Effective Linear Discriminant Analysis for High Dimensional, Low Sample Size Data

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

Abstract— **In the so-called high dimensional, low sample size (HDLSS) settings, 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.**

In this paper, a novel approach to overcome the data piling problem is introduced. It incorporates variable selection into LDA. 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. Experiments on both synthetic and real data sets show that the proposed method is effective in overcoming the data piling and overfitting problem of LDA while improving the out-of-sample classification performance.

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. It finds the projection directions such that for the projected data, the between-class variance is maximized relative to the within-class variance. 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. These discriminant variables can be used as inputs to any classification method, such as nearest centroid, k-nearest neighborhood, and support vector machines.

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 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 gene expression microarray analysis, chemometrics, medical image analysis, text classification, and face recognition. As pointed out by Marron et al. (2007), classical multivariate statistical methods often fail to give a meaningful analysis in HDLSS contexts.

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 in sample separation of classes, but it inevitably has bad generalization property.

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

[*] 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. Lan Zhou's work was supported by a training grant from the US National Cancer Institute.

[‡] Department of Statistics, Texas A&M University, 447 Blocker Building, College Station, TX 77843-3143, USA. Email:jianhua@stat.tamu.edu. Jianhua Z. Huang's work was partially supported by grants from the US National Science Foundation and the US National Cancer Institute.

ISBN:978-988-17012-3-7

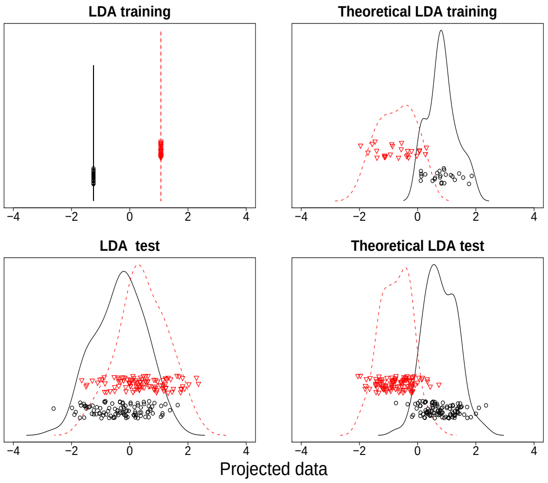

Figure 1: A simulated example with two classes. Plotted are the projected data using the estimated and theoretical LDA directions. Top panels are for training data; bottom panels for test data. Left panels use estimated LDA directions; right panels the theoretical directions. The in-sample and out-of-sample 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 training data set is 100 and there are 25 cases for each class.

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 within-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, the proposed sparse LDA method will be applied to the same data sets used in Figure 1. We will see that the projections to the sparse LDA direction will resemble the distributional behavior on the right panels of Figure 1 that are based on the theoretical LDA direction. 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 LDA method for constructing sparse discriminant vectors. Sections 4 and 5 illustrate the proposed method using a simulated data example and a real data set. Section 6 concludes.

2 Review of Fisher's LDA

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 $n = n_1 + \ldots, n_k$. Let

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

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 ith sub-vector y_i of y. Substituting y by Xa, we can rewrite the within-class sum of squares as

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

and the between-class sum of squares as

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

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 sequentially.

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.

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

$$(\sqrt{n_1}(\bar{x}_1 - \bar{x}), \ldots, \sqrt{n_k}(\bar{x}_k - \bar{x}))^T, \tag{2}$$

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

$$\Sigma_b \beta = \eta \Sigma_w \beta. \tag{3}$$

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. 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 applies penalized least squares with an L_1 penalty as in LASSO (Tibshirani, 1996). We refer to our method as sparse LDA.

3.1 Link of generalized eigenvalue problems to regressions

We will first consider the case that Σ_w is non-singular. The Σ_w singular case will be discussed in Section 3.3. The following theorem is crucial to our approach. The proof of the theorem can be found in Qiao (2006).

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 (2). Let $V_1, \ldots, V_q$ ($q \leq \min(p, k-1)$) denote the eigenvectors of problem (3) corresponding to the q largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$. 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*

$$\min_{A,B} \sum_{i=1}^{k} \|R_w^{-T} H_{b,i} - AB^T H_{b,i}\|^2 + \lambda \sum_{j=1}^{q} \beta_j^T \Sigma_w \beta_j, \tag{4}$$
$$\text{subject to } A^T A = I_{q\times q},$$

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 $V_j, j = 1, \ldots, q$.

The optimization problem (4) 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

$$\sum_{i=1}^{k} \|R_w^{-T} H_{b,i} - AB^T H_{b,i}\|^2 = \|H_b R_w^{-1} - H_b B A^T\|^2.$$

Since $A^T A = I$, the above expression equals

$$\operatorname{tr}\{H_b R_w^{-1} R_w^{-T} H_b^T + H_b B B^T H_b^T\} - 2 \operatorname{tr}\{B^T H_b^T H_b R_w^{-1} A\}$$

Thus, if B is fixed, the update of A maximizes $\operatorname{tr}\{B^T H_b^T H_b R_w^{-1} A\}$ 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

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

where U $(p \times q)$ has orthonormal columns and V $(q \times q)$ is orthogonal, and setting $\widehat{A} = UV^T$. (See Section 5.1 of Gower and Jijksterhuis, 2004).

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

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

If A is fixed, then the B that optimizes (4) solves

$$\min_B \sum_{j=1}^{q} \{\|H_b R_w^{-1} \alpha_j - H_b \beta_j\|^2 + \lambda \beta_j^T \Sigma_w \beta_j\}, \tag{5}$$

which is equivalent to q independent ridge regression problems.

3.2 Sparse eigenvectors

According to (5), the eigenvectors β_j are regression coefficient vectors. As in the LASSO, by adding an L_1 penalty to the objective function in the regression problem, we can obtain sparse regression coefficient vectors. Therefore we consider the optimization problem

$$\min_{A,B} \sum_{j=1}^{q} \{ \| H_b R_w^{-1} \alpha_j - H_b \beta_j \|^2 + \lambda \beta_j^T \Sigma_w \beta_j + \lambda_{1,j} \|\beta_j\|_1 \},$$

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

$$\begin{aligned} \min_{A,B} \sum_{i=1}^{k} \| R_w^{-T} H_{b,i} - AB^T H_{b,i} \|^2 \\ + \lambda \sum_{j=1}^{q} \beta_j \Sigma_w \beta_j + \sum_{j=1}^{q} \lambda_{1,j} \|\beta_j\|_1, \end{aligned} \tag{6}$$

subject to $A^T A = I_{q \times q}$. Whereas the same λ is used for all q directions, different $\lambda_{1,j}$'s are allowed to penalize the loadings of different discriminant directions.

The optimization problem (6) can be numerically solved by alternating optimization over A and B.

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

$$\min_{\beta_j} \|Y_j^* - H_b \beta_j\|^2 + \lambda \beta_j^T \Sigma_w \beta_j + \lambda_{1,j} \|\beta_j\|_1, \quad j = 1, \ldots, q. \tag{7}$$

- **A given B**: For fixed B, compute the singular value decomposition

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

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

Using the Cholesky decomposition $\Sigma_w = R_w^T R_w$, we see that for each j, (7) 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 which can be solved by an efficient computation algorithm (Zou et al. 2006).

Remarks: 1. Theorem 1 implies that the solution of the optimization problem (4) is independent of the value of λ. This does not imply that the solution of the regularized problem (6) 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. 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.

3.3 Sparse regularized LDA

When the within-class covariance matrix is singular, regularized LDA (rLDA for short) can be used to circumvent the singularity problem as in ridge regression. Specifically, one version of regularized LDA replaces Σ_w by $\widetilde{\Sigma}_w = \Sigma_w + (\gamma/p)\,\mathrm{tr}\,(\Sigma_w) I$ in the standard LDA, where $\gamma > 0$ is a tuning parameter that controls the strength of regularization of the within-class covariance matrix. The identity matrix is scaled by $\mathrm{tr}\,(\Sigma_w)/p$ here so that the matrices Σ_w and $\{\mathrm{tr}\,(\Sigma_w)/p\} I$ have the same trace. There is a straightforward extension of sparse LDA to regularized LDA: One just replaces Σ_w by $\widetilde{\Sigma}_w$ when compute the Cholesky factor R_w in Theorem 1. We refer to the resulting method as sparse regularized LDA (sparse rLDA for short).

Remark: In our empirical studies, 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

We illustrate our method using a simulated data example which 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 variables are irrelevant for discrimination. The distribution of each class is

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

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

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

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

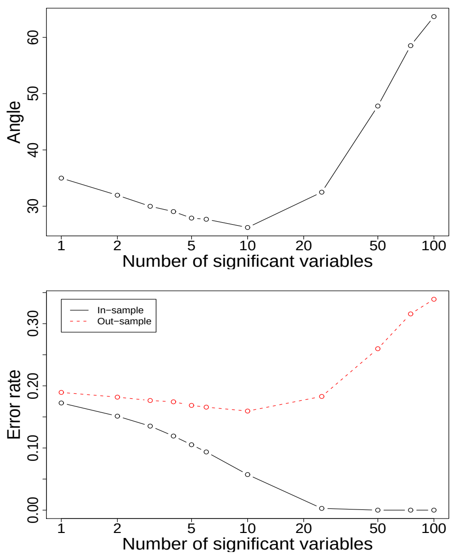

Figure 2: A simulated example with two classes. Top panel: The average of angles between the estimated and theoretical directions as a function of the number of variables used. Bottom panel: Average classification error rates using least centroid on the projected data. Based on 50 simulations.

and the within-class covariance matrix is $\Sigma_{w,2}$. 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. 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 applied the sparse rLDA with penalty parameter $\gamma = 0.05$ 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 2. 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. That the smallest average angle is achieved when $m = 10$ instead of $m = 2$ is because of the insufficiency of training sample size, which causes the estimation of the covariance matrix Σ_w inaccurate and therefore the inclusion of more variables.

The closeness of estimated direction to the theoretical direction also translates into out-of-sample classification performance. The bottom panel of Figure 2 shows the in-sample and out-of-sample classification error rate using nearest centroid method applied to the projected data.

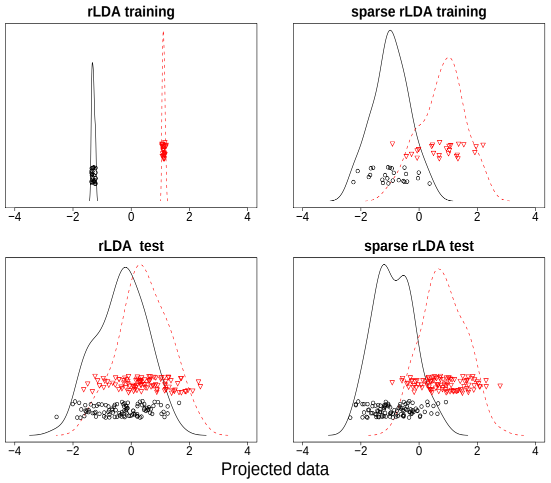

Figure 3: A simulated example with two classes. Top panels are the results of rLDA and sparse rLDA ($m = 5$) for the training data; bottom panels are the results for the test data. The in-sample and out-of-sample error rates are 0 and 32% for rLDA and 12% and 13.5% for sparse rLDA, when applying the nearest centroid method to the projected data. The dimension of the training data set is 100 and there are 25 cases for each class.

When all variables are used in constructing the discriminant vectors, the overfitting of training data is obvious, 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 2.

The discriminant power of the sparse discriminant projection is illustrated in Figure 3, where we plotted the projected, both training and test, data. Regularized LDA was used to obtain the discriminant direction for the left panels. 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. 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.

We have done more simulation studies of various number of classes and have reached the same conclusion as in the above example.

5 Gene expression microarray data

We use a gene expression microarray data to illustrate the sparse rLDA method. The Colon data set (Alon et al., 1999) contains 42 tumor and 20 normal colon tissue samples. For each sample there are 2000 gene expression

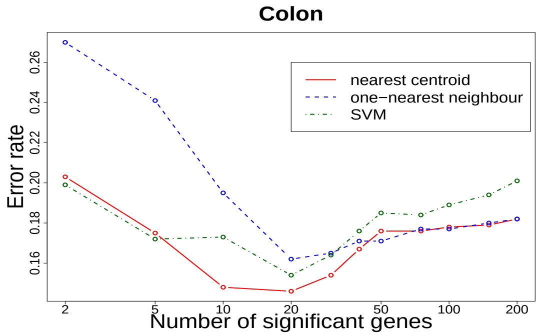

Figure 4: Colon data. The average test error rate as a function of the number of significant genes for the nearest centroid, 1-nearest neighbor and support vector machine, applied to the reduced data obtained from sparse rLDA. Based on 50 (2:1) training-test partition of the original data set.

level measurements. The goal of the analysis is classification of tumor and normal samples based on the gene expression measurements.

We first reduce the dimensionality of the data by projecting the data to the discriminant direction obtained using 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 the 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 within-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 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. Figure 4 plots 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 the three methods all decrease first and then rise. The nearest centroid method has the best overall classification performance. The beneficial effect of the variable selection in sparse rLDA is clear: The classification using reduced data based on a sparse discriminant vector performs better than that based on a non-sparse discriminant vector. 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 leads to larger classification error rate.

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 simulated and real data examples results 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.

References

[1] 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*, Vol 96, pp. 503-511.

[2] Gower, J. C. and Dijksterhuis, G. B., 2004, *Procrustes Problems*, Oxford University Press.

[3] Marron, J. S., Todd, M. J. and Ahn, J., 2007, "Distance weighted discrimination," *Journal of American Statistical Association*, Vol 480, pp. 1267-1271.

[4] Qiao, Z., 2006, *Variable selection in multivariate data analysis using regularization*, Ph.D. thesis, The Wharton School, University of Pennsylvania.

[5] Tibshirani, R., 1996, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, Vol 58, pp. 267-288.

[6] Zou, H., Hastie, T. and Tibshirani, R., 2006, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, Vol 15, pp. 157-177.