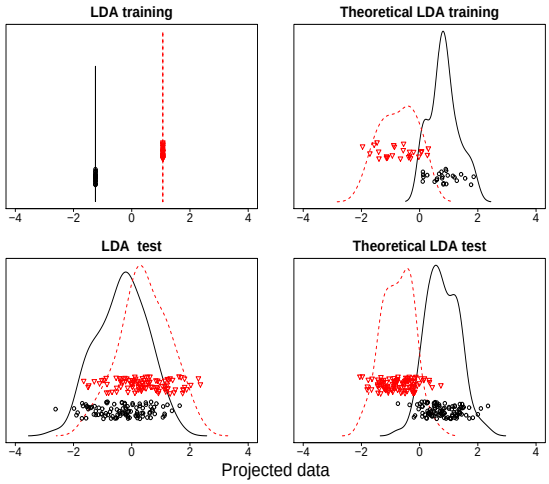

Figure 1: A simulated example with two classes. Plotted are the projected data using the estimated and theoretical LDA directions. Top panels are for training data; bottom panels for test data. Left panels use estimated LDA directions; right panels the theoretical directions. The in-sample and out-of-sample error rates are 0 and 32% respectively, when applying the nearest centroid method to the data projected to the estimated LDA direction. The dimension of the data is 100 and there are 25 cases for each class.

we incorporate variable selection in LDA.

We find that variable selection may provide a promising approach to deal with a very challenging case of data mining: the high dimensional, low sample size (HDLSS, Marron et al. 2007) settings. The HDLSS means that the dimension of the data vectors is larger (often much larger) than the sample size (the number of data vectors available). HDLSS data occur in many applied areas such as genetic micro-array analysis, chemometrics, medical image analysis, text classification, and face recognition. As pointed out by Marron et al., classical multivariate statistical methods often fail to give a meaningful analysis in HDLSS contexts.

Ahn and Marron (2007) and Marron et al. (2007) discovered an interesting phenomenon called "data piling" for discriminant analysis in HDLSS settings. Data piling means that when the data are projected onto some projection direction, many of the projections are exactly the same, that is, the data pile up on top of each other. Data piling is not a useful property for discrimination, because the corresponding direction vector is driven by very particular aspects of the realization of the training data at hand. Data piling direction provides perfect separation of classes in sample, but it inevitably has bad generalization property.

The Fisher's LDA is not applicable to HDLSS settings since the within-class covariance matrix is singular. Several extensions of LDA that can overcome the singularity problem, including pseudo-inverse LDA, Uncorrelated LDA (Ye et al. 2006), and Orthogonal LDA (Ye 2005), all possess the data piling problem. As an illustration of the data piling problem, Figure 1 provides views of two simulated data sets, one of which serves as a training data set, shown in the first row, the other the test data set, shown in the second row. The data are projected onto some direction vector and the projections are represented as a "jitter plot", with the horizontal coordinate representing the projection, and with a random vertical coordinate used for visual separation of the points. A kernel density estimate is also shown in each plot to reveal the structure of the projected data. Two methods are considered to find a projection direction in Figure 1. Fisher's LDA (using pseudo-inverse of the with-in class covariance matrix) is applied to the training data set to obtain the projection direction for the left panels, while the theoretical LDA direction, which is based on the knowledge of the true within-class and between-class covariance matrices, is used for the right panels. The LDA direction estimated using training data possesses obvious data piling and overfitting. The perfect class separation in sample does not translate to good separation out of sample. In contrast, the projections to the theoretical LDA direction for the two data sets have similar distributional properties.

One contribution of the present paper is to offer a method to deal with the "data piling" problem in HDLSS settings. If a small number of significant variables suffice for discrimination, then identifying these variables may help prevent "data piling" in the training data and consequently yield good out-of-sample classification. In Section 4.1, the same data sets will be projected to the sparse LDA direction estimated using the training data. We will see that these projections will resemble the distributional behavior on the right panels of Figure 1 that are based on the theoretical LDA directions. The main message is that without variable selection, LDA is subject to data piling and leads to bad out-of-sample classification; with variable selection, data piling on training data is prevented and thereby good classification on test data is obtained.

The rest of the paper is organized as follows. Section 2 reviews Fisher's LDA and also serves the purpose of introducing necessary notations for subsequent sections. In Section 3, we describe our sparse regularized LDA method for constructing sparse discriminant vectors. Numerical algorithm for finding these vectors is also provided. Sections 4 and 5 illustrate the proposed method using two simulated data examples and two real data sets. Section 6 concludes. Some technical proofs are given in the Appendix.

2 Review of Fisher's LDA

Discriminant analysis has been a standard topic in any multivariate analysis text book (e.g., Mardia, et al., 1979). A common approach to discriminant analysis is to apply the decision theory framework. In this framework, one assumes a parametric form of the population distribution and a prior probability for each class, then derives the Bayesian decision rule for classification. If the assumed population distribution for each class is multivariate normal and the covariances are common across different classes, the resulting decision rule is based on a linear function of the input data and therefore called linear discriminant analysis (LDA). Although the strong assumptions used in this derivation of LDA are not true in many applications, LDA has been proven very effective. This is mainly due to the fact that a simple, linear model is more robust against noise, and less likely to overfit.

An alternative approach to discrimination analysis can be made by merely looking for a "sensible" rule to discriminate the classes without assuming any particular parametric form for the distribution of the populations. Fisher's LDA looks for the linear function $a^T x$ such that the ratio of the between-class sum of squares to the within-class sum of squares is maximized. Formally, suppose there are k classes and let $x_{ij}, j = 1, \ldots, n_i$, be vectors of observations from the i-th class, $i = 1, \ldots, k$. Set $n = n_1 + \ldots, n_k$ and let $\bar{x}_i$ denote the mean of the i-th class. Let

$$X_{n\times p} = (x_{11}, \ldots, x_{1n_1}, \ \ldots, \ x_{k1}^T, \ldots, x_{kn_k}^T)^T$$

and $y = Xa$, then Fisher's LDA solves

$$\max_a \frac{\sum_{i=1}^k n_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}, \tag{1}$$

where $\bar{y}_i$ is the mean of the i-th sub-vector y_i of y, corresponding to the i-th class. Substituting y by Xa, we can rewrite the within-class sum of squares as

$$\sum_{i=1}^k \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = a^T \sum_{i=1}^k \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T a \stackrel{\text{def}}{=} a^T \Sigma_w a,$$

and the between-class sum of squares as

$$\sum_{i=1}^k n_i(\bar{y}_i - \bar{y})^2 = \sum_{i=1}^k n_i\{a^T(\bar{x}_i - \bar{x})\}^2 = a^T \sum_{i=1}^k n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T a \stackrel{\text{def}}{=} a^T \Sigma_b a.$$

Therefore the ratio is given by

$$a^T \Sigma_b a / a^T \Sigma_w a.$$

If a_1 is the vector that maximizes the ratio, one can find the next direction a_2 orthogonal in Σ_w to a_1, such that the ratio is maximized; and the additional directions can be computed in sequence similarly. The projection directions a_i are usually called discriminant coordinates and the linear functions $a_i^T x$ are called Fisher's linear discriminant functions. Fisher's criterion is intuitively sound because it is relatively easy to tell the classes apart if the between-class sum of squares for y is large relative to the within-class sum of squares. Alternatively, Fisher's criterion can be understood as dimension reduction using principal components analysis on the class centroids standardized by the common within-class covariance matrix.

The problem (1) was Fisher's original formulation of LDA. Another formulation of LDA popular in the pattern recognition literature (i.e., Fukunaga, 1990) is to solve the optimization problem

$$\max_A \{\text{tr}\,(A^T \Sigma_w A)^{-1} A^T \Sigma_b A\} \tag{2}$$

subject to $A^T A = I$. Both (1) and (2) are equivalent to finding a's that satisfy $\Sigma_b a = \eta \Sigma_w a$, for $\eta \neq 0$. This is a generalized eigenvalue problem. There are no more than $\min(p, k-1)$ eigenvectors corresponding to nonzero eigenvalues, since the rank of the matrix Σ_b is bounded from above by $\min(p, k-1)$.

In this paper, we view LDA as a supervised dimension reduction tool that searches for suitable projection directions, and therefore refer to eigenvectors a_i's as the discriminant directions or discriminant vectors. These discriminant directions/vectors are useful for data visualization and also for classification. By projecting the p-dimensional data onto the q-dimensional space spanned by the first q ($q \leq \min(p, k-1)$) discriminant vectors, we reduce the p-dimensional data to q-dimensional data. The low dimensional data can be easily visualized, using for example pairwise scatterplots. Any methods of discriminant analysis, such as nearest centroid method, nearest neighborhood method, and the support vector machines, can be applied to the reduced data to develop classification rules.

To facilitate subsequent discussion, we introduce some notations here. Define $n \times p$ matrices

$$H_w = X - \begin{pmatrix} e^{n_1}\bar{x}_1^T \\ \vdots \\ e^{n_k}\bar{x}_k^T \end{pmatrix} \text{ and } H_b = \begin{pmatrix} e^{n_1}(\bar{x}_1 - \bar{x})^T \\ \vdots \\ e^{n_k}(\bar{x}_k - \bar{x})^T \end{pmatrix},$$

where e^{n_i} is a column vector of ones with length n_i and e is a column vector of ones with length n. It is clear that with these notations, we have

$$\Sigma_w = H_w^T H_w \quad \text{and} \quad \Sigma_b = H_b^T H_b.$$

Notice that the matrix H_b can be reduced to a lower dimension $(k \times p)$ matrix

$$(\sqrt{n_1}(\bar{x}_1 - \bar{x}), \ldots, \sqrt{n_k}(\bar{x}_k - \bar{x}))^T, \tag{3}$$

which also satisfies $\Sigma_b = H_b^T H_b$. In the discussion that follows, this latter form of H_b is used throughout without further mentioning.

3 Sparse Discriminant Vectors

When Σ_w is positive definite, the first discriminant direction vector a in Fisher's LDA is the eigenvector corresponding to the largest eigenvalue of the following generalized eigenvalue problem

$$\Sigma_b \beta = \eta \Sigma_w \beta. \tag{4}$$

To incorporate variable selection in LDA corresponds to making the eigenvector a sparse. Here "sparsity" means that the eigenvector a has only a few nonzero components or it has lots of zero components. It is not so obvious how to achieve this. However, variable selection methods are well studied for linear regression problems and those methods are useful in suggesting a feasible approach to extracting sparse eigenvectors.

LASSO (Tibshirani, 1996) is a penalized least squares method that imposes a constraint on the L_1 norm of regression coefficients. Specifically, the LASSO solves the following optimization problem

$$\min_{\beta} \|Y - X\beta\|^2 + \lambda \|\beta\|_1,$$

where Y, X are the response vector and design matrix respectively, and β is the vector of regression coefficients. Due to the nature of the L_1 penalty, some components of β will be shrunk to exact zero if λ is large enough. Therefore the LASSO can produce a sparse coefficient vector β, which makes it a variable selection method.

3.1 Link of generalized eigenvalue problems to regressions

Our approach for obtaining sparse discriminant vectors is an extension of the sparse PCA method of Zou et al. (2006). It first relates the discriminant vector to a regression coefficient vector by transforming the generalized eigenvalue problem to a regression-type problem, and then apply penalized least squares with an L_1 penalty. The following theorem will serve our purpose.

Theorem 1. *Suppose Σ_w is positive definite and denote its Cholesky decomposition as $\Sigma_w = R_w^T R_w$, where $R_w \in \mathbb{R}^{p\times p}$ is an upper triangular matrix. Let $H_b \in \mathbb{R}^{k\times p}$ be defined as in (3). Let $V_1, \ldots, V_q$ $(q \le \min(p, k-1))$ denote the eigenvectors of problem (4) corresponding to the q largest eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_q$. Let $A_{p\times q} = [\alpha_1, \ldots, \alpha_q]$ and $B_{p\times q} = [\beta_1, \ldots, \beta_q]$. For $\lambda > 0$, let $\widehat{A}$ and $\widehat{B}$ be the solution to the following problem*

$$\min_{A,B} \sum_{i=1}^{k} \|R_w^{-T} H_{b,i} - AB^T H_{b,i}\|^2 + \lambda \sum_{j=1}^{q} \beta_j^T \Sigma_w \beta_j, \quad \text{subject to } A^T A = I_{q\times q}, \tag{5}$$

where $H_{b,i} = \sqrt{n_i}(\bar{x}_i - \bar{x})^T$ is the i-th row of H_b. Then $\hat{\beta}_j, j = 1, \ldots, q$, span the same linear space as $V_j, j = 1, \ldots, q$.

To prove Theorem 1, we need the following Lemma.

Lemma 1. *Let M be a $p \times p$ symmetric positive semi-definite matrix. Assume $q < p$ and the eigenvalues of M satisfy $d_{11} \ge \cdots \ge d_{qq} > d_{q+1,q+1} \ge \cdots \ge d_{pp} \ge 0$. The $p \times q$ matrix A that maximizes $\mathrm{tr}\,(A^T M A)$ under the constrain that $A^T A = I$ has the form $A = V_1 U_1$ where V_1 consists the first q eigenvectors of M and U_1 is an arbitrary $q \times q$ orthogonal matrix.*

Proof of Lemma 1. Let the eigenvalue decomposition of M be $M = VDV^T$, where V is orthogonal and $D = \mathrm{diag}(d_{11}, \ldots, d_{pp})$ is diagonal, both are $p \times p$ matrices. Note that $\mathrm{tr}\,(A^T M A) = \mathrm{tr}\,(A^T V D V^T A)$. Let $U = V^T A$, which is a $p \times q$ matrix. Then $A = VU$. Moreover, $U^T U = A^T V V^T A = I$, which means that U has orthonormal columns. Denote the rows of U by $u_1^T, \ldots, u_p^T$. Then

$$\mathrm{tr}\,(A^T M A) = \mathrm{tr}\,(U^T D U) = \sum_{i=1}^{p} d_{ii} \mathrm{tr}\,(u_i u_i^T) = \sum_{i=1}^{p} d_{ii} |u_i|^2.$$

Since U has orthonormal columns, $|u_i|^2 \le 1$ for $i = 1, \ldots, p$ and $\sum_{i=1}^{p} |u_i|^2 = q$. The problem reduces to maximizing $\sum_{i=1}^{p} d_{ii}|u_i|^2$ subject to the constraints that $|u_i|^2 \le 1$ for $i = 1, \ldots, p$ and $\sum_{i=1}^{p} |u_i|^2 = q$. Note that $d_{11} \ge \cdots \ge d_{qq} > d_{q+1,q+1} \cdots \ge d_{pp} > 0$ are arranged in decreasing order. It is clear that the optimization problem is solved by $|u_1| = \cdots = |u_q| = 1$ and $|u_{q+1}| = \cdots = |u_p| = 0$. This implies that the first q rows of U form a $q \times q$ orthogonal matrix, denoted as U_1, and the rest rows of U consist of only zeros. Partition $V = (V_1, V_2)$ where V_1 is $p \times q$. Then we have $A = V_1 U_1$, which is the desired result. □

Now we are ready to prove Theorem 1.

Proof of Theorem 1. Using (9) we see that the optimal $B = [\beta_1, \ldots, \beta_q]$ for fixed A are

$$\hat{\beta}_j = (\Sigma_b + \lambda \Sigma_w)^{-1} \Sigma_b R^{-1} \alpha_j,$$

or equivalently

$$\widehat{B} = (\Sigma_b + \lambda \Sigma_w)^{-1} \Sigma_b R_w^{-1} A. \tag{6}$$

Then we substitute $\widehat{B}$ into the objective function of (5) and find that we need to maximize the object

$$\mathrm{tr}\,\{A^T R_w^{-T}\Sigma_b(\Sigma_b+\lambda\Sigma_w)^{-1}\Sigma_b R_w^{-1}A\} \tag{7}$$

as a function of A subject to $A^TA=I$.

Denote the q leading eigenvectors of $R_w^{-T}\Sigma_b R_w^{-1}$ by $E=[\eta_1,\ldots,\eta_q]$ so that $R_w^{-T}\Sigma_b R_w^{-1}=E\Lambda E^T$ where Λ is an $q\times q$ diagonal matrix of eigenvalues. The columns of E are also the q leading eigenvectors of the matrix

$$\begin{aligned}&R_w^{-T}\Sigma_b(\Sigma_b+\lambda\Sigma_w)^{-1}\Sigma_b R_w^{-1}\\ &= R_w^{-T}\Sigma_b R_w^{-1}(R_w^{-T}\Sigma_b R_w^{-1}+\lambda I)^{-1}R_w^{-T}\Sigma_b R_w^{-1}.\end{aligned}$$

Thus, according Lemma 1, the $\widehat{A}$ that maximizes (7) satisfies $\widehat{A}=EP$ where P is an arbitrary $q\times q$ orthogonal matrix. Substituting this $\widehat{A}$ into equation (6) results in

$$\begin{aligned}\widehat{B} &= R_w^{-1}(R_w^{-T}\Sigma_b R_w^{-1}+\lambda I)^{-1}R_w^{-T}\Sigma_b R_w^{-1}\widehat{A}\\ &= R_w^{-1}(E\Lambda E^T+\lambda I)^{-1}E\Lambda E^T EP\\ &= R_w^{-1}E(\Lambda+\lambda I)^{-1}\Lambda P\end{aligned}$$

Note that the q leading eigenvectors of the generalized eigenvalue problem (4) are columns of $V=R_w^{-1}E$. Therefore, $\widehat{B}=V(\Lambda+\lambda I)^{-1}\Lambda P$. The desired result follows. □

From theorem 1, we know that if W is positive definite, then the $B_\beta=(\beta_1,\ldots,\beta_q)$ that solves the optimization problem (5) contains the first q discriminant vectors. The optimization problem (5) can be solved by iteratively minimizing over A and B. The update of A for fixed B is a Procrustes problem (Gower and Dijksterhuis 2004). To see this, note that

$$\begin{aligned}&\sum_{i=1}^k \|R_w^{-T}H_{b,i}-AB^TH_{b,i}\|^2\\ &= \|H_bR_w^{-1}-H_bBA^T\|^2\\ &= \mathrm{tr}\{(H_bR_w^{-1}-H_bBA^T)(H_bR_w^{-1}-H_bBA^T)^T\}\\ &= \mathrm{tr}\{H_bR_w^{-1}R_w^{-T}H_b^T+H_bBB^TH_b^T\}\\ &\qquad -2\,\mathrm{tr}\{B^TH_b^TH_bR_w^{-1}A\};\end{aligned} \tag{8}$$

we have used $A^TA=I$ to obtain the last equality. Thus, if B is fixed, the update of A maximizes $\mathrm{tr}\{B^TH_b^TH_bR_w^{-1}A\}$ subject to the constraint that A has orthonormal columns. This is an inner-product version of projection Procrustes that has an analytical solution. The solution is given by computing the singular value decomposition

$$R_w^{-T}(H_b^TH_b)B=UDV^T,$$

where U $(p\times q)$ has orthonormal columns and V $(q\times q)$ is orthogonal, and setting $\widehat{A}=UV^T$. (See Cliff, 1966, Section 3 of Gower and Jijksterhuis, 2004, or Theorem 4 of Zou et al. 2006).

The update of B for fixed A is a regression-type problem. To see this, let $A_\perp$ be an orthogonal matrix such that $[A;A_\perp]$ is $p\times p$ orthogonal; this is feasible since A has orthonormal columns. Then we have that

$$\begin{aligned}&\|H_bR_w^{-1}-H_bBA^T\|^2\\ &= \|H_bR_w^{-1}[A;A_\perp]-H_bBA^T[A;A_\perp]\|^2\\ &= \|H_bR_w^{-1}A-H_bB\|^2+\|H_bR_w^{-1}A_\perp\|^2\\ &= \sum_{j=1}^q\|H_bR_w^{-1}\alpha_j-H_b\beta_j\|^2+\|H_bR_w^{-1}A_\perp\|^2.\end{aligned}$$

If A is fixed, then the B that optimizes (5) solves

$$\min_B\sum_{j=1}^q\{\|H_bR_w^{-1}\alpha_j-H_b\beta_j\|^2+\lambda\beta_j^T\Sigma_w\beta_j\}, \tag{9}$$

which is equivalent to q independent ridge regression problems.

3.2 Sparse eigenvectors

The connection to a regression-type problem of the optimization problem for extracting the discriminant vectors suggests an approach to produce sparse discriminant vectors. As in the LASSO, by adding an L_1 penalty to the objective function in the regression problem, we can obtain sparse regression coefficients. Therefore we consider the optimization problem

$$\begin{aligned}\min_{A,B}\sum_{j=1}^q\{&\|H_bR_w^{-1}\alpha_j-H_b\beta_j\|^2\\ &+\lambda\beta_j^T\Sigma_w\beta_j+\lambda_{1,j}\|\beta_j\|_1\},\end{aligned} \tag{10}$$

subject to $A^TA=I_{q\times q}$, where $\|\beta_j\|_1$ is the 1-norm of the vector β_j, or equivalently,

$$\begin{aligned}\min_{A,B}&\sum_{i=1}^k\|R_w^{-T}H_{b,i}-AB^TH_{b,i}\|^2\\ &+\lambda\sum_{j=1}^q\beta_j\Sigma_w\beta_j+\sum_{j=1}^q\lambda_{1,j}\|\beta_j\|_1,\end{aligned} \tag{11}$$

subject to $A^TA=I_{q\times q}$. Whereas the same λ is used for all q directions, different $\lambda_{1,j}$'s are allowed to penalize different discriminant directions.

The optimization problem (10) or (11) can be numerically solved by alternating optimization over A and B.

- **B given A:** For each j, let $Y_j^*=H_bR_w^{-1}\alpha_j$. For fixed A, B is solved by q independent LASSO prob-

lems

$$\min_{\beta_j} \|Y_j^* - H_b\beta_j\|^2 + \lambda\beta_j^T\Sigma_w\beta_j + \lambda_{1,j}\|\beta_j\|_1, \quad j = 1, \ldots, q. \tag{12}$$

- **A given B**: For fixed B, we can ignore the penalty term in (11) and need only minimize

$$\sum_{i=1}^{k} \|R_w^{-T}H_{b,i} - AB^TH_{b,i}\|^2 = \|H_bR_w^{-1} - H_bBA^T\|^2$$

subject to $A^TA = I_{q\times q}$. The solution is obtained by computing the singular value decomposition

$$R_w^{-T}(H_b^TH_b)B = UDV^T$$

and letting $\widehat{A} = UV^T$.

Using the Cholesky decomposition $\Sigma_w = R_w^TR_w$, we see that for each j, (12) is equivalent to minimization of

$$\|\widetilde{Y}_j - \widetilde{W}\beta_j\|^2 + \lambda_{1,j}\|\beta_j\|_1,$$

where $\widetilde{Y}_j = (Y_j^{*T}, 0_{p\times p})^T$ and $\widetilde{W} = (H_b^T, R_w^T)^T$. This is a LASSO-type optimization problem where efficient implementations exist such as LARS (Efron et al. 2004).

Algorithm 1 summarizes the steps of our Sparse LDA procedure described above.

Remarks: 1. Theorem 1 implies that the solution of the optimization problem (5) is independent of the value of λ. This does not necessarily imply that the solution of the regularized problem (11) is also independent of λ. However, our empirical study suggests that the solution is very stable when λ varies in a wide range, for example in $(0.01, 10000)$.

2. We can use K-fold cross validation (CV) to select the optimal tuning parameters $\{\lambda_{1,j}\}$. We use the error rate of a specified classification method such as the nearest centroid or nearest neighbor method applied on the projected data to generate the cross validation score. Specifically, we randomly split the data into K parts. Fixing one part at a time as the test data, we apply the Sparse LDA to the rest of the data (treated as training data) to find the sparse discriminant vectors. Then we project all the data onto these discriminant vectors, apply a given classification method to the training data, and calculate the classification error using the test data. After we repeat this K times, with one part as test data at a time, we combine the classification errors as the selection criterion. When the dimension of the input data is very large, the numerical algorithm becomes time-consuming and we can let $\lambda_{1,1} = \cdots = \lambda_{1,q}$ to expedite computation.

Algorithm 1 *Sparse LDA Algorithm*

1. Form the matrices $H_b \in \mathbb{R}^{k\times p}$ and $H_w \in \mathbb{R}^{n\times p}$ from the data as follows

$$H_w = X - \begin{pmatrix} e^{n_1}\bar{x}_1^T \\ \vdots \\ e^{n_k}\bar{x}_k^T \end{pmatrix} \text{ and } H_b = \begin{pmatrix} \sqrt{n_1}(\bar{x}_1 - \bar{x}) \\ \vdots \\ \sqrt{n_k}(\bar{x}_k - \bar{x}) \end{pmatrix}.$$

2. Compute the upper triangular matrix $R_w \in \mathbb{R}^{p\times p}$ from the Cholesky decomposition of $H_w^TH_w$ such that $H_w^TH_w = R_w^TR_w$.

3. Solve q independent LASSO problems

$$\min_{\beta_j} \beta_j^T(\widetilde{W}^T\widetilde{W})\beta_j - 2\tilde{y}^T\widetilde{W}\beta_j + \lambda_1\|\beta_j\|_1,$$

where

$$\widetilde{W}_{(n+p)\times p} = \begin{pmatrix} H_b \\ \sqrt{\lambda}R_w \end{pmatrix}, \tilde{y}_{(n+p)\times 1} = \begin{pmatrix} H_bR_w^{-1}\alpha_j \\ 0 \end{pmatrix}.$$

4. Compute the singular value decomposition $R_w^{-T}(H_b^TH_b)B = UDV^T$ and let $A = UV^T$.

5. Repeat steps 3 and 4 until converges.

3.3 Sparse regularized LDA

When the number of variables exceeds the sample size, i.e., the high dimensional, low sample size (HDLSS) settings according to Marron et al. (2005), the within-class covariance matrix is singular and the method proposed above breaks down. One method to circumvent this singularity problem is to regularize the within-class covariance, similar to the regularization method as used in ridge regression.

Consider a standard linear regression with the design matrix X. When X is collinear or close to being collinear, the normal equation $X^TX\beta = X^TY$ is ill-conditioned and can not produce a stable solution. To stabilize the solution, ridge regression adds a positive multiple of the identity matrix to the Gram matrix X^TX in forming the normal equation, that is, $(X^TX + \gamma I)\beta = X^TY$. Similarly, when the within-class covariance matrix Σ_w is singular, we can replace it by $\Sigma_w + \gamma I$ when applying LDA. This idea has been proposed in the past, see for example, Champbell (1980), Peck and Van Ness (1982), Friedman (1989), and Rayens (1990). We adopt this idea but furthermore introduce sparsity in the discriminant vectors.

Algorithm 2 *Sparse rLDA Algorithm*

1. Form the matrix $H_B \in \mathbb{R}^{k\times p}$ and $H_W \in \mathbb{R}^{n\times p}$ from the data as follows

$$H_W = X - \begin{pmatrix} e^{n_1}\bar{x}_1^T \\ \vdots \\ e^{n_k}\bar{x}_k^T \end{pmatrix}$$

$$H_B = (\sqrt{n_1}(\bar{x}_1 - \bar{x}), \ldots, \sqrt{n_k}(\bar{x}_k - \bar{x}))^T$$

2. Compute upper triangular matrix $R_w \in \mathbb{R}^{p\times p}$ from the Cholesky decomposition of $\Sigma_w + (\gamma/p)\,\mathrm{tr}\,(\Sigma_w)I$ such that $\Sigma_w + (\gamma/p)\,\mathrm{tr}\,(\Sigma_w)I = R_w^T R_w$.

3. Solve q independent LASSO problems

$$\min_{\beta_j} \beta_j^T(\widetilde{W}^T\widetilde{W})\beta_j - 2\tilde{y}^T\widetilde{W}\beta_j + \lambda_1\|\beta_j\|_1, \qquad (13)$$

where

$$\widetilde{W}_{(n+p)\times p} = \begin{pmatrix} H_B \\ \sqrt{\lambda}R_W \end{pmatrix}, \; \tilde{y}_{(n+p)\times 1} = \begin{pmatrix} H_B R_W^{-1}\alpha_j \\ 0 \end{pmatrix}$$

4. Compute the singular value decomposition $R_W^{-T}(H_B^T H_B)B = UDV^T$ and let $A = UV^T$.

5. Repeat steps 3 and 4 until converges.

Consider the generalized eigenvalue problem

$$\Sigma_b\beta = \eta\left(\Sigma_w + \gamma\frac{\mathrm{tr}\,(\Sigma_w)}{p}I\right)\beta,$$

where γ is a regularization parameter. The identity matrix is scaled by $\mathrm{tr}\,(\Sigma_w)/p$ so that the matrices Σ_w and $\{\mathrm{tr}\,(\Sigma_w)/p\}I$ have the same trace. We refer to this problem as regularized LDA (rLDA for short). Following the same development as in Section 3.2, we see that the eigenvectors $\beta_1, \ldots, \beta_q$, associated with the first q largest eigenvalues of the generalized eigenvalue problem can be obtained up to a scaling constant as the solution to the following regression-type problem

$$\min_{\substack{A\in\mathbb{R}^{p\times q}\\ B\in\mathbb{R}^{p\times q}}} \sum_{i=1}^{k}\|R_w^{-T}H_{b,i} - AB^T H_{b,i}\|^2 + \lambda\sum_{j=1}^{q}\beta_j^T\left(\Sigma_w + \gamma\frac{\mathrm{tr}\,(\Sigma_w)}{p}I\right)\beta_j,$$

subject to $A^TA = I_{q\times q}$, where $B = [\beta_1, \ldots, \beta_q]$. This connection suggests using the L_1 penalty to obtain sparsity for the β_j's.

We define the first q sparse discriminant directions $\beta_1, \ldots, \beta_q$ as the solutions to the following optimization problem

$$\min_{\substack{A\in\mathbb{R}^{p\times q}\\ B\in\mathbb{R}^{p\times q}}} \sum_{i=1}^{k}\|R_w^{-T}H_{b,i} - AB^T H_{b,i}\|^2 + \lambda\sum_{j=1}^{q}\beta_j^T\left(\Sigma_w + \gamma\frac{\mathrm{tr}\,(\Sigma_w)}{p}I\right)\beta_j + \sum_{j=1}^{q}\lambda_{1,j}\|\beta_j\|_1, \qquad (14)$$

subject to $A^TA = I_{q\times q}$, where $B = [\beta_1, \ldots, \beta_q]$. Algorithm 1 can be modified to obtain the sparse rLDA directions. The resulting algorithm is summarized in Algorithm 2. The two algorithms only differ in step 2.

Remark: In (14), γ is a tuning parameter that controls the strength of regularization of the within-class covariance matrix. A large value of γ will bias too much the within-class covariance matrix towards identity matrix. There are two helpful criteria for choosing γ. First, if the sample size is small and the dimension is high, then the within-class covariance matrix is not accurately estimated and therefore we want to employ a high degree of regularization by using a relatively large value of γ. Second, we can exploit the graphical tools to choose a suitable γ with the aid of data visualization. For example, we seek a γ that yields a good separation of classes for the training data set. In our empirical studies, however, we find that the results of sparse rLDA are not sensitive to the choice of γ if a small value that is less than 0.1 is used. We shall use $\gamma = 0.05$ for the empirical results to be presented in Sections 4 and 5. More careful studies of choice of γ are left for future research.

4 Simulated Data

4.1 Two classes

Our first simulation example contains training data set of size 25 for each of the two classes and test data set of size 100 for each class. The input data X has dimension $p = 100$ so this is a HDLSS setting. Only the first two variables of X can distinguish the two classes, and the remaining 98 variables are irrelevant for discrimination. The distribution of each class is

$$x_i \sim \begin{pmatrix} N_2(\mu_i, \Sigma_{w,2}) \\ N_{p-2}(0, I_{p-2}) \end{pmatrix}, \quad i = 1, 2,$$

$$\mu_i = \begin{pmatrix} 0 \\ \pm 0.9 \end{pmatrix}, \qquad \Sigma_{w,2} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}.$$

There is only one discriminant direction of Fisher's LDA since we have two classes. Clearly, the theoretical discriminant direction depends only on the first two variables. Hence we can ignore the redundant variables in

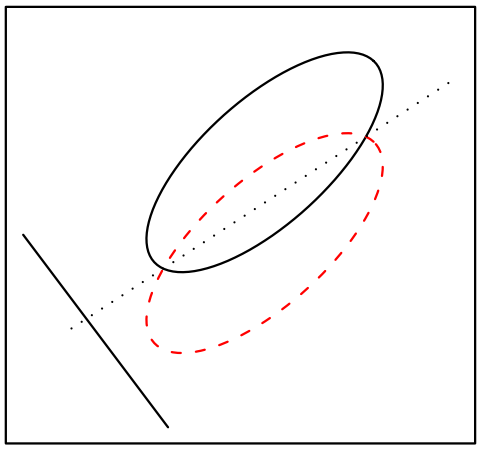

Figure 2: The theoretical projection direction and the ellipses of the population distributions of the two classes.

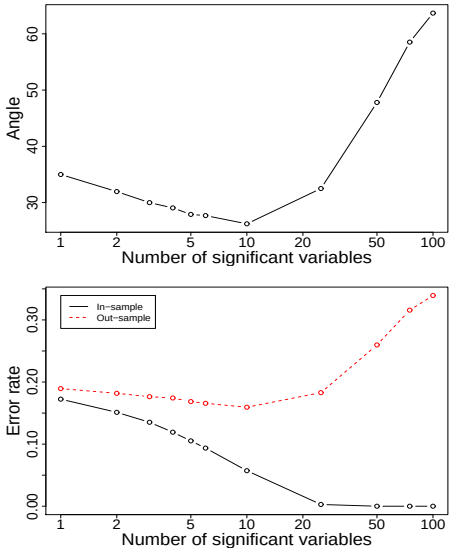

Figure 3: A simulated example with two classes. Top panel: The average of angles between the estimated and theoretical directions as a function of the number of variables used. Bottom panel: Average classification error rates using least centroid on the projected data. Based on 50 simulations.

deriving the theoretical direction. The between-class covariance matrix is given by

$$\Sigma_{w,2} = \sum_{i=1}^{2}(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^T = \frac{1}{2}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

and the within-class covariance matrix is $\Sigma_{b,2}$. The theoretical discriminant direction is the leading eigenvector of $\Sigma_{w,2}^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$, which is $(-0.57, 0.82)$ in this example. This projection direction and the ellipses of the population distributions of the two classes are plotted in Figure 2. The estimated direction will be compared with the theoretical direction derived here.

Since this is a HDLSS case, Σ_w is singular and therefore sparse LDA is not directly applicable. We thus applied the sparse rLDA to the simulated data sets. Denote the number of significant variables involved in specifying the discriminant direction to be m. For each of 50 simulated data sets, we applied sparse rLDA for $m = 1$, 2, 3, 4, 5, 10, 20, 30, 40, 50, 75, 100, and calculated the angles between the estimated and the true discriminant directions. The average angles as a function of m is plotted in the top panel of Figure 3. It is very clear that sparsity helps: Compare average angles around 30 degrees for $m = 2$–20 to an average angle about 60 degrees for $m = 100$. Sparse discriminant vectors are closer to the theoretical direction than the non-sparse ones.

The theoretical discriminant direction has only 2 nonzero components, while in Figure 3 the smallest average angle is achieved when $m = 10$. Although difference of the average angles between $m = 2$ and $m = 10$ is not significant, one may wonder why $m = 10$ is the best instead of $m = 2$. The main reason for this discrepancy is the insufficiency of training sample size, which causes the estimation of the covariance matrix Σ_w inaccurate and therefore the inclusion of more variables. We did some simulation experiments with increased sample size and observed that the optimal m indeed decrease and come closer to $m = 2$.

The closeness of estimated direction to the theoretical direction also translates into out-of-sample classification performance. The bottom panel of Figure 3 shows the in-sample and out-of-sample classification error rate using nearest centroid method applied to the projected data. When all variables are used in constructing the discriminant vectors, the overfitting of training data is apparent, and is associated with low in-sample error rate and high out-of-sample error rate. The out-of-sample error rate is minimized when the number of significant variables used in constructing the discriminant vectors is ten. It is also interesting to point out that the shape of the out-of-sample error rate curve resembles that of the average angle curve shown on the top panel of Figure 3.

The discriminant power of the sparse discriminant projection is illustrated in Figure 4, where we plotted the projected, both training and test, data. On the left panels, regularized LDA with penalty parameter $\gamma = 0.05$ was used to obtain the discriminant direction. Comparing with the upper left panel of Figure 1, we see that regularized LDA does help alleviate data piling slightly, but does not help improve out-of-sample classification.

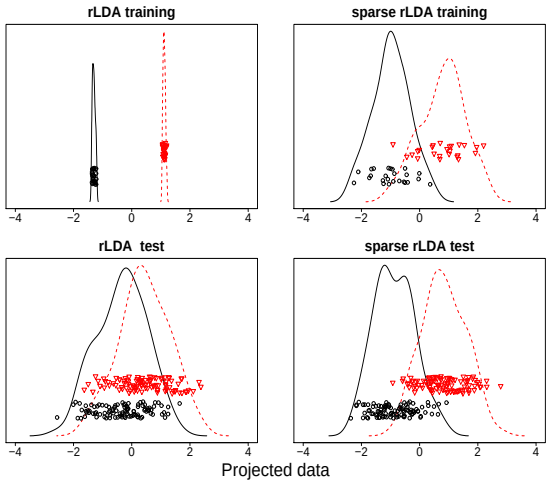

Figure 4: A simulated example with two classes. Top panels are the results of rLDA and sparse rLDA ($m = 5$) for the training data; bottom panels are the results for the test data. The in-sample and out-of-sample error rates are 0 and 32% for rLDA and 12% and 13.5% for sparse rLDA, when applying the nearest centroid method to the projected data. The dimension of the data is 100 and there are 25 cases for each class.

On the other hand, if sparsity is imposed in obtaining the discriminant direction, data piling of training set disappears and substantial improvement in test set classification is manifested.

4.2 Three classes

In this example we have three classes. The dimension of the input data is $p = 100$. For each class, there are 25 cases in the training data set and 200 cases in the test data set. The distribution of each class is

$$x_i \sim \begin{pmatrix} N_3(\mu_i, \Sigma_{w,3}) \\ N_{p-3}(0, I_{p-3}) \end{pmatrix}, \quad i = 1, 2, 3,$$

where

$$(\mu_1, \mu_2, \mu_3) = \begin{pmatrix} 0 & 0.9 & 0 \\ 0 & -0.9 & 0 \\ 1.6 & 1.1 & 0 \end{pmatrix}$$

and

$$\Sigma_{w,3} = \begin{pmatrix} 1 & 0 & 0.7 \\ 0 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{pmatrix}.$$

There are two discriminant directions of Fisher's LDA for this three-class problem. We first derive the theoretical directions. As in the previous example, we can ignore the redundant variables in this calculation. The between-class covariance matrix is given by

$$\Sigma_{b,3} = \frac{1}{2}\sum_{i=1}^{3}(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^T$$

and the within-class covariance matrix is $\Sigma_{w,3}$. The two true projection directions are the eigenvectors of $\Sigma_{w,3}^{-1}\Sigma_{b,3}$.

Since the within-class covariance matrix with all variables is singular, we applied sparse rLDA to the simulated data sets. There are two discriminant directions that form a 2-dimensional dimension reduction subspace. In estimating the sparse discriminant directions, we let the number of included variables m in the two projections to be the same. For a set of values of m, we calculated the angles between the estimated projection subspace and the theoretical discriminant subspace for 50 simulation runs. Here the angle between two q-dimensional spaces is defined by $\theta = \arccos(\rho)$, where ρ is the vector correlation coefficient (Hotelling 1936) defined in the following way. Suppose A and B are two matrices whose columns form the orthonormal bases of the two q-dimensional spaces under consideration. Then the vector correlation coefficient is

$$\rho = \Big(\prod_{i=1}^{q}\rho_i^2\Big)^{1/2},$$

where ρ_i^2 are the eigenvalues of the matrix B^TAA^TB. The top panel of Figure 5 shows the average angles as a function of m, while the bottom panel shows the error rate of nearest centroid applied to the projected data. Both the average angle and the out-of-sample error rate is minimized around $m = 30$. Use of only a small number of variables in constructing the discriminant directions does help estimate the theoretical directions more accurately and improve the classification performance.

To illustrate the discriminant power using estimated discriminant directions, we show in Figure 6 the projection of a training data set and a test data set on the subspace spanned by the two discriminant directions obtained using rLDA and sparse rLDA. Without incorporating sparsity in the discriminant vector, the data piling is apparent in training data set and projected data lead to terrible discrimination. As a contrast, if sparsity is imposed in estimating the discriminant directions, the data piling in the training set is avoided and good test set classification is achieved.

5 Real Data Examples

5.1 Wine data

The data, described in Forina et al. (1991) and available at The UCI Repository of Machine Learning Databases (Merz and Murphy, 1996), are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The chemical analysis determined the quantities of 13 constituents found in each of the three types of wines. Our statistical analy-

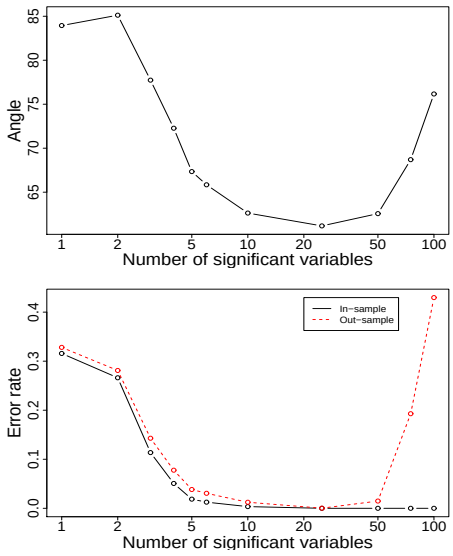

Figure 5: A simulated example with three classes. Top panel: The average of angles between the estimated and theoretical directions as a function of the number of variables used. Bottom panel: Average classification error rates using least centroid on the projected data. Based on 50 simulations.

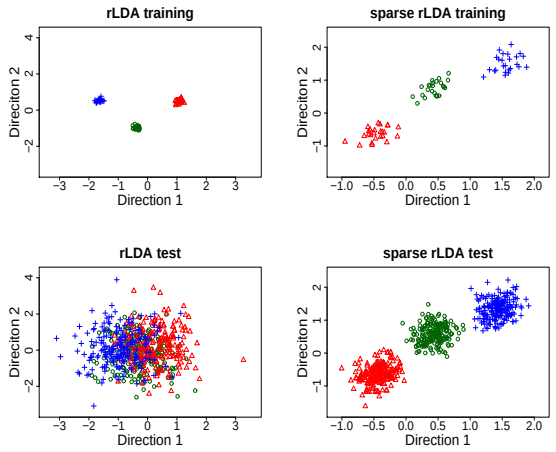

Figure 6: A simulated example with three classes. Top panels are the results of rLDA and sparse rLDA ($m = 10$) for the training data; bottom panels are the results for the test data. The in-sample and out-of-sample error rates are 0 and 44.3% for rLDA and 0 and 0.3% for sparse rLDA, when applying the nearest centroid method to the projected data. The dimension of the data is 100 and for each class, there are 25 cases in the training data and 200 cases in the test data.

sis is to find which type of wine a new sample belonging to based on its 13 attributes. The data consists of 178 instances, each belonging to one of three classes. This is not a HDLSS setting. Fisher's LDA projects the data to a two-dimensional subspace of $\mathbb{R}^{13}$. We would like to explore if it is possible to classify the wines using only part of the 13 constituents. In estimating the sparse discriminant directions, for simplicity, we let both direction vectors to have the same number of nonzero components m, which is set between 1 and 13.

We partitioned the data randomly into a training set with 2/3 of the data and a test set with 1/3 of the data. We did the random partition 50 times to average out the variability in the results due to the partition. For each partition, we used sparse LDA on the training data and obtained two sparse discriminant directions. Then we used the nearest centroid method on the projected data and computed the test set error rates. Figure 7 shows the average of error rates for 50 partitions as a function of the number of variables involved in each estimated sparse direction.

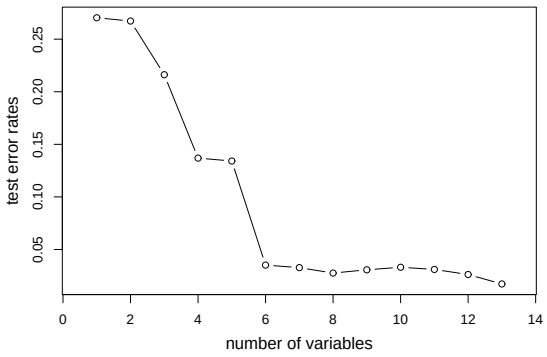

Figure 7: Wine data. The average of error rates as a function of the number of variables. Based on 50 random partitions of the dataset into training and test.

From the plot it is clear that if we specify $m = 6$ for each discriminate direction, sparse LDA can discriminate the classes fairly well. To check the stability of variable selection, we fixed $m = 6$, did 50 random partitions, and recorded the selected variables for each partition. Table 1 summarizes the frequency of each variable being selected. It shows that the variable selection is not sensitive to the random partition. Overall, eight variables are important in the discriminant analysis.

Finally, we picked one random partition and compared Fisher's LDA with the sparse LDA by plotting the projected data as in Figure 8. The results show that using six variables in each discriminant projection can separate

Table 1: Wine data. Frequency of selected variables.

variable ♯	1	2	3	4	5	6	7	8	9	10	11	12	13
projection 1	1	11	0	50	45	0	50	0	0	50	0	43	50
projection 2	50	49	2	48	45	0	1	0	4	49	0	2	50
importance	√	√		√	√		√			√		√	√

the data almost as well as using all the variables. It is not surprising that the separation is the best when using all the variables because, in this example, each variable is a constituent that characterizes the wine and would not be redundant. But the sparse LDA does suggest which constituents are the most important in the classification of wines.

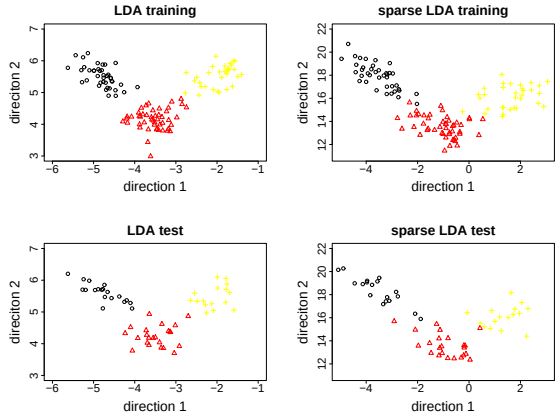

Figure 8: Wine data. Projection of the data (training and test separately) onto the two discriminant directions.

5.2 Gene expression data

We use two gene expression microarray data sets to illustrate the sparse rLDA method. Gene expression microarrays can be used to discriminate between multiple clinical and biological classes. It is a typical example of HDLSS settings because there are usually thousands of genes while the availability of the patients is very limited.

The first data set is the Colon data set (Alon et al., 1999), which contains 42 tumor and 20 normal colon tissue samples. For each sample there are 2000 gene expression level measurements. The second data set is the Prostate data set (Singh et al., 2002), which contains 52 tumor samples and 50 normal samples. For each sample there exist 6033 gene expression level measurements. For both data sets, the goal of the analysis is classification of tumor and normal samples based on the gene expression measurements.

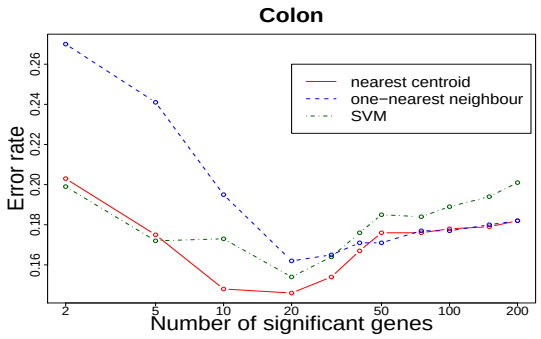

Figure 9: Colon data. The average test error rate as a function of the number of significant genes for the nearest centroid, 1-nearest neighbor and support vector machine, applied to the reduced data obtained from sparse rLDA. Based on 50 (2:1) training-test partition of the original data set.

For each data set, we first reduce the dimensionality of the data by projecting the data to the discriminant directions obtained using sparse rLDA, then the reduced data is used as an input to some standard classification methods. We shall examine the effect of gene selection on classification. Our sparse rLDA algorithm incorporates gene selection to constructing discriminant vector. To expedite computation, we implemented a two-step procedure. First we do a crude gene preselection using the Wilcoxon rank test statistic to obtain 200 significant genes. Then the preselected gene expressions are used as input to sparse rLDA. Note that even after gene preselection, we still have HDLSS settings, so regularization of with-in class covariance matrices is needed and the sparse rLDA instead of the sparse LDA algorithm should be applied.

In the absence of genuine test sets we performed our comparative study by repeated random splitting of the data into training and test sets. The data were partitioned into a balanced training set comprising two-thirds of the arrays, used for gene preselection, applying sparse rLDA for dimension reduction and fitting the classifiers. Then, the class labels of the remaining one-third of the experiments were predicted, compared with the true labels, and

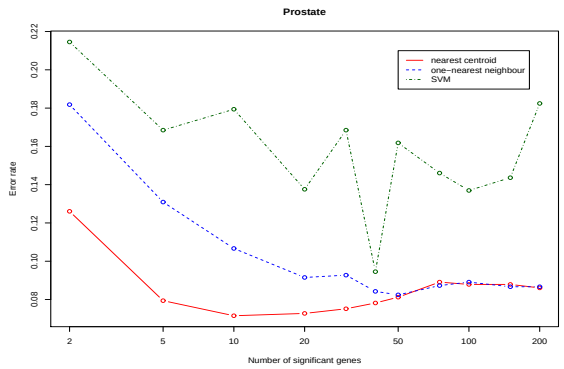

Figure 10: Prostate data. The average test error rate as a function of the number of significant genes for the nearest centroid, 1-nearest neighbor and support vector machine, applied to the reduced data obtained from sparse rLDA. Based on 50 (2:1) training-test partition of the original data set.

the misclassification error rate was computed. To reduce variability, the splitting into training and test sets were repeated 50 times and the error rate is averaged. It is important to note that, for reliable conclusion, all gene preselection, applying sparse rLDA and fitting classifiers were re-done on each of the 50 training sets.

Three classifiers, the nearest centroid, 1-nearest neighbor and support vector machine, have been applied to the reduced data for classification. Figures 9 and 10 plot the average test error rate as a function of significant genes used in sparse rLDA for the two data sets. The x-axis is plotted using the logarithmic scale to put less focus on large values. As the number of significant genes vary from 2 to 200, the error rates for three methods all decrease first and then rise. The nearest centroid method has the best overall classification performance. The beneficial effect of variable selection in sparse rLDA is clear: The classification using reduced data based on sparse discriminant vectors perform better than that based on non-sparse discriminant vectors. For example, if the nearest centroid method is used as the classifier, using the sparse discriminant vectors based on only 10-20 significant genes gives the best test set classification, while using all 200 genes is harmful to classification. Note that the 200 genes used are preselected significant genes, the benefit of using the sparse rLDA could be much bigger if the 200 genes were randomly selected.

6 Conclusions

In this paper, we propose a novel algorithm for constructing sparse discriminant vectors. The sparse discriminant vectors are useful for supervised dimension reduction for high dimensional data. Naive application of classical Fisher's LDA to high dimensional, low sample size settings suffers from the data piling problem. Introducing sparsity in the discriminant vectors is very effective in eliminating data piling and the associated overfitting problem. Our results on simulated and real data examples suggest that, in the presence of irrelevant or redundant variables, the sparse LDA method can select important variables for discriminant analysis and thereby yield improved classification.

Acknowledgments

Jianhua Z. Huang was partially supported by grants from the National Science Foundation (DMS-0606580) and the National Cancer Institute (CA57030). Lan Zhou was partially supported by a training grant from the National Cancer Institute.

Reference

1. Ahn, J. and Marron, J.S., 2007, The maximal data piling direction for discrimination, manuscript.

2. Alon, U., Barkai, N., Notterman, D., Gish, K., Mack, S. and Levine, J., 1999, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by ologonucleotide arrays, *Proc. Natl. Acad. Sci, USA*, **96**, 503-511.

3. Campbell, N.A., 1980, Shrunken estimator in discriminant and canonical variate analysis, *Applied Statistics*, **29**, 5-14.

4. Cliff, N., 1966, Orthogonal rotation to congruence. *Psychometrika*, **31**, 33-42.

5. Efron B., Hastie T., Johnstone, I., and Tibshirani, R., 2004, Least angle regression, *Ann. Statist.*, **32**, 407-499.

6. Fisher, R. A., 1936, The use of multiple measurements in taxonomic problems. *Ann. Eugen*, **7**, 179-188.

7. Friedman, J. H., 1989, Regularized discriminant analysis, *Journal of the American Statistical Association*, **84**, 165-175.

8. Gower, J. C. and Dijksterhuis, G. B., 2004, Procrustes Problems, Oxford University Press, New York.

9. Hotelling, H., 1933, Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, **24**, 417-441, 498–520.

10. Hotelling, H., 1936, Relations between two sets of variates, *Biometrika*, **28**, 321-377.

11. Fukunaga, K., 1990, Introduction to Statistical Pattern Classification. Academic Press, San Diego, California, USA, 1990.

12. Mardia, K.V., Kent, J.T. and Bibby J.M., 1979, Multivariate analysis, *Academic Press*.

13. Marron, J. S., Todd, M. and Ahn, J., 2007, Distance weighted discrimination, *Journal of American Statistical Association*, **102**, 1267–1271.

14. Pearson, K., 1901, On lines and planes of closest fit to systems of points in space, *Philosophical Magazine*, **2**, 559-572.

15. 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**, 531-537.

16. Rayens, W.S., 1990, A Role for covariance stabilization in the construction of the classical mixture surface, *Journal of Chemometrics*, **4**, 159-169.

17. Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A.V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P., W., Golub, T. R., and Sellers, W. R., 2002, Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, **1**, 203-209.

18. Tibshirani, R., 1996, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.

19. Ye, J., 2005, Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems, *Journal of Machine Learning Research*, **6**, 483-502.

20. Ye, J., Janardan, R., Li, Q. and Park H., 2006, Feature reduction via generalized uncorrelated linear discriminant analysis, emphIEEE Transactions on Knowledge and Data Engineering, **18**, 1312-1322.

21. Zou, H., Hastie, T. and Tibshirani, R., 2006, Sparse principal component analysis, *Journal of Computational and Graphical Statistics*, **15**, 157-177.