Penalized classification using Fisher's linear discriminant

Daniela M. Witten †

Department of Biostatistics, University of Washington, USA

Robert Tibshirani

Department of Health Research & Policy, and Statistics, Stanford University, USA

Summary. We consider the supervised classification setting, in which the data consist of p features measured on n observations, each of which belongs to one of K classes. Linear discriminant analysis (LDA) is a classical method for this problem. However, in the high-dimensional setting where $p \gg n$, LDA is not appropriate for two reasons. First, the standard estimate for the within-class covariance matrix is singular, and so the usual discriminant rule cannot be applied. Second, when p is large, it is difficult to interpret the classification rule obtained from LDA, since it involves all p features. We propose *penalized LDA*, a general approach for penalizing the discriminant vectors in Fisher's discriminant problem in a way that leads to greater interpretability. The discriminant problem is not convex, so we use a minorization-maximization approach in order to efficiently optimize it when convex penalties are applied to the discriminant vectors. In particular, we consider the use of L_1 and fused lasso penalties. Our proposal is equivalent to recasting Fisher's discriminant problem as a biconvex problem. We evaluate the performances of the resulting methods on a simulation study, and on three gene expression data sets. We also survey past methods for extending LDA to the high-dimensional setting, and explore their relationships with our proposal.

Keywords: classification; feature selection; high dimensional; lasso; linear discriminant analysis; supervised learning.

1. Introduction

In this paper, we consider the classification setting. The data consist of a $n \times p$ matrix $\mathbf{X}$ with p features measured on n observations, each of which belongs to one of K classes. Linear discriminant analysis (LDA) is a well-known method for this problem in the classical setting where $n > p$. However, in high dimensions (when the number of features is large relative to the number of observations) LDA faces two problems:

(a) The maximum likelihood estimate of the within-class covariance matrix is approximately singular (if p is almost as large as n) or singular (if $p > n$). Even if the estimate is not singular, the resulting classifer can suffer from high variance, resulting in poor performance.

†Corresponding Author. Email: dwitten@u.washington.edu. Mailing Address: University of Washington, Department of Biostatistics, F-600 Health Sciences Building, Box 357232, Seattle WA 98195-7232, USA.

(b) When p is large, the resulting classifier is difficult to interpret, since the classification rule involves a linear combination of all p features.

The LDA classifier can be derived in three different ways, which we will refer to as the *normal model*, the *optimal scoring problem*, and *Fisher's discriminant problem* (see e.g. Mardia et al., 1979; Hastie et al., 2009). In recent years, a number of papers have extended LDA to the high-dimensional setting in such a way that the resulting classifier involves a sparse linear combination of the features (see e.g. Tibshirani et al., 2002, 2003; Grosenick et al., 2008; Leng, 2008; Clemmensen et al., 2011). These methods involve *regularizing* or *penalizing* the log likelihood for the normal model, or the optimal scoring problem, by applying an L_1 or lasso penalty (Tibshirani, 1996).

In this paper, we instead approach the problem through Fisher's discriminant framework, which is in our opinion the most natural of the three problems that result in LDA. The resulting problem is nonconvex. We overcome this difficulty using a minorization-maximization approach (see e.g. Lange et al., 2000; Hunter and Lange, 2004; Lange, 2004), which allows us to solve the problem efficiently when convex penalties are applied to the discriminant vectors. This is equivalent to recasting Fisher's discriminant problem as a biconvex problem that can be optimized using a simple iterative algorithm, and is closely related to the sparse principal components analysis proposal of Witten et al. (2009).

To our knowledge, our approach to penalized LDA is novel. Clemmensen et al. (2011) state the same criterion that we use, but then go on to solve instead a closely related optimal scoring problem. Trendafilov and Jolliffe (2007) consider a closely related problem, but they propose a specialized algorithm that can be applied only in the case of L_1 penalties on the discriminant vectors; moreover, they do not consider the high-dimensional setting. In this paper, we take a more general approach that has a number of attractive features:

(a) It results from a natural criterion for which a simple optimization strategy is provided.
(b) A reduced rank solution can be obtained.
(c) It provides a natural way to enforce a diagonal estimate for the within-class covariance matrix, which has been shown to yield good results in the high-dimensional setting (see e.g. Dudoit et al., 2001; Tibshirani et al., 2003; Bickel and Levina, 2004).
(d) It yields interpretable discriminant vectors, where the concept of interpretability can be chosen based on the problem at hand. Interpretability is achieved via application of convex penalties to the discriminant vectors. For instance, if L_1 penalties are used, then the resulting discriminant vectors are sparse.

This paper is organized as follows. We review Fisher's discriminant problem in Section 2, we review the principle behind minorization-maximization algorithms in Section 3, and we propose our approach for penalized classification using Fisher's linear discriminant in Section 4. A simulation study and applications to gene expression data are presented in Section 5. Since many proposals have been made for sparse LDA, we review past work and discuss the relationships between various approaches in Section 6. In Section 7, we discuss connections between our proposal and past work. Section 8 contains the Discussion.

2. Fisher's discriminant problem

2.1. Fisher's discriminant problem with full rank within-class covariance

Let $\mathbf{X}$ be a $n \times p$ matrix with observations on the rows and features on the columns. We assume that the features are centered to have mean zero, and we let $\mathbf{X}_j$ denote feature/column j and $\mathbf{x}_i$ denote observation/row i. $C_k \subset \{1, \ldots, n\}$ contains the indices of the observations in class k, and $n_k = |C_k|, \sum_{k=1}^{K} n_k = n$. The standard estimate for the *within-class covariance matrix* $\mathbf{\Sigma}_w$ is given by

$$\hat{\mathbf{\Sigma}}_w = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T \tag{1}$$

where $\hat{\boldsymbol{\mu}}_k$ is the sample mean vector for class k. In this section, we assume that $\hat{\mathbf{\Sigma}}_w$ is non-singular. Furthermore, the standard estimate for the *between-class covariance matrix* $\mathbf{\Sigma}_b$ is given by

$$\hat{\mathbf{\Sigma}}_b = \frac{1}{n}\mathbf{X}^T\mathbf{X} - \hat{\mathbf{\Sigma}}_w = \frac{1}{n} \sum_{k=1}^{K} n_k \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T. \tag{2}$$

In later sections, we will make use of the fact that $\hat{\mathbf{\Sigma}}_b = \frac{1}{n}\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}$, where $\mathbf{Y}$ is a $n \times K$ matrix with Y_{ik} an indicator of whether observation i is in class k.

Fisher's discriminant problem seeks a low-dimensional projection of the observations such that the between-class variance is large relative to the within-class variance. That is, we sequentially solve

$$\text{maximize}_{\boldsymbol{\beta}_k \in \mathbb{R}^p} \{\boldsymbol{\beta}_k^T \hat{\mathbf{\Sigma}}_b \boldsymbol{\beta}_k\} \text{ subject to } \boldsymbol{\beta}_k^T \hat{\mathbf{\Sigma}}_w \boldsymbol{\beta}_k \leq 1, \boldsymbol{\beta}_k^T \hat{\mathbf{\Sigma}}_w \boldsymbol{\beta}_i = 0 \ \ \forall i < k. \tag{3}$$

Note that the problem (3) is generally written with the inequality constraint replaced with an equality constraint, but the two are equivalent if $\hat{\mathbf{\Sigma}}_w$ has full rank, as is shown in the Appendix. We will refer to the solution $\hat{\boldsymbol{\beta}}_k$ to (3) as the *kth discriminant vector.* In general, there are $K - 1$ nontrivial discriminant vectors.

A classification rule is obtained by computing $\mathbf{X}\hat{\boldsymbol{\beta}}_1, \ldots, \mathbf{X}\hat{\boldsymbol{\beta}}_{K-1}$ and assigning each observation to its nearest centroid in this transformed space. Alternatively, one can transform the observations using only the first $k < K - 1$ discriminant vectors in order to perform *reduced rank classification.* LDA derives its name from the fact that the classification rule involves a linear combination of the features.

One can solve (3) by substituting $\tilde{\boldsymbol{\beta}}_k = \hat{\mathbf{\Sigma}}_w^{\frac{1}{2}} \boldsymbol{\beta}_k$, where $\hat{\mathbf{\Sigma}}_w^{\frac{1}{2}}$ is the symmetric matrix square root of $\hat{\mathbf{\Sigma}}_w$. Then, Fisher's discriminant problem is reduced to a standard eigenproblem. In fact, from (2), it is clear that Fisher's discriminant problem is closely related to principal components analysis on the class centroid matrix.

2.2. Existing methods for extending Fisher's discriminant problem to the $p > n$ setting

In high dimensions, there are two reasons that problem (3) does not lead to a suitable classifier:

(a) $\hat{\mathbf{\Sigma}}_w$ is singular. Any discriminant vector that is in the null space of $\hat{\mathbf{\Sigma}}_w$ but not in the null space of $\hat{\mathbf{\Sigma}}_b$ can result in an arbitrarily large value of the objective.
(b) The resulting classifier is not interpretable when p is very large, because the discriminant vectors contain p elements that have no particular structure.

A number of modifications to Fisher's discriminant problem have been proposed to address the singularity problem. Krzanowski et al. (1995) consider modifying (3) by instead seeking a unit vector $\boldsymbol{\beta}$ that maximizes $\boldsymbol{\beta}^T\hat{\mathbf{\Sigma}}_b\boldsymbol{\beta}$ subject to $\boldsymbol{\beta}^T\hat{\mathbf{\Sigma}}_w\boldsymbol{\beta} = 0$, and Tebbens and Schlesinger (2007) further require that the solution does not lie in the null space of $\hat{\mathbf{\Sigma}}_b$. Others have proposed modifying (3) by using a positive definite estimate of $\mathbf{\Sigma}_w$. For instance, Friedman (1989), Dudoit et al. (2001), and Bickel and Levina (2004) consider the use of the diagonal estimate

$$\text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2), \tag{4}$$

where $\hat{\sigma}_j^2$ is the jth diagonal element of $\hat{\mathbf{\Sigma}}_w$ (1). Other positive definite estimates for $\mathbf{\Sigma}_w$ are suggested in Krzanowski et al. (1995) and Xu et al. (2009). The resulting criterion is

$$\text{maximize}_{\boldsymbol{\beta}_k \in \mathbb{R}^p}\{\boldsymbol{\beta}_k^T\hat{\mathbf{\Sigma}}_b\boldsymbol{\beta}_k\} \text{ subject to } \boldsymbol{\beta}_k^T\tilde{\mathbf{\Sigma}}_w\boldsymbol{\beta}_k \leq 1, \boldsymbol{\beta}_k^T\tilde{\mathbf{\Sigma}}_w\boldsymbol{\beta}_i = 0 \;\; \forall i < k, \tag{5}$$

where $\tilde{\mathbf{\Sigma}}_w$ is a positive definite estimate for $\mathbf{\Sigma}_w$. The criterion (5) addresses the singularity issue, but not the interpretability issue.

In this paper, we extend (5) so that the resulting discriminant vectors are interpretable. We will make use of the following proposition, which provides a reformulation of (5) that results in the same solution:

Proposition 1. *The solution $\hat{\boldsymbol{\beta}}_k$ to (5) also solves the problem*

$$\text{maximize}_{\boldsymbol{\beta}_k}\{\boldsymbol{\beta}_k^T\hat{\mathbf{\Sigma}}_b^k\boldsymbol{\beta}_k\} \text{ subject to } \boldsymbol{\beta}_k^T\tilde{\mathbf{\Sigma}}_w\boldsymbol{\beta}_k \leq 1 \tag{6}$$

where

$$\hat{\mathbf{\Sigma}}_b^k = \frac{1}{n}\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{P}_k^{\perp}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{X}. \tag{7}$$

$\mathbf{P}_k^{\perp}$ *is defined as follows:* $\mathbf{P}_1^{\perp} = \mathbf{I}$*, and for* $k > 1$*,* $\mathbf{P}_k^{\perp}$ *is an orthogonal projection matrix into the space that is orthogonal to* $(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}}_i$ *for all* $i < k$*.*

Throughout this paper, $\hat{\mathbf{\Sigma}}_w$ will always refer to the standard maximum likelihood estimate of $\mathbf{\Sigma}_w$ (1), whereas $\tilde{\mathbf{\Sigma}}_w$ will refer to some positive definite estimate of $\mathbf{\Sigma}_w$ for which the specific form will depend on the context.

3. A brief review of minorization algorithms

In this paper, we will make use of a *minorization-maximization* (or simply *minorization*) algorithm, as described for instance in Lange et al. (2000), Hunter and Lange (2004), and Lange (2004). Consider the problem

$$\text{maximize}_{\boldsymbol{\beta}}\{f(\boldsymbol{\beta})\}. \tag{8}$$

If f is a concave function, then standard tools from convex optimization (see e.g. Boyd and Vandenberghe, 2004) can be used to solve (8). If not, solving (8) can be difficult. (We note here that minimization of a convex function is a *convex problem*, as is maximization of a concave function. Hence, (8) is a convex problem if $f(\boldsymbol{\beta})$ is concave in $\boldsymbol{\beta}$. For non-concave $f(\boldsymbol{\beta})$ – for instance if $f(\boldsymbol{\beta})$ is convex – (8) is not a convex problem.)

Minorization refers to a general strategy for maximizing non-concave functions. The function $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ is said to minorize the function $f(\boldsymbol{\beta})$ at the point $\boldsymbol{\beta}^{(m)}$ if

$$f(\boldsymbol{\beta}^{(m)}) = g(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)}), \quad f(\boldsymbol{\beta}) \geq g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})\ \forall \boldsymbol{\beta}. \tag{9}$$

A minorization algorithm for solving (8) initializes $\boldsymbol{\beta}^{(0)}$, and then iterates:

$$\boldsymbol{\beta}^{(m+1)} = \text{argmax}_{\boldsymbol{\beta}}\{g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})\}. \tag{10}$$

Then by (9),

$$f(\boldsymbol{\beta}^{(m+1)}) \geq g(\boldsymbol{\beta}^{(m+1)}|\boldsymbol{\beta}^{(m)}) \geq g(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)}) = f(\boldsymbol{\beta}^{(m)}). \tag{11}$$

This means that in each iteration the objective is nondecreasing. However, in general we do not expect to arrive at the global optimum of (8) using a minorization approach: global optima for non-convex problems are very hard to obtain, and a local optimum is the best we can hope for except in specific special cases. Different initial values for $\boldsymbol{\beta}^{(0)}$ can be tried and the solution resulting in the largest objective value can be chosen. A good minorization function is one for which (10) is easily solved. For instance, if $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ is concave in $\boldsymbol{\beta}$ then standard convex optimization tools can be applied.

In the next section, we use a minorization approach to develop an algorithm for our proposal for penalized LDA.

4. The penalized LDA proposal

4.1. The general form of penalized LDA

We would like to modify the problem (5) by imposing penalty functions on the discriminant vectors. We define the *first penalized discriminant vector* $\hat{\boldsymbol{\beta}}_1$ to be the solution to the problem

$$\text{maximize}_{\boldsymbol{\beta}_1}\{\boldsymbol{\beta}_1^T\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta}_1 - P_1(\boldsymbol{\beta}_1)\} \text{ subject to } \boldsymbol{\beta}_1^T\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta}_1 \leq 1, \tag{12}$$

where $\tilde{\boldsymbol{\Sigma}}_w$ is a positive definite estimate for $\boldsymbol{\Sigma}_w$ and where P_1 is a convex penalty function. In this paper, we will be most interested in the case where $\tilde{\boldsymbol{\Sigma}}_w$ is the diagonal estimate (4), since it has been shown that using a diagonal estimate for $\boldsymbol{\Sigma}_w$ can lead to good classification results when $p \gg n$ (see e.g. Tibshirani et al., 2002; Bickel and Levina, 2004). Note that (12) is closely related to penalized principal components analysis, as described for instance in Jolliffe et al. (2003) and Witten et al. (2009) – in fact, it would be exactly penalized principal components analysis if $\tilde{\boldsymbol{\Sigma}}_w$ were the identity.

To obtain multiple discriminant vectors, rather than requiring that subsequent discriminant vectors be orthogonal with respect to $\tilde{\boldsymbol{\Sigma}}_w$ - a difficult task for a general convex penalty function - we instead make use of Proposition 1. We define the *kth penalized discriminant vector* $\hat{\boldsymbol{\beta}}_k$ to be the solution to

$$\text{maximize}_{\boldsymbol{\beta}_k} \{\boldsymbol{\beta}_k^T \hat{\boldsymbol{\Sigma}}_b^k \boldsymbol{\beta}_k - P_k(\boldsymbol{\beta}_k)\} \text{ subject to } \boldsymbol{\beta}_k^T \tilde{\boldsymbol{\Sigma}}_w \boldsymbol{\beta}_k \leq 1, \tag{13}$$

where $\hat{\boldsymbol{\Sigma}}_b^k$ is given by (7), with $\mathbf{P}_k^{\perp}$ an orthogonal projection matrix into the space that is orthogonal to $(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}}_i$ for all $i < k$, and $\mathbf{P}_1^{\perp} = \mathbf{I}$. Here P_k is a convex penalty function on the kth discriminant vector. Note that (12) follows from (13) with $k = 1$.

In general, the problem (13) cannot be solved using tools from convex optimization, because it involves maximizing an objective function that is not concave. We apply a minorization algorithm to solve it. For any positive semidefinite matrix $\mathbf{A}$, $f(\boldsymbol{\beta}) = \boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}$ is convex in $\boldsymbol{\beta}$. Thus, for a fixed value of $\boldsymbol{\beta}^{(m)}$,

$$f(\boldsymbol{\beta}) \geq f(\boldsymbol{\beta}^{(m)}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^{(m)})^T \nabla f(\boldsymbol{\beta}^{(m)}) = 2\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m)^T}\mathbf{A}\boldsymbol{\beta}^{(m)} \tag{14}$$

for any $\boldsymbol{\beta}$, and equality holds when $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$. Therefore,

$$g(\boldsymbol{\beta}_k|\boldsymbol{\beta}^{(m)}) = 2\boldsymbol{\beta}_k^T\hat{\boldsymbol{\Sigma}}_b^k\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m)^T}\hat{\boldsymbol{\Sigma}}_b^k\boldsymbol{\beta}^{(m)} - P_k(\boldsymbol{\beta}_k) \tag{15}$$

minorizes the objective of (13) at $\boldsymbol{\beta}^{(m)}$. Moreover, since P_k is a convex function, $g(\boldsymbol{\beta}_k|\boldsymbol{\beta}^{(m)})$ is concave in $\boldsymbol{\beta}_k$ and hence can be maximized using convex optimization tools. We can use (15) as the basis for a minorization algorithm to find the kth penalized discriminant vector. The algorithm assumes that the first $k-1$ penalized discriminant vectors have already been computed.

Algorithm 1: Obtaining the kth penalized discriminant vector.

(a) If $k > 1$, define an orthogonal projection matrix $\mathbf{P}_k^{\perp}$ that projects onto the space that is orthogonal to $(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}}_i$ for all $i < k$. Let $\mathbf{P}_1^{\perp} = \mathbf{I}$.
(b) Let $\hat{\boldsymbol{\Sigma}}_b^k = \frac{1}{n}\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{P}_k^{\perp}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{X}$. Note that $\hat{\boldsymbol{\Sigma}}_b^1 = \hat{\boldsymbol{\Sigma}}_b$.
(c) Let $\boldsymbol{\beta}_k^{(0)}$ be the first eigenvector of $\tilde{\boldsymbol{\Sigma}}_w^{-1}\hat{\boldsymbol{\Sigma}}_b^k$.
(d) For $m = 1, 2, \ldots$ until convergence: let $\boldsymbol{\beta}_k^{(m)}$ be the solution to

$$\text{maximize}_{\boldsymbol{\beta}_k} \{2\boldsymbol{\beta}_k^T\hat{\boldsymbol{\Sigma}}_b^k\boldsymbol{\beta}_k^{(m-1)} - P_k(\boldsymbol{\beta}_k)\} \text{ subject to } \boldsymbol{\beta}_k^T\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta}_k \leq 1. \tag{16}$$

Let $\hat{\boldsymbol{\beta}}_k$ denote the solution at convergence.

Of course, the solution to (16) will depend on the form of the convex function P_k. In the next section, we will consider two specific forms for P_k.

Once the penalized discriminant vectors have been computed, classification is straightforward: as in the case of classical LDA, we compute $\mathbf{X}\hat{\boldsymbol{\beta}}_1, \ldots, \mathbf{X}\hat{\boldsymbol{\beta}}_{K-1}$ and assign each observation to its nearest centroid in this transformed space. To perform reduced rank classification, we transform the observations using only the first $k < K-1$ penalized discriminant vectors.

4.2. Penalized LDA-L_1 and penalized LDA-FL

4.2.1. Penalized LDA-L_1

We define *penalized LDA-L_1* to be the solution to (13) with an L_1 penalty,

$$\text{maximize}_{\boldsymbol{\beta}_k}\{\boldsymbol{\beta}_k^T\hat{\boldsymbol{\Sigma}}_b^k\boldsymbol{\beta}_k - \lambda_k\sum_{j=1}^{p}|\hat{\sigma}_j\beta_{kj}|\} \text{ subject to } \boldsymbol{\beta}_k^T\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta}_k \le 1. \tag{17}$$

When the tuning parameter λ_k is large, some elements of the solution $\hat{\boldsymbol{\beta}}_k$ will be exactly equal to zero. In (17), $\hat{\sigma}_j$ is the within-class standard deviation for feature j; the inclusion of $\hat{\sigma}_j$ in the penalty has the effect that features that vary more within each class undergo greater penalization. Penalized LDA-L_1 is appropriate if we want to obtain a sparse classifier - that is, a classifier for which the decision rule involves only a subset of the features. In particular, the resulting discriminant vectors are sparse, so penalized LDA-L_1 amounts to projecting the data onto a low-dimensional subspace that involves only a subset of the features.

To solve (17), we use the minorization approach outlined in Algorithm 1. Step (d) can be written as

$$\text{maximize}_{\boldsymbol{\beta}_k}\{2\boldsymbol{\beta}_k^T\hat{\boldsymbol{\Sigma}}_b^k\boldsymbol{\beta}_k^{(m-1)} - \lambda_k\sum_{j=1}^{p}|\hat{\sigma}_j\beta_{kj}|\} \text{ subject to } \boldsymbol{\beta}_k^T\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta}_k \le 1. \tag{18}$$

The solution to (18) is given in Proposition 2 in Section 4.2.3.

4.2.2. Penalized LDA-FL

We define *penalized LDA-FL* to be the solution to the problem (13) with a fused lasso penalty (Tibshirani et al., 2005):

$$\begin{aligned}&\text{maximize}_{\boldsymbol{\beta}_k}\{\boldsymbol{\beta}_k^T\hat{\boldsymbol{\Sigma}}_b^k\boldsymbol{\beta}_k - \lambda_k\sum_{j=1}^{p}|\hat{\sigma}_j\beta_{kj}| - \gamma_k\sum_{j=2}^{p}|\hat{\sigma}_j\beta_{kj} - \hat{\sigma}_{j-1}\beta_{k,j-1}|\}\\&\text{subject to } \boldsymbol{\beta}_k^T\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta}_k \le 1.\end{aligned} \tag{19}$$

When the nonnegative tuning parameter λ_k is large then the resulting discriminant vector will be sparse in the features, and when the nonnegative tuning parameter γ_k is large then the discriminant vector will be piecewise constant. This classifier is appropriate if the features are ordered on a line, and one believes that the true underlying signal is sparse and piecewise constant.

To solve (13), we again apply Algorithm 1. Step (d) can be written as

$$\begin{aligned}&\text{maximize}_{\boldsymbol{\beta}_k}\{2\boldsymbol{\beta}_k^T\hat{\boldsymbol{\Sigma}}_b^k\boldsymbol{\beta}_k^{(m-1)} - \lambda_k\sum_{j=1}^{p}|\hat{\sigma}_j\beta_{kj}| - \gamma_k\sum_{j=2}^{p}|\hat{\sigma}_j\beta_{kj} - \hat{\sigma}_{j-1}\beta_{k,j-1}|\}\\&\text{subject to } \boldsymbol{\beta}_k^T\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta}_k \le 1.\end{aligned} \tag{20}$$

Proposition 2 in Section 4.2.3 provides the solution to (20).

4.2.3. The minorization step for penalized LDA-L_1 and penalized LDA-FL
Now we present Proposition 2, which provides a solution to (18) and (20). In other words, Proposition 2 provides details for performing Step (d) in Algorithm 1 for penalized LDA-L_1 and penalized LDA-FL.

Proposition 2. *(a) To solve (18), we first solve the problem*

$$\text{minimize}_{\mathbf{d}\in\mathbb{R}^p}\{\mathbf{d}^T\tilde{\mathbf{\Sigma}}_w\mathbf{d} - 2\mathbf{d}^T\hat{\mathbf{\Sigma}}_b^k\boldsymbol{\beta}_k^{(m-1)} + \lambda_k\sum_j|\hat{\sigma}_j d_j|\}. \tag{21}$$

If $\hat{\mathbf{d}} = 0$ then $\hat{\boldsymbol{\beta}}_k = 0$. Otherwise, $\hat{\boldsymbol{\beta}}_k = \hat{\mathbf{d}}/\sqrt{\hat{\mathbf{d}}^T\tilde{\mathbf{\Sigma}}_w\hat{\mathbf{d}}}$.

(b) To solve (20), we first solve the problem

$$\text{minimize}_{\mathbf{d}\in\mathbb{R}^p}\{\mathbf{d}^T\tilde{\mathbf{\Sigma}}_w\mathbf{d}-2\mathbf{d}^T\hat{\mathbf{\Sigma}}_b^k\boldsymbol{\beta}_k^{(m-1)}+\lambda_k\sum_{j=1}^p|\hat{\sigma}_j d_j|+\gamma_k\sum_{j=2}^p|\hat{\sigma}_j d_j-\hat{\sigma}_{j-1}d_{j-1}|\}. \tag{22}$$

If $\hat{\mathbf{d}} = 0$ then $\hat{\boldsymbol{\beta}}_k = 0$. Otherwise, $\hat{\boldsymbol{\beta}}_k = \hat{\mathbf{d}}/\sqrt{\hat{\mathbf{d}}^T\tilde{\mathbf{\Sigma}}_w\hat{\mathbf{d}}}$.

The proof is given in the Appendix. Some comments on Proposition 2 are as follows:

- If $\tilde{\mathbf{\Sigma}}_w$ is the diagonal estimate (4), then the solution to (21) is

$$\hat{d}_j = \frac{1}{\hat{\sigma}_j^2}S((\hat{\mathbf{\Sigma}}_b^k\boldsymbol{\beta}_k^{(m-1)})_j, \lambda_k\hat{\sigma}_j/2) \tag{23}$$

where S is the soft-thresholding operator, defined as

$$S(x, a) = \text{sign}(x)(|x| - a)_+ \tag{24}$$

and applied componentwise. To see why, note that differentiating (21) with respect to d_j indicates that the solution will satisfy

$$2\hat{\sigma}_j^2 d_j = 2(\hat{\mathbf{\Sigma}}_b^k\boldsymbol{\beta}_k^{(m-1)})_j - \lambda_k\hat{\sigma}_j\Gamma_j, \tag{25}$$

where Γ_j is the subgradient of $|d_j|$, defined as

$$\Gamma_j = \begin{cases} 1 & \text{if } d_j > 0 \\ -1 & \text{if } d_j < 0 \\ a & \text{if } d_j = 0 \end{cases} \tag{26}$$

where a is some number between 1 and -1. Then (23) follows from (25).

- On the other hand, if $\tilde{\mathbf{\Sigma}}_w$ is a non-diagonal positive definite estimate of $\mathbf{\Sigma}_w$, then one can solve (21) by coordinate descent (see e.g. Friedman et al., 2007). (21) is in that case closely related to the lasso, but may involve more demanding computations. This is due to the fact that when $p \gg n$ the standard lasso can be implemented by storing the $n\times p$ matrix $\mathbf{X}$ rather than the entire $p\times p$ matrix $\mathbf{X}^T\mathbf{X}$. But if $\tilde{\mathbf{\Sigma}}_w$ is a $p \times p$ matrix without special structure then one must store it in full in order to solve (21).

- If $\tilde{\mathbf{\Sigma}}_w$ is a diagonal estimate for $\mathbf{\Sigma}_w$ then (22) is a diagonal fused lasso problem, for which fast algorithms have been proposed (see e.g. Hoefling, 2009; Johnson, 2010).

Table 1. Timing results for penalized LDA-L_1 (with $\lambda = 0.005$) and penalized LDA-FL (with $\lambda = \gamma = 0.005$) for various values of n and p, with 4-class data. Mean (and standard error) of running time (in seconds), over 25 repetitions. The diagonal estimate (4) was used for $\tilde{\boldsymbol{\Sigma}}_w$.

		p=20	p=200	p=2000	p=20000
Penalized LDA-L_1	n=20	0.049(0)	0.059(0.002)	0.199(0.022)	5.1(0.851)
	n=200	0.062(0)	0.147(0.001)	1.182(0.014)	11.835(0.417)
Penalized LDA-FL	n=20	0.064(0.003)	0.108(0.007)	1.018(0.102)	118.61(9.915)
	n=200	0.075(0.001)	0.219(0.012)	1.835(0.102)	118.557(8.895)

4.2.4. Comments on tuning parameter selection

We now consider the problem of selecting the tuning parameter λ_k for the penalized LDA-L_1 problem (17). The simplest approach would be to take $\lambda_k = \lambda$, i.e. the same tuning parameter value for all components. However, this results in effectively penalizing each component more than the previous components, since the unpenalized objective value of (17), which is equal to the largest eigenvalue of $\tilde{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_b^k\tilde{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}}$, is nonincreasing in k. So instead, we take the following approach. We first fix a nonnegative constant λ, and then we take $\lambda_k = \lambda||\tilde{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_b^k\tilde{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}}||$ where $||\cdot||$ indicates the largest eigenvalue. Note that when $p \gg n$, this largest eigenvalue can be quickly computed using the fact that $\hat{\boldsymbol{\Sigma}}_b^k$ has low rank. The value of λ can be chosen by cross-validation.

In the case of the penalized LDA-FL problem (19), instead of choosing λ_k and γ_k directly, we instead fix nonnegative constants λ and γ. Then, we take $\lambda_k = \lambda||\tilde{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_b^k\tilde{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}}||$ and $\gamma_k = \gamma||\tilde{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_b^k\tilde{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}}||$. λ and γ can be chosen by cross-validation.

4.2.5. Timing results for penalized LDA

We now comment on the computations involved in the algorithms proposed earlier in this section. We used a very simple simulation corresponding to no signal in the data: $X_{ij} \sim N(0,1)$ and there were four equally-sized classes. Table 1 summarizes the computational times required to perform penalized LDA-L_1 and penalized LDA-FL with the diagonal estimate (4) used for $\tilde{\boldsymbol{\Sigma}}_w$. The `R` library `penalizedLDA` was used. Timing depends critically on the convergence criterion used; we determine that the algorithm has “converged” when subsequent iterations lead to a relative improvement in the objective of no more than 10^{-6}; that is, $|r_i - r_{i+1}|/r_{i+1} < 10^{-6}$ where r_i is the objective obtained at the ith iteration. Of course, computational times will be shorter if a less strict convergence threshold is used. All timings were carried out on a AMD Opteron 848 2.20 GHz processor.

4.3. Recasting penalized LDA as a biconvex problem

Rather than using a minorization approach to solve the nonconvex problem (12), one could instead recast it as a biconvex problem. Consider the problem

$$\text{maximize}_{\boldsymbol{\beta},\mathbf{u}}\{\frac{2}{\sqrt{n}}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{u} - P(\boldsymbol{\beta}) - \mathbf{u}^T\mathbf{u}\} \text{ subject to } \boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta} \leq 1. \quad (27)$$

Partially optimizing (27) with respect to $\mathbf{u}$ reveals that the $\boldsymbol{\beta}$ that solves it also solves (12). Moreover, (27) is a *biconvex* problem (see e.g. Gorski et al., 2007): that is, with $\boldsymbol{\beta}$ held fixed, it is convex in $\mathbf{u}$, and with $\mathbf{u}$ held fixed, it is convex in $\boldsymbol{\beta}$. This suggests a a simple iterative approach for solving it.

Algorithm 4: A biconvex formulation for penalized LDA.

(a) Let $\boldsymbol{\beta}^{(0)}$ be the first eigenvector of $\tilde{\boldsymbol{\Sigma}}_w^{-1}\hat{\boldsymbol{\Sigma}}_b$.

(b) For $m = 1, 2, \ldots$ until convergence:

(i) Let $\mathbf{u}^{(m)}$ solve

$$\text{maximize}_{\mathbf{u}}\{\frac{2}{\sqrt{n}}\boldsymbol{\beta}^{(m-1)^T}\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{u} - \mathbf{u}^T\mathbf{u}\}. \tag{28}$$

(ii) Let $\boldsymbol{\beta}^{(m)}$ solve

$$\text{maximize}_{\boldsymbol{\beta}}\{\frac{2}{\sqrt{n}}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{u}^{(m)} - P(\boldsymbol{\beta})\} \text{ subject to } \boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta} \leq 1. \tag{29}$$

Combining Steps (b)(i) and (b)(ii), we see that $\boldsymbol{\beta}^{(m)}$ solves

$$\text{maximize}_{\boldsymbol{\beta}}\{2\boldsymbol{\beta}^T\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta}^{(m-1)} - P(\boldsymbol{\beta})\} \text{ subject to } \boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta} \leq 1. \tag{30}$$

Comparing (30) to (16), we see that the biconvex formulation (27) results in the same update step as the minorization approach outlined in Algorithm 1. This biconvex formulation is very closely related to the sparse principal components analysis proposal of Witten et al. (2009), which corresponds to the case where $\tilde{\boldsymbol{\Sigma}}_w = \mathbf{I}$ and a bound form is used for the penalty $P(\boldsymbol{\beta})$. Since $\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}$ is a weighted version of the class centroid matrix, our penalized LDA proposal is closely related to performing sparse principal components analysis on the class centroids matrix.

5. Examples

5.1. Methods included in comparisons

In the examples that follow, penalized LDA-L_1 and penalized LDA-FL were performed using the diagonal estimate (4) for $\tilde{\boldsymbol{\Sigma}}_w$, as implemented in the `R` package `penalizedLDA`. The nearest shrunken centroids (NSC; Tibshirani et al., 2002, 2003) method was performed using the `R` package `pamr`, and the shrunken centroids regularized discriminant analysis (RDA; Guo et al., 2007) method was performed using the `rda` `R` package. Briefly, NSC results from using a diagonal estimate of $\boldsymbol{\Sigma}_w$ and imposing L_1 penalties on the class mean vectors under the normal model, and RDA combines a ridge-type penalty in estimating $\boldsymbol{\Sigma}_w$ with soft-thresholding of $\tilde{\boldsymbol{\Sigma}}_w^{-1}\hat{\boldsymbol{\mu}}_k$. These methods are discussed further in Section 6.

The tuning parameters for each of the methods considered were as follows. For penalized LDA-L_1, λ described in Section 4.2.4 was a tuning parameter. For penalized LDA-FL, we treated $\lambda = \gamma$ (see Section 4.2.4) as a single tuning parameter in order to avoid performing tuning parameter selection on a two-dimensional grid. Moreover, penalized LDA had an additional tuning parameter, the number of discriminant vectors to include in the classifier. NSC has a single tuning parameter, which corresponds to the amount of soft-thresholding performed. RDA has two tuning parameters, one of which controls the number of features used and the other controls the ridge penalty used to regularize the estimate of $\mathbf{\Sigma}_w$.

5.2. A simulation study

We compare penalized LDA to NSC and RDA in a simulation study. Four simulations were considered. In each simulation, there are 1200 observations, equally split between the classes. Of these 1200 observations, 100 belong to the training set, 100 belong to the validation set, and 1000 are in the test set. Each simulation consists of measurements on 500 features, of which 100 differ between classes.

> **Simulation 1. Mean shift with independent features.** There are four classes. If observation i is in class k, then $\mathbf{x}_i \sim N(\boldsymbol{\mu}_k, \mathbf{I})$, where $\mu_{1j} = 0.7 \times 1_{(1 \le j \le 25)}, \mu_{2j} = 0.7 \times 1_{(26 \le j \le 50)}, \mu_{3j} = 0.7 \times 1_{(51 \le j \le 75)}, \mu_{4j} = 0.7 \times 1_{(75 \le j \le 100)}$.
>
> **Simulation 2. Mean shift with dependent features.** There are two classes. For $i \in C_1$, $\mathbf{x}_i \sim N(0, \mathbf{\Sigma})$ and for $i \in C_2$, $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$, $\mu_j = 0.6 \times 1_{(j \le 200)}$. The covariance structure is block diagonal, with 5 blocks each of dimension 100×100. The blocks have (j, j') element $0.6^{|j-j'|}$. This covariance structure is intended to mimic gene expression data, in which genes are positively correlated within a pathway and independent between pathways.
>
> **Simulation 3. One-dimensional mean shift with independent features.** There are four classes, and the features are independent. For $i \in C_k$, $X_{ij} \sim N(\frac{k-1}{3}, 1)$ if $j \le 100$, and $X_{ij} \sim N(0,1)$ otherwise. Note that a one-dimensional projection of the data fully captures the class structure.
>
> **Simulation 4. Mean shift with independent features and no linear ordering.** There are four classes. If observation i is in class k, then $\mathbf{x}_i \sim N(\boldsymbol{\mu}_k, \mathbf{I})$. The mean vectors are defined as follows: $\mu_{1j} \sim N(0, 0.3^2)$ if $1 \le j \le 25$ and $\mu_{1j} = 0$ otherwise, $\mu_{2j} \sim N(0, 0.3^2)$ if $26 \le j \le 50$ and $\mu_{2j} = 0$ otherwise, $\mu_{3j} \sim N(0, 0.3^2)$ if $51 \le j \le 75$ and $\mu_{3j} = 0$ otherwise, $\mu_{4j} \sim N(0, 0.3^2)$ if $75 \le j \le 100$ and $\mu_{4j} = 0$ otherwise.

Figure 1 displays the class mean vectors for each simulation.

For each method, models were fit on the training set using a range of tuning parameter values. Tuning parameter values were then selected to minimize the validation set error. Finally, the training set models with appropriate tuning parameter values were evaluated on the test set. Penalized LDA-FL was performed in Simulations 1-3 but not in Simulation 4, since in Simulation 4 the features do not have a linear ordering as assumed by the fused lasso penalty (see Figure 1).

Test set errors and the numbers of nonzero features used are reported in Table 2. For penalized LDA, the numbers of discriminant vectors used are also reported. Penalized LDA-FL has by far the best performance in the first three simulations, since it exploits the fact that the important features have a linear ordering. Of

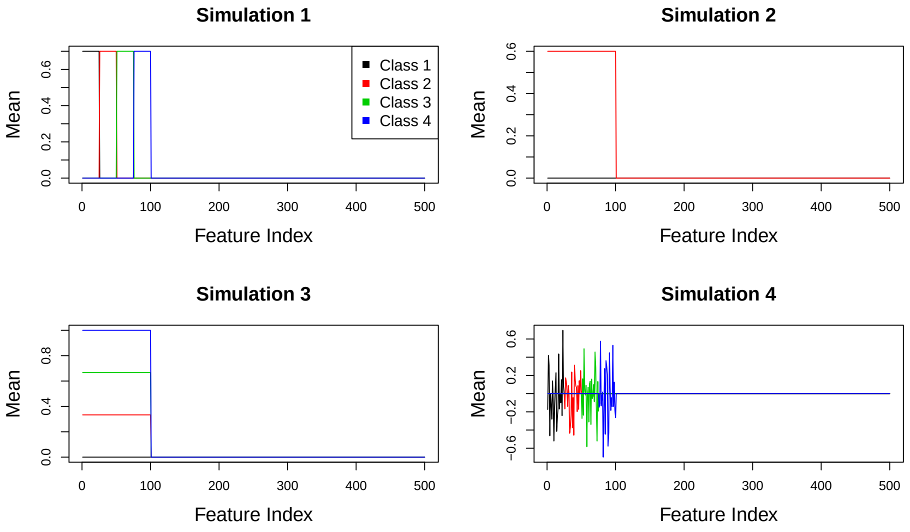

Fig. 1. Class mean vectors for each simulation.

course, in real data applications, penalized LDA-FL can only be applied if such an ordering is present. Note that penalized LDA tends to use fewer than three components in Simulation 3, in which a one-dimensional projection is sufficient to explain the class structure.

5.3. Application to gene expression data

We compare penalized LDA-L_1, NSC, and RDA on three gene expression data sets:

> **Ramaswamy data**. A data set consisting of 16,063 gene expression measurements and 198 samples belonging to 14 distinct cancer subtypes (Ramaswamy

Table 2. Simulation results. Mean (and standard errors), computed over 25 repetitions, of test set errors, number of nonzero features, and number of discriminant vectors used.

		Pen. LDA-L_1	Pen. LDA-FL	NSC	RDA
Sim 1	Errors	117.48(3)	38.4(2)	88.96(2.6)	96.8(3.4)
	Features	301.16(20.1)	159.28(15.8)	290.28(16.7)	226.6(15.7)
	Components	3(0)	3(0)	-	-
Sim 2	Errors	90.04(2.8)	77(1.9)	88.44(2.7)	112.2(5.8)
	Features	229.36(20.4)	170.16(18.4)	341.28(24.8)	414.84(32.6)
	Components	1(0)	1(0)	-	-
Sim 3	Errors	150.8(5.4)	83.44(2.3)	276.64(4)	291(4.8)
	Features	147.84(7.1)	115.92(9.1)	439.6(10.7)	349.32(24.5)
	Components	1(0)	1(0)	-	-
Sim 4	Errors	60.56(1.1)	-	58.28(1.2)	57(0.9)
	Features	311.4(22.1)	-	135.4(22.6)	98(7.3)
	Components	3(0)	-	-	-

Table 3. Results obtained on gene expression data over 10 training/test set splits. Quantities reported are the mean (and standard deviation) of test set errors and nonzero coefficients.

		NSC	Penalized LDA-L_1	RDA
Ramaswamy	Errors	16.3(4.16)	18.8(3.05)	24(17.45)
	Features	2336.9(2292.03)	14873.5(720.29)	5022.5(2503.35)
Nakayama	Errors	4.2(2.15)	4.4(1.51)	2.8(1.23)
	Features	5908(7131.5)	10478.7(2116.27)	22283(0)
Sun	Errors	15(4.29)	15.2(3.29)	15.7(4.52)
	Features	30004.9(18557.68)	21634.8(7443.21)	54183.4(693.23)

et al., 2001). The data set has been studied in a number of papers (see e.g. Zhu and Hastie, 2004; Guo et al., 2007; Witten and Tibshirani, 2009) and is available from `http://www-stat.stanford.edu/~hastie/glmnet/glmnetData/`.

Nakayama data. A data set consisting of 105 samples from 10 types of soft tissue tumors, each with 22,283 gene expression measurements (Nakayama et al., 2007). We limited the analysis to five tumor types for which at least 15 samples were present in the data; the resulting subset of the data contained 86 samples. The data are available on Gene Expression Omnibus (Barrett et al., 2005) with accession number GDS2736.

Sun data. A data set consisting of 180 samples and 54,613 expression measurements (Sun et al., 2006). The samples fall into four classes: one non-tumor class and three types of glioma. The data are available on Gene Expression Omnibus with accession number GDS1962.

Each data set was split into a training set containing 75% of the samples and a test set containing 25% of the samples. Cross-validation was performed on the training set and test set error rates were evaluated. The process was repeated ten times, each with a random choice of training set and test set. Results are reported in Table 3. The results suggest that the three methods tend to have roughly comparable performance. A reviewer pointed out that there is substantial variability in the number of features used by each classifier across each training/test set split. Indeed, this instability in the set of genes selected likely reflects the fact that in the analysis of many real data types, sparsity is simply an approximation, rather than a property that we expect to hold exactly.

Penalized LDA-L_1 has the added advantage over RDA and NSC of yielding penalized discriminant vectors that can be used to visualize the observations, as in Figure 2.

6. The normal model, optimal scoring, and extensions to high dimensions

In this section, we review the normal model and the optimal scoring problem, which lead to the same classification rule as Fisher's discriminant problem. We also review past extensions of LDA to the high-dimensional setting.

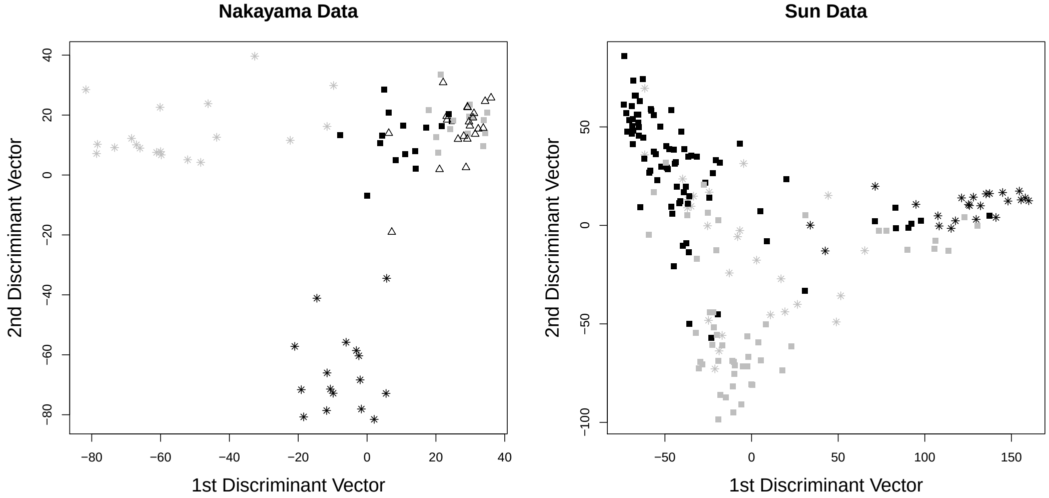

Fig. 2. For the Nakayama and Sun data, the samples were projected onto the first two penalized discriminant vectors. The samples in each class are shown using a distinct symbol.

6.1. The normal model

Suppose that the observations are independent and normally distributed with a common within-class covariance matrix $\mathbf{\Sigma}_w \in \mathbb{R}^{p \times p}$ and a class-specific mean vector $\boldsymbol{\mu}_k \in \mathbb{R}^p$. The log likelihood under this model is

$$\sum_{k=1}^{K} \sum_{i \in C_k} \{-\frac{1}{2} \log |\mathbf{\Sigma}_w| - \frac{1}{2} \mathrm{tr}[\mathbf{\Sigma}_w^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T]\} + c \tag{31}$$

where c is a constant. If the classes have equal prior probabilities, then by Bayes' theorem, a new observation $\mathbf{x}$ is assigned to the class for which the discriminant function

$$\delta_k(\mathbf{x}) = \mathbf{x}^T \hat{\mathbf{\Sigma}}_w^{-1} \hat{\boldsymbol{\mu}}_k - \frac{1}{2} \hat{\boldsymbol{\mu}}_k^T \hat{\mathbf{\Sigma}}_w^{-1} \hat{\boldsymbol{\mu}}_k \tag{32}$$

is maximal. One can show that this is the same as the classification rule obtained from Fisher's discriminant problem.

6.2. The optimal scoring problem

Let $\mathbf{Y}$ be a $n \times K$ matrix, with $Y_{ik} = 1_{i \in C_k}$. Then, optimal scoring involves sequentially solving

$$\begin{aligned} &\mathrm{minimize}_{\boldsymbol{\beta}_k \in \mathbb{R}^p, \boldsymbol{\theta}_k \in \mathbb{R}^K} \{\frac{1}{n} ||\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k||^2\} \\ &\text{subject to } \boldsymbol{\theta}_k^T \mathbf{Y}^T \mathbf{Y} \boldsymbol{\theta}_k = 1, \boldsymbol{\theta}_k^T \mathbf{Y}^T \mathbf{Y} \boldsymbol{\theta}_i = 0 \ \ \forall i < k \end{aligned} \tag{33}$$

for $k = 1, \ldots, K-1$. This amounts to recasting the classification problem as a regression problem, where a quantitative coding $\boldsymbol{\theta}_k$ of the K classes must be chosen along with the regression coefficient vector $\boldsymbol{\beta}_k$. The solution $\hat{\boldsymbol{\beta}}_k$ to (33) is proportional to the solution to (3). Somewhat involved proofs of this fact are given in Breiman and Ihaka (1984) and Hastie et al. (1995). We present a simpler proof in the Appendix.

6.3. LDA in high dimensions

In recent years, a number of authors have proposed extensions of LDA to the high-dimensional setting in order to achieve sparsity (Tibshirani et al., 2002, 2003; Guo et al., 2007; Trendafilov and Jolliffe, 2007; Grosenick et al., 2008; Leng, 2008; Fan and Fan, 2008; Shao et al., 2011; Clemmensen et al., 2011). In Section 4, we proposed penalizing Fisher's discriminant problem. Here we briefly review some past proposals that have involved penalizing the log likelihood under the normal model, and the optimal scoring problem.

The *nearest shrunken centroids (NSC)* proposal (Tibshirani et al., 2002, 2003) assigns an observation $\mathbf{x}^*$ to the class that minimizes

$$\sum_{j=1}^{p} \frac{(x_j^* - \bar{\mu}_{kj})^2}{\hat{\sigma}_j^2}, \tag{34}$$

where $\bar{\mu}_{kj} = S(\hat{\mu}_{kj}, \lambda\hat{\sigma}_j\sqrt{\frac{1}{n_k} + \frac{1}{n}})$, S is the soft-thresholding operator (24), and we have assumed equal prior probabilities for each class. This classification rule approximately follows from estimating the class mean vectors via maximization of an L_1-penalized version of the log likelihood (31), and assuming independence of the features (Hastie et al., 2009). The *shrunken centroids regularized discriminant analysis (RDA)* proposal (Guo et al., 2007) arises instead from applying the normal model approach with covariance matrix $\tilde{\boldsymbol{\Sigma}}_w = \hat{\boldsymbol{\Sigma}}_w + \rho\mathbf{I}$ and performing soft-thresholding in order to obtain a classifier that is sparse in the features.

Several authors have proposed penalizing the optimal scoring criterion (33) by imposing penalties on $\boldsymbol{\beta}_k$ (see e.g. Grosenick et al., 2008; Leng, 2008). For instance, the *sparse discriminant analysis* (SDA) proposal (Clemmensen et al., 2011) involves sequentially solving

$$\begin{aligned} &\text{minimize}_{\boldsymbol{\beta}_k,\boldsymbol{\theta}_k}\{\frac{1}{n}||\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k||^2 + \boldsymbol{\beta}_k^T\boldsymbol{\Omega}\boldsymbol{\beta}_k + \lambda||\boldsymbol{\beta}_k||_1\} \\ &\text{subject to } \boldsymbol{\theta}_k^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta}_k = 1, \boldsymbol{\theta}_k^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta}_i = 0 \ \ \forall i < k \end{aligned} \tag{35}$$

where λ is a nonnegative tuning parameter and $\boldsymbol{\Omega}$ is a positive definite penalization matrix. If $\boldsymbol{\Omega} = \gamma\mathbf{I}$ for $\gamma > 0$, then this is an elastic net penalty (Zou and Hastie, 2005). The resulting discriminant vectors will be sparse if λ is sufficiently large. If $\lambda = 0$, then this reduces to the *penalized discriminant analysis* proposal of Hastie et al. (1995). The criterion (35) can be optimized in a simple iterative fashion: we optimize with respect to $\boldsymbol{\beta}_k$ holding $\boldsymbol{\theta}_k$ fixed, and we optimize with respect to $\boldsymbol{\theta}_k$ holding $\boldsymbol{\beta}_k$ fixed. In fact, if any convex penalties are applied to the discriminant vectors in the optimal scoring criterion (33), then an iterative approach can be developed that decreases the objective at each step. However, the optimal scoring problem is a somewhat indirect formulation for LDA.

Table 4. Advantages and disadvantages of using the normal model (NM), optimal scoring (OS), and Fisher's discriminant analysis (FD) as the basis for penalized LDA with an L_1 penalty.

	Advantages	Disadvantages	Citation
NM	Sparse class means if diagonal estimate of $\mathbf{\Sigma}_w$ used. Computations are fast.	Does not give sparse discriminant vectors. No reduced-rank classification.	Tibshirani et al. (2002)
OS	Sparse discriminant vectors.	Difficult to enforce diagonal estimate for $\mathbf{\Sigma}_w$, which is useful if $p > n$. Computations can be slow.	Grosenick et al. (2008) Leng (2008) Clemmensen et al. (2011)
FD	Sparse discriminant vectors. Simple to enforce diagonal estimate of $\mathbf{\Sigma}_w$. Computations are fast using diagonal estimate of $\mathbf{\Sigma}_w$.	Computations can be slow when p is large, unless diagonal estimate of $\mathbf{\Sigma}_w$ is used.	This work.

Our penalized LDA proposal is instead a direct extension of Fisher's discriminant problem (3). Trendafilov and Jolliffe (2007) consider a problem very similar to penalized LDA-L_1. But they discuss only the $p < n$ case. Their algorithm is more complex than ours, and does not extend to general convex penalty functions.

A summary of proposals that extend LDA to the high-dimensional setting through the use of L_1 penalties is given in Table 4. In the next section, we will explore how our penalized LDA-L_1 proposal relates to the NSC and SDA methods.

7. Connections with existing methods

7.1. Connection with sparse discriminant analysis

Consider the SDA criterion (35) with $k = 1$. We drop the subscripts on $\boldsymbol{\beta}_1$ and $\boldsymbol{\theta}_1$ for convenience. Partially optimizing (35) with respect to $\boldsymbol{\theta}$ reveals that for any $\boldsymbol{\beta}$ for which $\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} \neq 0$, the optimal $\boldsymbol{\theta}$ equals $\frac{(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta}}}$. So (35) can be rewritten as

$$\text{maximize}_{\boldsymbol{\beta}}\{\frac{2}{\sqrt{n}}\sqrt{\boldsymbol{\beta}^T\hat{\mathbf{\Sigma}}_b\boldsymbol{\beta}} - \boldsymbol{\beta}^T(\hat{\mathbf{\Sigma}}_b + \hat{\mathbf{\Sigma}}_w + \mathbf{\Omega})\boldsymbol{\beta} - \lambda||\boldsymbol{\beta}||_1\}. \quad (36)$$

Assume that each feature has been standardized to have within-class standard deviation equal to 1. Take $\tilde{\mathbf{\Sigma}}_w = \hat{\mathbf{\Sigma}}_w + \mathbf{\Omega}$, where $\mathbf{\Omega}$ is chosen so that $\tilde{\mathbf{\Sigma}}_w$ is positive definite. Then, the following proposition holds.

Proposition 3. *Consider the penalized LDA-L_1 problem (17) where $\lambda_1 > 0$ and $k = 1$. Suppose that at the solution $\boldsymbol{\beta}^*$ to (17), the objective is positive. Then, there exists a positive tuning parameter λ_2 and a positive scalar c such that $c\boldsymbol{\beta}^*$ corresponds to a zero of the generalized gradient of the SDA objective (36).*

A proof is given in the Appendix. Note that the assumption that the objective is positive at the solution $\boldsymbol{\beta}^*$ is not very taxing - it simply means that $\boldsymbol{\beta}^*$ results in a higher value of the objective than does a vector of zeros. Proposition 3 states that if the same positive definite estimate for $\mathbf{\Sigma}_w$ is used for both problems, then the solution of the penalized LDA-L_1 problem corresponds to a point where the generalized gradient of the SDA problem is zero. But since the SDA problem is not convex, this does not imply that there is a correspondence between the solutions of the two problems. Penalized LDA-L_1 has some advantages over SDA. Unlike SDA, penalized LDA-L_1 has a clear relationship with Fisher's discriminant problem. Moreover, unlike SDA, it provides a natural way to enforce a diagonal estimate of $\mathbf{\Sigma}_w$.

7.2. Connection with nearest shrunken centroids

The following proposition indicates that in the case of two equally-sized classes, NSC is closely related to the penalized LDA-L_1 problem with the diagonal estimate (4) for $\mathbf{\Sigma}_w$.

Proposition 4. *Suppose that $K = 2$ and $n_1 = n_2 = \frac{n}{2}$. Let $\hat{\boldsymbol{\beta}}$ denote the solution to the problem*

$$\text{maximize}_{\boldsymbol{\beta}}\{\sqrt{\boldsymbol{\beta}^T\hat{\mathbf{\Sigma}}_b\boldsymbol{\beta}} - \lambda\sum_{j=1}^{p}|\beta_j\hat{\sigma}_j|\} \text{ subject to } \boldsymbol{\beta}^T\tilde{\mathbf{\Sigma}}_w\boldsymbol{\beta} \leq 1 \qquad (37)$$

where $\tilde{\mathbf{\Sigma}}_w$ is the diagonal estimate (4). Consider the classification rule obtained by computing $\mathbf{X}\hat{\boldsymbol{\beta}}$ and assigning each observation to its nearest centroid in this transformed space. This is the same as the NSC classification rule (34).

Note that (37) is simply a modified version of the penalized LDA-L_1 criterion, in which the between-class variance term has been replaced with its square root. Therefore, penalized LDA-L_1 with a diagonal estimate of $\mathbf{\Sigma}_w$ and NSC are closely connected when $K = 2$. This connection does not hold for larger values of K, since NSC penalizes the elements of the $p \times K$ class centroid matrix, whereas penalized LDA-L_1 penalizes the eigenvectors of this matrix. A proof of Proposition 4 is given in the Appendix.

8. Discussion

We have extended Fisher's discriminant problem to the high-dimensional setting by imposing penalties on the discriminant vectors. The penalty function is chosen based on the problem at hand, and can result in an interpretable classifier. A potentially useful but unexplored area of application for our proposal is fMRI data, for which one could use a penalty that incorporates the spatial structure of the voxels.

There is a strong connection between our penalized LDA proposal and previous work on penalized principal components analysis (PCA). When P_k is an L_1 penalty, (12) is closely related to the *SCoTLASS* proposal for sparse PCA (Jolliffe et al., 2003). The criterion (12) and Algorithm 1 for optimizing it are closely related to the penalized principal components algorithms considered by a number of authors (see e.g. Zou et al., 2006; Shen and Huang, 2008; Witten et al., 2009). This connection stems from the fact that Fisher's discriminant problem is simply a generalized eigenproblem.

The R language software package penalizedLDA implementing penalized LDA-L_1 and penalized LDA-FL will be made available on CRAN, http://cran.r-project.org/.

Acknowledgments

We thank two anonymous reviewers for helpful suggestions, and we thank Line Clemmensen for responses to our inquiries. Trevor Hastie provided helpful comments that improved the quality of this manuscript. Robert Tibshirani was partially supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183.

Appendix

Equivalence between (3) and standard formulation for LDA

We have stated Fisher's discriminant problem as (3), but a more standard formulation is

$$\text{maximize}_{\boldsymbol{\beta}_k \in \mathbb{R}^p} \{\boldsymbol{\beta}_k^T \hat{\boldsymbol{\Sigma}}_b \boldsymbol{\beta}_k\} \text{ subject to } \boldsymbol{\beta}_k^T \hat{\boldsymbol{\Sigma}}_w \boldsymbol{\beta}_k = 1, \boldsymbol{\beta}_k^T \hat{\boldsymbol{\Sigma}}_w \boldsymbol{\beta}_i = 0 \;\; \forall i < k. \tag{38}$$

We now show that (3) and (38) are equivalent, provided that the solution is not in the null space of $\hat{\boldsymbol{\Sigma}}_b$. It suffices to show that if $\boldsymbol{\alpha}$ solves (3), then $\boldsymbol{\alpha}^T \hat{\boldsymbol{\Sigma}}_w \boldsymbol{\alpha} = 1$.

We proceed with a proof by contradiction. Suppose that $\boldsymbol{\alpha}$ solves (3) and $\boldsymbol{\alpha}^T \hat{\boldsymbol{\Sigma}}_w \boldsymbol{\alpha} < 1$, $\boldsymbol{\alpha}^T \hat{\boldsymbol{\Sigma}}_b \boldsymbol{\alpha} > 0$. Let $c = 1/\sqrt{\boldsymbol{\alpha}^T \hat{\boldsymbol{\Sigma}}_w \boldsymbol{\alpha}}$. Since $c > 1$, it follows that $(c\boldsymbol{\alpha})^T \hat{\boldsymbol{\Sigma}}_b (c\boldsymbol{\alpha}) > \boldsymbol{\alpha}^T \hat{\boldsymbol{\Sigma}}_b \boldsymbol{\alpha}$. And $c\boldsymbol{\alpha}$ is in the feasible set for (3). This contradicts the assumption that $\boldsymbol{\alpha}$ solves (3). Hence, any solution to (3) that is not in the null space of $\hat{\boldsymbol{\Sigma}}_b$ also solves (38).

Note that we do not concern ourselves with solutions that are in the null space of $\hat{\boldsymbol{\Sigma}}_b$, as these are not useful for the purpose of discrimination and will arise only if too many discriminant vectors are used.

Proof of Proposition 1

Proof. Letting $\tilde{\boldsymbol{\Sigma}}_w^{\frac{1}{2}}$ denote the symmetric matrix square root of $\tilde{\boldsymbol{\Sigma}}_w$ and $\tilde{\boldsymbol{\beta}}_k = \tilde{\boldsymbol{\Sigma}}_w^{\frac{1}{2}} \boldsymbol{\beta}_k$, (6) becomes

$$\begin{aligned} &\text{maximize}_{\tilde{\boldsymbol{\beta}}_k} \{\tilde{\boldsymbol{\beta}}_k^T \tilde{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}} \mathbf{X}^T \mathbf{Y} (\mathbf{Y}^T \mathbf{Y})^{-\frac{1}{2}} \mathbf{P}_k^{\perp} (\mathbf{Y}^T \mathbf{Y})^{-\frac{1}{2}} \mathbf{Y}^T \mathbf{X} \tilde{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}} \tilde{\boldsymbol{\beta}}_k\} \\ &\text{subject to } ||\tilde{\boldsymbol{\beta}}_k||^2 \leq 1, \end{aligned} \tag{39}$$

which is equivalent to

$$\text{maximize}_{\tilde{\boldsymbol{\beta}}_k, \mathbf{u}_k} \{\tilde{\boldsymbol{\beta}}_k^T \mathbf{A} \mathbf{P}_k^{\perp} \mathbf{u}_k\} \text{ subject to } ||\tilde{\boldsymbol{\beta}}_k||^2 \leq 1, ||\mathbf{u}_k||^2 \leq 1, \tag{40}$$

where $\mathbf{A} = \tilde{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}} \mathbf{X}^T \mathbf{Y} (\mathbf{Y}^T \mathbf{Y})^{-\frac{1}{2}}$. Equivalence of (40) and (39) can be seen from partially optimizing (40) with respect to $\mathbf{u}_k$.

We claim that $\tilde{\boldsymbol{\beta}}_k$ and $\mathbf{u}_k$ that solve (40) are the kth left and right singular vectors of $\mathbf{A}$. By inspection, the claim holds when $k = 1$. Now, suppose that the claim holds for all $i < k$, where $k > 1$. Partially optimizing (40) with respect to $\tilde{\boldsymbol{\beta}}_k$ yields

$$\text{maximize}_{\mathbf{u}_k} \{\mathbf{u}_k^T \mathbf{P}_k^{\perp} \mathbf{A}^T \mathbf{A} \mathbf{P}_k^{\perp} \mathbf{u}_k\} \text{ subject to } ||\mathbf{u}_k||^2 \leq 1. \tag{41}$$

By definition, $\mathbf{P}_k^{\perp}$ is an orthogonal projection matrix into the space orthogonal to

$$(\mathbf{Y}^T \mathbf{Y})^{-\frac{1}{2}} \mathbf{Y}^T \mathbf{X} \hat{\boldsymbol{\beta}}_i = (\mathbf{Y}^T \mathbf{Y})^{-\frac{1}{2}} \mathbf{Y}^T \mathbf{X} \tilde{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}} \tilde{\boldsymbol{\beta}}_i = \mathbf{A}^T \tilde{\boldsymbol{\beta}}_i \propto \mathbf{u}_i \tag{42}$$

for all $i < k$, where proportionality follows from the fact that $\tilde{\boldsymbol{\beta}}_i$ and $\mathbf{u}_i$ are the ith singular vectors of $\mathbf{A}$ for all $i < k$. Hence, $\mathbf{P}_k^{\perp} = \mathbf{I} - \sum_{i=1}^{k-1} \mathbf{u}_i \mathbf{u}_i^T$. Therefore, by (41), $\mathbf{u}_k$ is the kth eigenvector of $\mathbf{A}^T\mathbf{A}$, or equivalently the kth right singular vector of $\mathbf{A}$. So by (40), $\tilde{\boldsymbol{\beta}}_k$ is the kth left singular vector of $\mathbf{A}$, or equivalently the kth eigenvector of $\mathbf{A}\mathbf{A}^T = n\tilde{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_b\tilde{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}}$. Therefore, the solution to (6) is the kth discriminant vector.

Proof of Proposition 2

For (18), the Karush-Kuhn-Tucker (KKT) conditions (Boyd and Vandenberghe, 2004) are given by

$$2\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta}^{(m-1)} - \lambda\boldsymbol{\Gamma} - 2\delta\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta} = 0, \;\; \delta \geq 0, \;\; \delta(\boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta} - 1) = 0, \;\; \boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta} \leq 1, \tag{43}$$

where we have dropped the "k" subscripts and superscripts for ease of notation, and where $\boldsymbol{\Gamma}$ is a p-vector of which the jth element is the subgradient of $\sum_{j=1}^{p} |\hat{\sigma}_j\beta_j|$ with respect to β_j; i.e. $\Gamma_j = \hat{\sigma}_j$ if $\beta_j > 0$, $\Gamma_j = -\hat{\sigma}_j$ if $\beta_j < 0$, and Γ_j is in between $\hat{\sigma}_j$ and $-\hat{\sigma}_j$ if $\beta_j = 0$.

First, suppose that for some j, $|(2\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta}^{(m-1)})_j| > \lambda\hat{\sigma}_j$. Then it must be the case that $2\delta\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta} \neq 0$. So $\delta > 0$ and $\boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta} = 1$. Then the KKT conditions simplify to

$$2\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta}^{(m-1)} - \lambda\boldsymbol{\Gamma} - 2\delta\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta} = 0, \;\; \boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta} = 1, \delta > 0. \tag{44}$$

Substituting $\mathbf{d} = \delta\boldsymbol{\beta}$, this is equivalent to solving (21) and then dividing the solution $\hat{\mathbf{d}}$ by $\sqrt{\hat{\mathbf{d}}^T\tilde{\boldsymbol{\Sigma}}_w\hat{\mathbf{d}}}$.

Now, suppose instead that $|(2\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta}^{(m-1)})_j| \leq \lambda\hat{\sigma}_j$ for all j. Then, by (43), it follows that $\hat{\boldsymbol{\beta}} = 0$ solves (18). By inspection of the subgradient equation for (21), we see that in this case $\hat{\mathbf{d}} = 0$ solves (21) as well. Therefore, the solution to (18) is as given in Proposition 2.

The same set of arguments applied to (20) lead to Proposition 2(b).

Proof of Proposition 3

Proof. Consider (17) with tuning parameter λ_1 and $k = 1$. Then by Theorem 6.1.1 of Clarke (1990), if there is a nonzero solution $\boldsymbol{\beta}^*$, then there exists $\mu \geq 0$ such that

$$0 \in 2\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta}^* - \lambda_1\boldsymbol{\Gamma}(\boldsymbol{\beta}^*) - 2\mu\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta}^*, \tag{45}$$

where $\boldsymbol{\Gamma}(\boldsymbol{\beta})$ is the subdifferential of $||\boldsymbol{\beta}||_1$. The subdifferential is the set of subgradients of $||\boldsymbol{\beta}||_1$; the jth element of a subgradient equals $\text{sign}(\beta_j)$ if $\beta_j \neq 0$ and is between -1 and 1 if $\beta_j = 0$. Left-multiplying (45) by $\boldsymbol{\beta}^*$ yields $0 = 2\boldsymbol{\beta}^{*T}\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta}^* - \lambda_1||\boldsymbol{\beta}^*||_1 - 2\mu\boldsymbol{\beta}^{*T}\tilde{\boldsymbol{\Sigma}}_w\boldsymbol{\beta}^*$. Since the sum of the first two terms is positive (since $\boldsymbol{\beta}^*$ is a nonzero solution), it follows that $\mu > 0$.

Now, define a new vector that is proportional to $\boldsymbol{\beta}^*$:

$$\hat{\boldsymbol{\beta}} = \frac{\mu}{(1+\mu)a}\boldsymbol{\beta}^* = c\boldsymbol{\beta}^* \tag{46}$$

where $a = \sqrt{n\boldsymbol{\beta}^{*T}\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta}^*}$. By inspection, $a \neq 0$, since otherwise $\boldsymbol{\beta}^*$ would not be a nonzero solution. Also, let $\lambda_2 = \lambda_1(\frac{1-ca}{a})$. Note that $1 - ca = \frac{1}{1+\mu} > 0$, so $\lambda_2 > 0$.

The generalized gradient of (36) with tuning parameter λ_2 evaluated at $\hat{\boldsymbol{\beta}}$ is proportional to

$$2\hat{\boldsymbol{\Sigma}}_b\hat{\boldsymbol{\beta}} - \lambda_2\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}})(\frac{\sqrt{n\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\Sigma}}_b\hat{\boldsymbol{\beta}}}}{1-\sqrt{n\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\Sigma}}_b\hat{\boldsymbol{\beta}}}}) - 2\tilde{\boldsymbol{\Sigma}}_w\hat{\boldsymbol{\beta}}(\frac{\sqrt{n\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\Sigma}}_b\hat{\boldsymbol{\beta}}}}{1-\sqrt{n\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\Sigma}}_b\hat{\boldsymbol{\beta}}}}), \tag{47}$$

or equivalently,

$$\begin{aligned}2c\hat{\mathbf{\Sigma}}_b\boldsymbol{\beta}^* - \lambda_2\mathbf{\Gamma}(\boldsymbol{\beta}^*)\frac{ac}{1-ac} - 2c\tilde{\mathbf{\Sigma}}_w\boldsymbol{\beta}^*\frac{ac}{1-ac} &= 2c\hat{\mathbf{\Sigma}}_b\boldsymbol{\beta}^* - \lambda_1 c\mathbf{\Gamma}(\boldsymbol{\beta}^*) - 2c\tilde{\mathbf{\Sigma}}_w\boldsymbol{\beta}^*\frac{ac}{1-ac}\\ &= 2c\hat{\mathbf{\Sigma}}_b\boldsymbol{\beta}^* - \lambda_1 c\mathbf{\Gamma}(\boldsymbol{\beta}^*) - 2c\mu\tilde{\mathbf{\Sigma}}_w\boldsymbol{\beta}^*\\ &= c(2\hat{\mathbf{\Sigma}}_b\boldsymbol{\beta}^* - \lambda_1\mathbf{\Gamma}(\boldsymbol{\beta}^*) - 2\mu\tilde{\mathbf{\Sigma}}_w\boldsymbol{\beta}^*). \qquad (48)\end{aligned}$$

Comparing (45) to (48), we see that 0 is contained in the generalized gradient of the SDA objective evaluated at $\hat{\boldsymbol{\beta}}$.

Proof of Proposition 4

PROOF. Since $n_1 = n_2$, NSC assigns an observation $\mathbf{x} \in \mathbb{R}^p$ to the class that maximizes

$$\sum_{j=1}^{p} \frac{x_j S(\overline{\mathbf{X}}_{kj}, \hat{\sigma}_j\lambda)}{\hat{\sigma}_j^2} \qquad (49)$$

where $\overline{\mathbf{X}}_{kj}$ is the mean of feature j in class k, and the soft-thresholding operator S is given by (24). On the other hand, the classification rule resulting from (37) assigns $\mathbf{x}$ to the class that minimizes

$$\left|\sum_{j=1}^{p} \frac{\overline{\mathbf{X}}_{kj} S(\overline{\mathbf{X}}_{1j}, \hat{\sigma}_j\lambda)}{\hat{\sigma}_j^2} - \sum_{j=1}^{p} \frac{x_j S(\overline{\mathbf{X}}_{1j}, \hat{\sigma}_j\lambda)}{\hat{\sigma}_j^2}\right|. \qquad (50)$$

This follows from the fact that (37) reduces to

$$\text{maximize}_{\boldsymbol{\beta}}\{\boldsymbol{\beta}^T\overline{\mathbf{X}}_1 - \lambda\sum_{j=1}^{p}|\beta_j\hat{\sigma}_j|\} \text{ subject to } \sum_{j=1}^{p}\beta_j^2\hat{\sigma}_j^2 \le 1, \qquad (51)$$

since $\frac{1}{\sqrt{n}}\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}} = \overline{\mathbf{X}}_1[\frac{1}{\sqrt{2}} \quad -\frac{1}{\sqrt{2}}]$ and $\hat{\mathbf{\Sigma}}_b = \frac{1}{n}\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}$.

Since the first term in (50) is positive if $k = 1$ and negative if $k = 2$, (37) classifies to class 1 if $\sum_{j=1}^{p}\frac{x_j S(\overline{\mathbf{X}}_{1j}, \hat{\sigma}_j\lambda)}{\hat{\sigma}_j^2} > 0$ and classifies to class 2 if $\sum_{j=1}^{p}\frac{x_j S(\overline{\mathbf{X}}_{1j}, \hat{\sigma}_j\lambda)}{\hat{\sigma}_j^2} < 0$. Because $\overline{\mathbf{X}}_{1j} = -\overline{\mathbf{X}}_{2j}$, by inspection of (49), the two methods result in the same classification rule.

Proof of equivalence of Fisher's LDA and optimal scoring

PROOF. Consider the following two problems:

$$\text{maximize}_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\boldsymbol{\beta}^T\hat{\mathbf{\Sigma}}_b\boldsymbol{\beta}\} \text{ subject to } \boldsymbol{\beta}^T(\hat{\mathbf{\Sigma}}_w + \mathbf{\Omega})\boldsymbol{\beta} = 1 \qquad (52)$$

and

$$\text{minimize}_{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\theta}\in\mathbb{R}^K}\{\frac{1}{n}||\mathbf{Y}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta}||^2 + \boldsymbol{\beta}^T\mathbf{\Omega}\boldsymbol{\beta}\} \text{ subject to } \boldsymbol{\theta}^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta} = 1. \qquad (53)$$

In Hastie et al. (1995), a somewhat challenging proof is given of the fact that the solutions $\hat{\boldsymbol{\beta}}$ to the two problems are proportional to each other. Here, we present a more direct argument. In (52) and (53), $\mathbf{\Omega}$ is a matrix such that $\hat{\mathbf{\Sigma}}_w + \mathbf{\Omega}$ is positive definite; if $\mathbf{\Omega} = 0$ then these two problems reduce to Fisher's LDA and optimal scoring. Optimizing (53) with respect to $\boldsymbol{\theta}$, we see that the $\boldsymbol{\beta}$ that solves (53) also solves

$$\text{minimize}_{\boldsymbol{\beta}}\{-\frac{2}{\sqrt{n}}\sqrt{\boldsymbol{\beta}^T\hat{\mathbf{\Sigma}}_b\boldsymbol{\beta}} + \boldsymbol{\beta}^T\hat{\mathbf{\Sigma}}_b\boldsymbol{\beta} + \boldsymbol{\beta}^T(\hat{\mathbf{\Sigma}}_w + \mathbf{\Omega})\boldsymbol{\beta}\}. \qquad (54)$$

For notational convenience, let $\tilde{\boldsymbol{\beta}} = (\hat{\boldsymbol{\Sigma}}_w + \boldsymbol{\Omega})^{\frac{1}{2}}\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\Sigma}}_b = (\hat{\boldsymbol{\Sigma}}_w + \boldsymbol{\Omega})^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_b(\hat{\boldsymbol{\Sigma}}_w + \boldsymbol{\Omega})^{-\frac{1}{2}}$. Then, the problems become

$$\text{maximize}_{\tilde{\boldsymbol{\beta}}}\{\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\Sigma}}_b\tilde{\boldsymbol{\beta}}\} \text{ subject to } \tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\beta}} = 1 \tag{55}$$

and

$$\text{minimize}_{\tilde{\boldsymbol{\beta}}}\{-\frac{2}{\sqrt{n}}\sqrt{\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\Sigma}}_b\tilde{\boldsymbol{\beta}}} + \tilde{\boldsymbol{\beta}}^T(\tilde{\boldsymbol{\Sigma}}_b + \mathbf{I})\tilde{\boldsymbol{\beta}}\}. \tag{56}$$

It is easy to see that the solution to (55) is the first eigenvector of $\tilde{\boldsymbol{\Sigma}}_b$. Let $\hat{\boldsymbol{\beta}}$ denote the solution to (56). Consequently, $\hat{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\Sigma}}_b\hat{\boldsymbol{\beta}} > 0$. So $\hat{\boldsymbol{\beta}}$ satisfies

$$\tilde{\boldsymbol{\Sigma}}_b\hat{\boldsymbol{\beta}}(1 - \frac{1}{\sqrt{n\hat{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\Sigma}}_b\hat{\boldsymbol{\beta}}}}) + \hat{\boldsymbol{\beta}} = 0, \tag{57}$$

and therefore $\sqrt{n\hat{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\Sigma}}_b\hat{\boldsymbol{\beta}}} < 1$. Now (57) indicates that $\hat{\boldsymbol{\beta}}$ is an eigenvector of $\tilde{\boldsymbol{\Sigma}}_b$ with eigenvalue $\lambda = \frac{\sqrt{n\hat{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\Sigma}}_b\hat{\boldsymbol{\beta}}}}{1-\sqrt{n\hat{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\Sigma}}_b\hat{\boldsymbol{\beta}}}}$; it remains to show that $\hat{\boldsymbol{\beta}}$ is in fact the first eigenvector. Notice that if we let $w = \hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\beta}}$ then $\lambda = \frac{\sqrt{n\lambda w}}{1-\sqrt{n\lambda w}}$, and so $w = \frac{\lambda}{n(1+\lambda)^2}$. Then the objective of (56) evaluated at $\hat{\boldsymbol{\beta}}$ equals

$$-\frac{2}{\sqrt{n}}\sqrt{\lambda w} + \lambda w + w = \frac{-2\lambda}{n(1+\lambda)} + \frac{\lambda}{n(1+\lambda)} = -\frac{\lambda}{n(1+\lambda)}. \tag{58}$$

The minimum occurs when λ is large. So the solution to (56) is the largest eigenvector of $\tilde{\boldsymbol{\Sigma}}_b$.

This argument can be extended to show that subsequent solutions to Fisher's discriminant problem and the optimal scoring problem are proportional to each other.

References

Barrett, T., Suzek, T., Troup, D., Wilhite, S., Ngau, W., Ledoux, P., Rudnev, D., Lash, A., Fujibuchi, W. and Edgar, R. (2005), 'NCBI GEO: mining millions of expression profiles–database and tools', *Nucleic Acids Research* **33**, D562–D566.

Bickel, P. and Levina, E. (2004), 'Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations', *Bernoulli* **10(6)**, 989–1010.

Boyd, S. and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge University Press.

Breiman, L. and Ihaka, R. (1984), Nonlinear discriminant analysis via scaling and ACE, Technical report, Univ. California, Berkeley.

Clarke, F. (1990), *Optimization and nonsmooth analysis*, SIAM, Troy, New York.

Clemmensen, L., Hastie, T., Witten, D. and Ersboll, B. (2011), 'Sparse discriminant analysis'.

Dudoit, S., Fridlyand, J. and Speed, T. (2001), 'Comparison of discrimination methods for the classification of tumors using gene expression data', *J. Amer. Statist. Assoc.* **96**, 1151–1160.

Fan, J. and Fan, Y. (2008), 'High-dimensional classification using features annealed independence rules', *Annals of Statistics* **36(6)**, 2605–2637.

Friedman, J. (1989), 'Regularized discriminant analysis', *Journal of the American Statistical Association* **84**, 165–175.

Friedman, J., Hastie, T., Hoefling, H. and Tibshirani, R. (2007), 'Pathwise coordinate optimization', *Annals of Applied Statistics* **1**, 302–332.

Gorski, J., Pfeuffer, F. and Klamroth, K. (2007), 'Biconvex sets and optimization with biconvex functions: a survey and extensions', *Mathematical Methods of Operations Rsearch* **66**, 373–407.

Grosenick, L., Greer, S. and Knutson, B. (2008), 'Interpretable classifiers for fMRI improve prediction of purchases', *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **16(6)**, 539–547.

Guo, Y., Hastie, T. and Tibshirani, R. (2007), 'Regularized linear discriminant analysis and its application in microarrays', *Biostatistics* **8**, 86–100.

Hastie, T., Buja, A. and Tibshirani, R. (1995), 'Penalized discriminant analysis', *Annals of Statistics* **23**, 73–102.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer Verlag, New York.

Hoefling, H. (2009), 'A path algorithm for the fused lasso signal approximator', *arXiv:0910.0526* .

Hunter, D. and Lange, K. (2004), 'A tutorial on MM algorithms', *The American Statistician* **58**, 30–37.

Johnson, N. (2010), 'A dynamic programming algorithm for the fused lasso and l0-segmentation'.

Jolliffe, I., Trendafilov, N. and Uddin, M. (2003), 'A modified principal component technique based on the lasso', *Journal of Computational and Graphical Statistics* **12**, 531–547.

Krzanowski, W., Jonathan, P., McCarthy, W. and Thomas, M. (1995), 'Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data', *Journal of the Royal Statistical Society, Series C* **44**, 101–115.

Lange, K. (2004), *Optimization*, Springer, New York.

Lange, K., Hunter, D. and Yang, I. (2000), 'Optimization transfer using surrogate objective functions', *Journal of Computational and Graphical Statistics* **9**, 1–20.

Leng, C. (2008), 'Sparse optimal scoring for multiclass cancer diagnosis and biomarker detection using microarray data', *Computational Biology and Chemistry* **32**, 417–425.

Mardia, K., Kent, J. and Bibby, J. (1979), *Multivariate Analysis*, Academic Press.

Nakayama, R., Nemoto, T., Takahashi, H., Ohta, T., Kawai, A., Yoshida, T., Toyama, Y., Ichikawa, H. and Hasegama, T. (2007), 'Gene expression analysis of soft tissue sarcomas: characterization and reclassification of malignant fibrous histiocytoma', *Modern Pathology* **20(7)**, 749–759.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E. and Golub, T. (2001), 'Multiclass cancer diagnosis using tumor gene expression signature', *PNAS* **98**, 15149–15154.

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011), 'Sparse linear discriminant analysis by thresholding for high dimensional data', *Annals of Statistics* .

Shen, H. and Huang, J. Z. (2008), 'Sparse principal component analysis via regularized low rank matrix approximation', *Journal of Multivariate Analysis* **101**, 1015–1034.

Sun, L., Hui, A., Su, Q., Vortmeyer, A., Kotliarov, Y., Pastorino, S., Passaniti, A., Menon, J., Walling, J., Bailey, R., Rosenblum, M., Mikkelsen, T. and Fine, H. (2006), 'Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain', *Cancer Cell* **9**, 287–300.

Tebbens, J. and Schlesinger, P. (2007), 'Improving implementation of linear discriminant analysis for the high dimension / small sample size problem', *Computational Statistics and Data Analysis* **52**, 423–437.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *J. Royal. Statist. Soc. B.* **58**, 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002), 'Diagnosis of multiple cancer types by shrunken centroids of gene expression', *Proc. Natl. Acad. Sci.* **99**, 6567–6572.

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.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005), 'Sparsity and smoothness via the fused lasso', *J. Royal. Statist. Soc. B.* **67**, 91–108.

Trendafilov, N. and Jolliffe, I. (2007), 'DALASS: Variable selection in discriminant analysis via the LASSO', *Computational Statistics and Data Analysis* **51**, 3718–3736.

Witten, D. and Tibshirani, R. (2009), 'Covariance-regularized regression and classification for high-dimensional problems', *J. Royal. Stat. Soc. B.* **71(3)**, 615–636, PMCID:PMC2806603.

Witten, D., Tibshirani, R. and Hastie, T. (2009), 'A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis', *Biostatistics* **10(3)**, 515–534.

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.

Zhu, J. and Hastie, T. (2004), 'Classification of gene microarrays by penalized logistic regression', *Biostatistics* **5**(2), 427–443.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *J. Royal. Stat. Soc. B.* **67**, 301–320.

Zou, H., Hastie, T. and Tibshirani, R. (2006), 'Sparse principal component analysis', *Journal of Computational and Graphical Statistics* **15**, 265–286.