Journal of the
Royal Statistical Society

SERIES B
Statistical
Methodology

J. R. Statist. Soc. B (2011)
73, *Part* 5, *pp.* 753–772

Penalized classification using Fisher's linear discriminant

Daniela M. Witten

University of Washington, Seattle, USA

and Robert Tibshirani

Stanford University, USA

[Received August 2010. Final revision March 2011]

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 that is obtained from LDA, since it involves all p features. We propose *penalized LDA*, which is 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 to optimize it efficiently 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 problems; Lasso; Linear discriminant analysis; Supervised learning

1. Introduction

In this paper, we consider the classification setting. The data consist of an $n \times p$ 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.
(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 shall refer to as the *normal model*, the *optimal scoring problem* and *Fisher's discriminant problem* (see for example

Address for correspondence: Daniela M. Witten, Department of Biostatistics, University of Washington, F-600 Health Sciences Building, Box 357232, Seattle, WA 98195-7232, USA.
E-mail: dwitten@u.washington.edu

1369–7412/11/73753

Mardia *et al.* (1979) and Hastie *et al.* (2009)). In recent years, several references 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 for example Tibshirani *et al.* (2002, 2003), Grosenick *et al.* (2008), Leng (2008) and 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 non-convex. We overcome this difficulty by using a minorization–maximization approach (see for example Lange *et al.* (2000), Hunter and Lange (2004) and 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 by 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) considered a closely related problem, but they proposed a specialized algorithm that can be applied only in the case of L_1-penalties on the discriminant vectors; moreover, they did not consider the high dimensional setting. In this paper, we take a more general approach that has several 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 for example Dudoit *et al.* (2001), Tibshirani *et al.* (2003) and Bickel and Levina (2004)).
(d) It yields interpretable discriminant vectors, where the concept of interpretability can be chosen on the basis of 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 by 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 an $n \times p$ matrix with observations on the rows and features on the columns. We assume that the features are centred to have mean 0, and we let $\mathbf{X}_j$ denote feature or column j and $\mathbf{x}_i$ denote observation or row i. $C_k \subset \{1, \ldots, n\}$ contains the indices of the observations in class k, and $n_k = |C_k|, \Sigma_{k=1}^K n_k = n$. The standard estimate for the *within-class covariance matrix* $\boldsymbol{\Sigma}_{\mathrm{w}}$ is given by

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

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

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

In later sections, we shall make use of the fact that

$$\hat{\boldsymbol{\Sigma}}_\mathrm{b} = \frac{1}{n}\mathbf{X}^\mathrm{T}\mathbf{Y}(\mathbf{Y}^\mathrm{T}\mathbf{Y})^{-1}\mathbf{Y}^\mathrm{T}\mathbf{X},$$

where $\mathbf{Y}$ is an $n \times K$ 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, i.e. we sequentially solve

$$\text{maximize}_{\boldsymbol{\beta}_k \in \mathbb{R}^p}(\boldsymbol{\beta}_k^\mathrm{T}\hat{\boldsymbol{\Sigma}}_\mathrm{b}\boldsymbol{\beta}_k) \text{ subject to } \boldsymbol{\beta}_k^\mathrm{T}\hat{\boldsymbol{\Sigma}}_\mathrm{w}\boldsymbol{\beta}_k \leqslant 1, \boldsymbol{\beta}_k^\mathrm{T}\hat{\boldsymbol{\Sigma}}_\mathrm{w}\boldsymbol{\beta}_i = 0 \qquad \forall i < k. \tag{3}$$

Problem (3) is generally written with the inequality constraint replaced with an equality constraint, but the two are equivalent if $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$ has full rank, as is shown in Appendix A. We shall refer to the solution $\hat{\boldsymbol{\beta}}_k$ to problem (3) as the kth *discriminant vector*. In general, there are $K-1$ non-trivial 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, we can transform the observations by using only the first $k < K-1$ discriminant vectors 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 problem (3) by substituting $\tilde{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\Sigma}}_\mathrm{w}^{1/2}\boldsymbol{\beta}_k$, where $\hat{\boldsymbol{\Sigma}}_\mathrm{w}^{1/2}$ is the symmetric matrix square root of $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$. Then, Fisher's discriminant problem is reduced to a standard eigenproblem. In fact, from equation (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 why problem (3) does not lead to a suitable classifier.

(a) $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$ is singular. Any discriminant vector that is in the null space of $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$ but not in the null space of $\hat{\boldsymbol{\Sigma}}_\mathrm{b}$ 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.

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

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

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

$$\text{maximize}_{\boldsymbol{\beta}_k \in \mathbb{R}^p}(\boldsymbol{\beta}_k^\mathrm{T}\hat{\boldsymbol{\Sigma}}_\mathrm{b}\boldsymbol{\beta}_k) \text{ subject to } \boldsymbol{\beta}_k^\mathrm{T}\tilde{\boldsymbol{\Sigma}}_\mathrm{w}\boldsymbol{\beta}_k \leqslant 1, \boldsymbol{\beta}_k^\mathrm{T}\tilde{\boldsymbol{\Sigma}}_\mathrm{w}\boldsymbol{\beta}_i = 0 \qquad \forall i < k, \tag{5}$$

where $\tilde{\boldsymbol{\Sigma}}_{\text{w}}$ is a positive definite estimate for $\boldsymbol{\Sigma}_{\text{w}}$. Criterion (5) addresses the singularity issue, but not the interpretability issue.

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

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

$$\text{maximize}_{\boldsymbol{\beta}_k}(\boldsymbol{\beta}_k^{\text{T}}\hat{\boldsymbol{\Sigma}}_{\text{b}}^k\boldsymbol{\beta}_k)\text{subject to } \boldsymbol{\beta}_k^{\text{T}}\tilde{\boldsymbol{\Sigma}}_{\text{w}}\boldsymbol{\beta}_k \leqslant 1 \tag{6}$$

where

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

$\mathbf{P}_k^{\perp}$ 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 $(\mathbf{Y}^{\text{T}}\mathbf{Y})^{-1/2}\mathbf{Y}^{\text{T}}\mathbf{X}\hat{\boldsymbol{\beta}}_i$ for all $i < k$.

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

3. Brief review of minorization algorithms

In this paper, we shall 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 for example Boyd and Vandenberghe (2004)) can be used to solve problem (8). If not, solving problem (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, problem (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—problem (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

$$\begin{aligned} f(\boldsymbol{\beta}^{(m)}) &= g(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)}),\\ f(\boldsymbol{\beta}) &\geqslant g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})\ \forall\boldsymbol{\beta}. \end{aligned} \tag{9}$$

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

$$\boldsymbol{\beta}^{(m+1)} = \arg\max_{\boldsymbol{\beta}}\{g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})\}. \tag{10}$$

Then, by expression (9),

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

This means that in each iteration the objective is non-decreasing. However, in general we do not expect to arrive at the global optimum of problem (8) by using a minorization approach: global optima for non-convex problems are very difficult to obtain, and a local optimum is the best that 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 equation (10) is easily solved. For instance, if $g(\beta|\beta^{(m)})$ is concave in β 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 linear discriminant analysis proposal

4.1. General form of penalized linear discriminant analysis

We would like to modify 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

$$\text{maximize}_{\beta_1}\{\beta_1^{\mathrm{T}}\hat{\Sigma}_{\mathrm{b}}\beta_1 - P_1(\beta_1)\} \text{ subject to } \beta_1^{\mathrm{T}}\tilde{\Sigma}_{\mathrm{w}}\beta_1 \leqslant 1, \tag{12}$$

where $\tilde{\Sigma}_{\mathrm{w}}$ is a positive definite estimate for Σ_{w} and where P_1 is a convex penalty function. In this paper, we shall be most interested in the case where $\tilde{\Sigma}_{\mathrm{w}}$ is the diagonal estimate (4), since it has been shown that using a diagonal estimate for Σ_{w} can lead to good classification results when $p \gg n$ (see for example Tibshirani *et al.* (2002) and Bickel and Levina (2004)). Note that problem (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 $\tilde{\Sigma}_{\mathrm{w}}$ were the identity.

To obtain multiple discriminant vectors, rather than requiring that subsequent discriminant vectors be orthogonal with respect to $\tilde{\Sigma}_{\mathrm{w}}$—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

$$\text{maximize}_{\beta_k}\{\beta_k^{\mathrm{T}}\hat{\Sigma}_{\mathrm{b}}^k\beta_k - P_k(\beta_k)\} \text{ subject to } \beta_k^{\mathrm{T}}\tilde{\Sigma}_{\mathrm{w}}\beta_k \leqslant 1, \tag{13}$$

where $\hat{\Sigma}_b^k$ is given by equation (7), with $\mathbf{P}_k^{\perp}$ an orthogonal projection matrix into the space that is orthogonal to $(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1/2}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\hat{\beta}_i$ 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 problem (12) follows from problem (13) with $k=1$.

In general, problem (13) cannot be solved by 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 $\mathbf{A}$, $f(\beta)=\beta^{\mathrm{T}}\mathbf{A}\beta$ is convex in β. Thus, for a fixed value of $\beta^{(m)}$,

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

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

$$g(\beta_k|\beta^{(m)}) = 2\beta_k^{\mathrm{T}}\hat{\Sigma}_{\mathrm{b}}^k\beta^{(m)} - \beta^{(m)\mathrm{T}}\hat{\Sigma}_{\mathrm{b}}^k\beta^{(m)} - P_k(\beta_k) \tag{15}$$

minorizes the objective of problem (13) at $\beta^{(m)}$. Moreover, since P_k is a convex function, $g(\beta_k|\beta^{(m)})$ is concave in β_k and hence can be maximized by using convex optimization tools. We can use equation (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.

4.1.1. *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 $(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1/2}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\hat{\beta}_i$ for all $i<k$. Let $\mathbf{P}_1^{\perp}=\mathbf{I}$.

(b) Let $\hat{\boldsymbol{\Sigma}}_b^k=(1/n)\mathbf{X}^{\mathrm{T}}\mathbf{Y}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1/2}\mathbf{P}_k^{\perp}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1/2}\mathbf{Y}^{\mathrm{T}}\mathbf{X}$. Note that $\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}^1=\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}$.

(c) Let $\boldsymbol{\beta}_k^{(0)}$ be the first eigenvector of $\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}^k$.

(d) For $m=1,2,\ldots$ until convergence: let $\boldsymbol{\beta}_k^{(m)}$ be the solution to

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

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

Of course, the solution to problem (16) will depend on the form of the convex function P_k. In the next section, we shall 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 by using only the first $k<K-1$ penalized discriminant vectors.

4.2. Penalized LDA-L_1 and penalized LDA-FL methods

4.2.1. Penalized LDA-L_1 method

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

$$\text{maximize}_{\boldsymbol{\beta}_k}\left(\boldsymbol{\beta}_k^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}^k\boldsymbol{\beta}_k-\lambda_k\sum_{j=1}^{p}|\hat{\sigma}_j\beta_{kj}|\right)\text{ subject to }\boldsymbol{\beta}_k^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}\boldsymbol{\beta}_k\leqslant 1. \tag{17}$$

When the tuning parameter λ_k is large, some elements of the solution $\hat{\boldsymbol{\beta}}_k$ will be exactly equal to 0. In problem (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. The penalized LDA-L_1 method is appropriate if we want to obtain a sparse classifier—i.e. a classifier for which the decision rule involves only a subset of the features. In particular, the resulting discriminant vectors are sparse, so the penalized LDA-L_1 method amounts to projecting the data onto a low dimensional subspace that involves only a subset of the features.

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

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

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

4.2.2. Penalized LDA-FL method

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

$$\text{maximize}_{\boldsymbol{\beta}_k}\left(\boldsymbol{\beta}_k^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}^k\boldsymbol{\beta}_k-\lambda_k\sum_{j=1}^{p}|\hat{\sigma}_j\beta_{kj}|-\gamma_k\sum_{j=2}^{p}|\hat{\sigma}_j\beta_{kj}-\hat{\sigma}_{j-1}\beta_{k,j-1}|\right)\text{subject to }\boldsymbol{\beta}_k^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}\boldsymbol{\beta}_k\leqslant 1. \tag{19}$$

When the non-negative tuning parameter λ_k is large then the resulting discriminant vector will be sparse in the features, and when the non-negative 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 problem (13), we again apply algorithm 1. Step (d) can be written as

$$\text{maximize}_{\beta_k}\left(2\boldsymbol{\beta}_k^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}^k\boldsymbol{\beta}_k^{(m-1)}-\lambda_k\sum_{j=1}^{p}|\hat{\sigma}_j\beta_{kj}|-\gamma_k\sum_{j=2}^{p}|\hat{\sigma}_j\beta_{kj}-\hat{\sigma}_{j-1}\beta_{k,j-1}|\right)\text{subject to } \boldsymbol{\beta}_k^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}\boldsymbol{\beta}_k\leqslant 1. \tag{20}$$

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

4.2.3. *Minorization step for penalized LDA-L_1 and penalized LDA-FL methods*

Now we present proposition 2, which provides a solution to problems (18) and (20). In other words, proposition 2 provides details for performing step (d) in algorithm 1 for the penalized LDA-L_1 and penalized LDA-FL methods.

Proposition 2.

(a) To solve problem (18), we first solve the problem

$$\text{minimize}_{\mathbf{d}\in\mathbb{R}^p}\left(\mathbf{d}^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}\mathbf{d}-2\mathbf{d}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}^k\boldsymbol{\beta}_k^{(m-1)}+\lambda_k\sum_j|\hat{\sigma}_jd_j|\right). \tag{21}$$

If $\hat{\mathbf{d}}=0$ then $\hat{\boldsymbol{\beta}}_k=0$. Otherwise, $\hat{\boldsymbol{\beta}}_k=\hat{\mathbf{d}}/\sqrt{}(\hat{\mathbf{d}}^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}\hat{\mathbf{d}})$.

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

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

If $\hat{\mathbf{d}}=0$ then $\hat{\boldsymbol{\beta}}_k=0$. Otherwise, $\hat{\boldsymbol{\beta}}_k=\hat{\mathbf{d}}/\sqrt{}(\hat{\mathbf{d}}^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}\hat{\mathbf{d}})$.

The proof is given in Appendix A. Some comments on proposition 2 are as follows.

(a) If $\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}$ is the diagonal estimate (4), then the solution to problem (21) is

$$\hat{d}_j=\frac{1}{\hat{\sigma}_j^2}S\left\{(\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}^k\boldsymbol{\beta}_k^{(m-1)})_j,\frac{\lambda_k\hat{\sigma}_j}{2}\right\} \tag{23}$$

where S is the soft thresholding operator, which is defined as

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

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

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

where Γ_j is the subgradient of $|d_j|$, which is 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 equation (23) follows from equation (25).

(b) In contrast, if $\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}$ is a non-diagonal positive definite estimate of $\boldsymbol{\Sigma}_{\mathrm{w}}$, then we can solve problem (21) by co-ordinate descent (see for example Friedman *et al.* (2007)). Problem (21) is in that case closely related to the lasso but may involve more demanding computations. This is because 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 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. But if $\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}$ is a $p\times p$ matrix without special structure then we must store it in full to solve problem (21).

(c) If $\tilde{\Sigma}_{\mathrm{w}}$ is a diagonal estimate for Σ_{w} then problem (22) is a diagonal fused lasso problem, for which fast algorithms have been proposed (see for example Hoefling (2010) and Johnson (2010)).

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 problem (17), which is equal to the largest eigenvalue of $\tilde{\Sigma}_{\mathrm{w}}^{-1/2}\hat{\Sigma}_{\mathrm{b}}^{k}\tilde{\Sigma}_{\mathrm{w}}^{-1/2}$, is non-increasing in k. So, instead, we take the following approach. We first fix a non-negative constant λ, and then we take

$$\lambda_k = \lambda \| \tilde{\Sigma}_{\mathrm{w}}^{-1/2}\hat{\Sigma}_{\mathrm{b}}^{k}\tilde{\Sigma}_{\mathrm{w}}^{-1/2} \|$$

where $\|\cdot\|$ indicates the largest eigenvalue. When $p \gg n$, this largest eigenvalue can be quickly computed by using the fact that $\hat{\Sigma}_{\mathrm{b}}^{k}$ 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 non-negative constants λ and γ. Then, we take $\lambda_k = \lambda \| \tilde{\Sigma}_{\mathrm{w}}^{-1/2}\hat{\Sigma}_{\mathrm{b}}^{k}\tilde{\Sigma}_{\mathrm{w}}^{-1/2} \|$ and $\gamma_k = \gamma \| \tilde{\Sigma}_{\mathrm{w}}^{-1/2}\hat{\Sigma}_{\mathrm{b}}^{k}\tilde{\Sigma}_{\mathrm{w}}^{-1/2} \|$. λ and γ can be chosen by cross-validation.

4.2.5. *Timing results for penalized linear discriminant analysis*

We now comment on the computations that are involved in the algorithms that were 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 $\tilde{\Sigma}_{\mathrm{w}}$. The R library `penalizedLDA` (Witten, 2011) was used. Timing depends critically on the convergence criterion that is 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}, i.e. $|r_i - r_{i+1}|/r_{i+1} < 10^{-6}$ 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.

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 four-class data†

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

†Mean (and standard error) of running time, over 25 repetitions. The diagonal estimate (4) was used for $\tilde{\Sigma}_{\mathrm{w}}$.

4.3. Recasting penalized linear discriminant analysis as a biconvex problem

Rather than using a minorization approach to solve the non-convex problem (12), we could instead recast it as a biconvex problem. Consider the problem

$$\text{maximize}_{\beta,\mathbf{u}}\left\{\frac{2}{\sqrt{n}}\beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1/2}\mathbf{u}-P(\beta)-\mathbf{u}^{\mathrm{T}}\mathbf{u}\right\}\text{ subject to }\beta^{\mathrm{T}}\tilde{\Sigma}_{\mathrm{w}}\beta\leqslant 1. \tag{27}$$

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

4.3.1. Algorithm 2: a biconvex formulation for penalized linear discriminant analysis

(a) Let $\beta^{(0)}$ be the first eigenvector of $\tilde{\Sigma}_{\mathrm{w}}^{-1}\hat{\Sigma}_{\mathrm{b}}$.

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

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

$$\text{maximize}_{\mathbf{u}}\left\{\frac{2}{\sqrt{n}}\beta^{(m-1)^{\mathrm{T}}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1/2}\mathbf{u}-\mathbf{u}^{\mathrm{T}}\mathbf{u}\right\}; \tag{28}$$

(ii) let $\beta^{(m)}$ solve

$$\text{maximize}_{\beta}\left\{\frac{2}{\sqrt{n}}\beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1/2}\mathbf{u}^{(m)}-P(\beta)\right\}\text{ subject to }\beta^{\mathrm{T}}\tilde{\Sigma}_{\mathrm{w}}\beta\leqslant 1. \tag{29}$$

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

$$\text{maximize}_{\beta}\{2\beta^{\mathrm{T}}\hat{\Sigma}_{\mathrm{b}}\beta^{(m-1)}-P(\beta)\}\text{ subject to }\beta^{\mathrm{T}}\tilde{\Sigma}_{\mathrm{w}}\beta\leqslant 1. \tag{30}$$

Comparing problem (30) with problem (16), we see that the biconvex formulation (27) results in the same update step as the minorization approach that was 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 $\tilde{\Sigma}_{\mathrm{w}}=\mathbf{I}$ and a bound form is used for the penalty $P(\beta)$. Since $\mathbf{X}^{\mathrm{T}}\mathbf{Y}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1/2}$ 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 by using the diagonal estimate (4) for $\tilde{\Sigma}_{\mathrm{w}}$, as implemented in the R package `penalizedLDA`. The nearest shrunken centroids (Tibshirani *et al.*, 2002, 2003) method NSC 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 Σ_{w} and imposing L_1-penalties on the class mean vectors under the normal model, and RDA combines a ridge-type penalty in estimating Σ_{w} with soft thresholding of $\tilde{\Sigma}_{\mathrm{w}}^{-1}\hat{\mu}_k$. These methods are discussed further in Section 6.

The tuning parameters for each of the methods considered were as follows. For the penalized LDA-L_1 method, λ described in Section 4.2.4 was a tuning parameter. For the penalized LDA-FL method, we treated $\lambda=\gamma$ (see Section 4.2.4) as a single tuning parameter 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. The NSC method 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 that is used to regularize the estimate of Σ_w.

5.2. Simulation study

We compare penalized LDA with the NSC and RDA methods 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.

(a) *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 $\mu_{1j} = 0.7$ if $1 \leqslant j \leqslant 25$, $\mu_{2j} = 0.7$ if $26 \leqslant j \leqslant 50$, $\mu_{3j} = 0.7$ if $51 \leqslant j \leqslant 75$, and $\mu_{4j} = 0.7$ if $75 \leqslant j \leqslant 100$ and $\mu_{kj} = 0$ otherwise.

(b) *Simulation 2: mean shift with dependent features*—there are two classes. For $i \in C_1$, $\mathbf{x}_i \sim N(0, \boldsymbol{\Sigma})$ and, for $i \in C_2$, $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\mu_j = 0.6$ if $j \leqslant 200$ and $\mu_j = 0$ otherwise. The covariance structure is block diagonal, with five 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.

(c) *Simulation 3: one-dimensional mean shift with independent features*—there are four classes, and the features are independent. For $i \in C_k$, $X_{ij} \sim N\{(k-1)/3, 1\}$ if $j \leqslant 100$, and $X_{ij} \sim N(0, 1)$ otherwise. Note that a one-dimensional projection of the data fully captures the class structure.

(d) *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 $1 \leqslant j \leqslant 25$ and $\mu_{1j} = 0$ otherwise, $\mu_{2j} \sim N(0, 0.3^2)$ if $26 \leqslant j \leqslant 50$

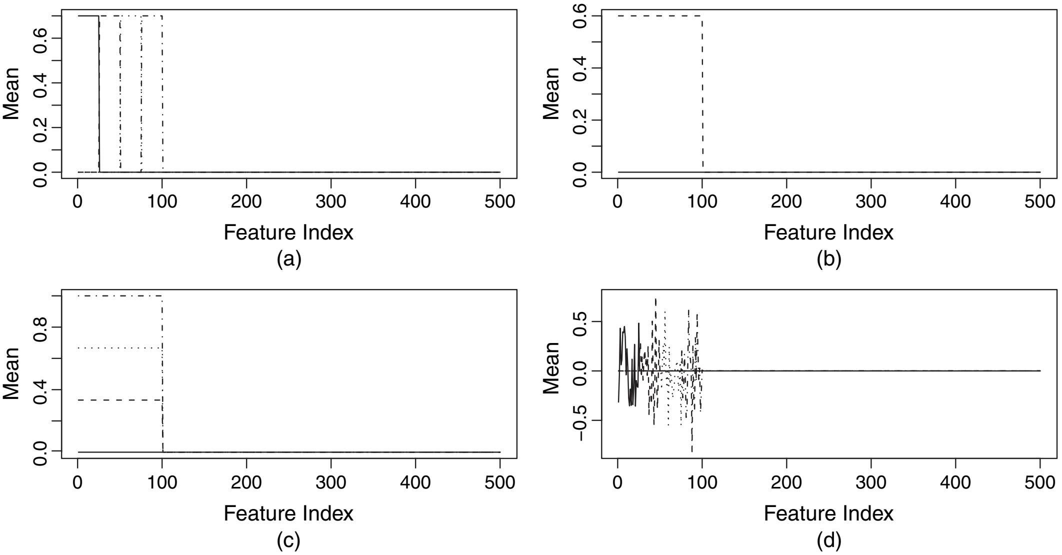

Fig. 1. Class mean vectors for each simulation (———, class 1; – – –, class 2; · · · · · · ·, class 3; · – · – ·, class 4): (a) simulation 1; (b) simulation 2; (c) simulation 3; (d) simulation 4

and $\mu_{2j} = 0$ otherwise, $\mu_{3j} \sim N(0, 0.3^2)$ if $51 \leqslant j \leqslant 75$ and $\mu_{3j} = 0$ otherwise, and $\mu_{4j} \sim N(0, 0.3^2)$ if $75 \leqslant j \leqslant 100$ and $\mu_{4j} = 0$ otherwise.

Fig. 1 displays the class mean vectors for each simulation.

For each method, models were fitted 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 Fig. 1).

Test set errors and the numbers of non-zero features that were used are reported in Table 2. For penalized LDA, the numbers of discriminant vectors that were used are also reported. The penalized LDA-FL method 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 course, in real data applications, the penalized LDA-FL method 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.

Table 2. Simulation results†

Simulation		*Results for the following methods:*			
		Penalized LDA-L_1	*Penalized LDA-FL*	*NSC*	*RDA*
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)	—	—
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)	—	—
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)	—	—
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)	—	—	—

†Mean (and standard errors), computed over 25 repetitions, of test set errors, number of non-zero features and number of discriminant vectors used.

Table 3. Results obtained on gene expression data over 10 training–test set splits†

Data set		*Results for the following methods:*		
		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)

†The quantities reported are the mean (and standard deviation) of test set errors and non-zero coefficients.

5.3. Application to gene expression data

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

(a) *Ramaswamy data*—a data set consisting of 16063 gene expression measurements and 198 samples belonging to 14 distinct cancer subtypes (Ramaswamy *et al.*, 2001) (the data set has been studied in several references (see for example Zhu and Hastie (2004), Guo *et al.* (2007) and Witten and Tibshirani (2009)) and is available from `http://www-stat.stanford.edu/~hastie/glmnet/glmnetData/`);

(b) *Nakayama data*—a data set consisting of 105 samples from 10 types of soft tissue tumours, each with 22283 gene expression measurements (Nakayama *et al.*, 2007) (we limited the

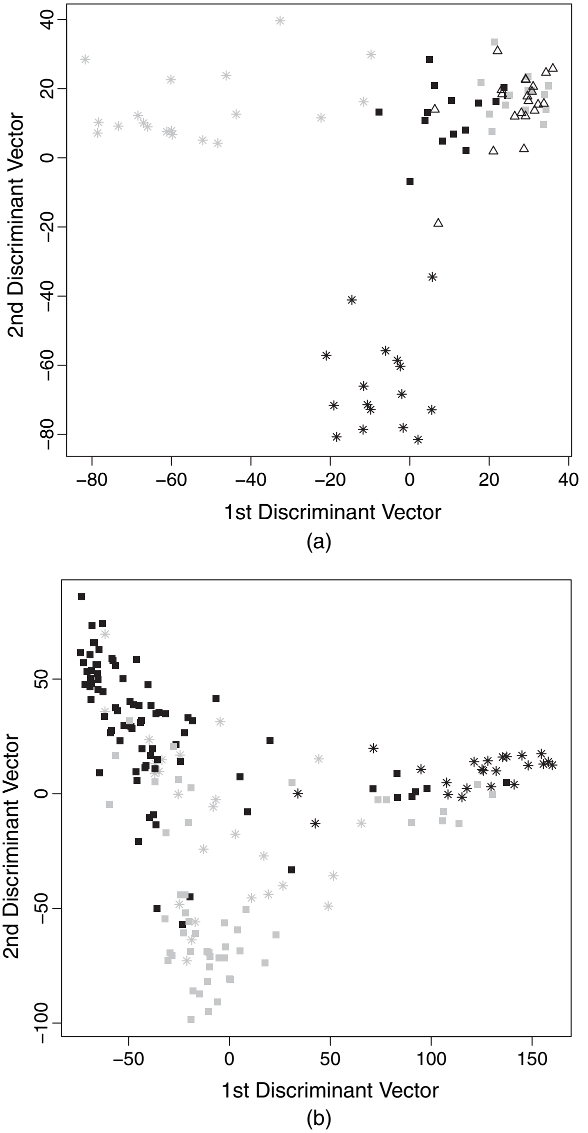

Fig. 2. For (a) the Nakayama and (b) the Sun data, the samples were projected onto the first two penalized discriminant vectors: the samples in each class are shown by using a distinct symbol

analysis to five tumour 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 from the *Gene Expression Omnibus* (Barrett *et al.*, 2005) with accession number GDS2736);

(c) *Sun data*—a data set consisting of 180 samples and 54613 expression measurements (Sun *et al.*, 2006). The samples fall into four classes: one non-tumour class and three types of glioma; the data are available from the *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 10 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 that were used by each classifier across each training–test set split. Indeed, this instability in the set of genes selected probably 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.

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

6. 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.

6.1. Normal model

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

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

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

$$\delta_k(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{\mathrm{w}}^{-1}\hat{\boldsymbol{\mu}}_k - \tfrac{1}{2}\hat{\boldsymbol{\mu}}_k^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{\mathrm{w}}^{-1}\hat{\boldsymbol{\mu}}_k \tag{32}$$

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

6.2. Optimal scoring problem

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

$$\mathrm{minimize}_{\beta_k\in\mathbb{R}^p,\boldsymbol{\theta}_k\in\mathbb{R}^K}\left(\frac{1}{n}\|\mathbf{Y}\boldsymbol{\theta}_k-\mathbf{X}\beta_k\|^2\right) \text{ subject to } \boldsymbol{\theta}_k^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\boldsymbol{\theta}_k=1, \boldsymbol{\theta}_k^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\boldsymbol{\theta}_i=0 \qquad \forall i<k \tag{33}$$

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 problem (33) is proportional to the solution to problem (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 Appendix A.

6.3. Linear discriminant analysis in high dimensions

In recent years, many researchers have proposed extensions of LDA to the high dimensional setting 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 to penalize 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 *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 $\bar{\mu}_{kj} = S\{\hat{\mu}_{kj}, \lambda\hat{\sigma}_j\surd(1/n_k + 1/n)\}$, 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 RDA proposal* (Guo *et al.*, 2007) arises instead from applying the normal model approach with covariance matrix $\tilde{\boldsymbol{\Sigma}}_{\text{w}} = \hat{\boldsymbol{\Sigma}}_{\text{w}} + \rho\mathbf{I}$ and performing soft thresholding to obtain a classifier that is sparse in the features.

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

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

where λ is a non-negative 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). 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.

Our penalized LDA proposal is instead a direct extension of Fisher's discriminant problem (3). Trendafilov and Jolliffe (2007) considered a problem that was very similar to penalized LDA-L_1. But they discussed 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 shall explore how our penalized LDA-L_1 proposal relates to the NSC and SDA methods.

Table 4. Advantages and disadvantages of using the normal model, optimal scoring and Fisher's discriminant problem as the basis for penalized LDA with an L_1-penalty

Method	*Advantages*	*Disadvantages*	*Reference*
Normal model	Sparse class means if diagonal estimate of Σ_w used; computations are fast	Does not give sparse discriminant vectors; no reduced rank classification	Tibshirani *et al.* (2002)

Optimal scoring	Sparse discriminant vectors	Difficult to enforce diagonal estimate for Σ_w, which is useful if $p>n$; computations can be slow	Grosenick *et al.* (2008);
			Leng (2008);
			Clemmensen *et al.* (2011)

Fisher's discriminant problem	Sparse discriminant vectors; simple to enforce diagonal estimate of Σ_w; computations are fast using diagonal estimate of Σ_w.	Computations can be slow when p is large, unless diagonal estimate of Σ_w is used	This work

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 β_1 and θ_1 for convenience. Partially optimizing criterion (35) with respect to θ reveals that, for any β for which $\mathbf{Y}^T\mathbf{X}\beta\neq 0$, the optimal θ equals

$$\frac{(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}\beta}{\surd\{\beta^T\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}\beta\}}.$$

So criterion (35) can be rewritten as

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

Assume that each feature has been standardized to have within-class standard deviation equal to 1. Take $\tilde{\Sigma}_w=\hat{\Sigma}_w+\Omega$, where Ω is chosen so that $\tilde{\Sigma}_w$ 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 β^* to problem (17) the objective is positive. Then, there is a positive tuning parameter λ_2 and a positive scalar c such that $c\beta^*$ corresponds to a zero of the generalized gradient of the SDA objective (36).

A proof is given in Appendix A. Note that the assumption that the objective is positive at the solution β^* is not very taxing—it simply means that β^* results in a higher value of the objective than does a vector of 0s. Proposition 3 states that, if the same positive definite estimate for Σ_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, the penalized LDA-L_1 method has a clear relationship with Fisher's discriminant problem. Moreover, unlike SDA, it provides a natural way to enforce a diagonal estimate of Σ_w.

7.2. Connection with nearest shrunken centroids

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

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

$$\text{maximize}_{\boldsymbol{\beta}}\left\{\sqrt{(\boldsymbol{\beta}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{\mathrm{b}}\boldsymbol{\beta})}-\lambda\sum_{j=1}^{p}|\beta_j\hat{\sigma}_j|\right\} \text{ subject to } \boldsymbol{\beta}^{\mathrm{T}}\tilde{\mathbf{\Sigma}}_{\mathrm{w}}\boldsymbol{\beta}\leqslant 1 \qquad (37)$$

where $\tilde{\mathbf{\Sigma}}_{\mathrm{w}}$ is the diagonal estimate (4). Consider the classification rule that is 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 problem (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, the penalized LDA-L_1 method with a diagonal estimate of $\mathbf{\Sigma}_{\mathrm{w}}$ and the NSC method are closely connected when $K=2$. This connection does not hold for larger values of K, since the NSC method penalizes the elements of the $p \times K$ class centroid matrix, whereas the penalized LDA-L_1 method penalizes the eigenvectors of this matrix. A proof of proposition 4 is given in Appendix A.

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 on the basis of the problem at hand and can result in an interpretable classifier. A potentially useful but unexplored area of application for our proposal is functional magnetic resonance imaging 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. When P_k is an L_1-penalty, problem (12) is closely related to the 'SCoTLASS' proposal for sparse principal components analysis (Jolliffe *et al.*, 2003). Criterion (12) and algorithm 1 for optimizing it are closely related to the penalized principal components algorithms that have been considered by various researchers (see for example Zou *et al.* (2006), Shen and Huang (2008) and 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 are available on the Comprehensive R Archive Network: `http://cran.r-project.org/`.

Acknowledgements

We thank two 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 A

A.1. Equivalence between problem (3) and standard formulation for linear discriminant analysis

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

$$\text{maximize}_{\boldsymbol{\beta}_k \in \mathbb{R}^p} (\boldsymbol{\beta}_k^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\mathrm{b}} \boldsymbol{\beta}_k) \text{ subject to } \boldsymbol{\beta}_k^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\mathrm{w}} \boldsymbol{\beta}_k = 1, \boldsymbol{\beta}_k^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\mathrm{w}} \boldsymbol{\beta}_i = 0 \qquad \forall i < k. \tag{38}$$

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

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

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

A.2. Proof of proposition 1

Letting $\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}^{1/2}$ denote the symmetric matrix square root of $\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}$ and $\tilde{\boldsymbol{\beta}}_k = \tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}^{1/2} \boldsymbol{\beta}_k$, problem (6) becomes

$$\text{maximize}_{\tilde{\boldsymbol{\beta}}_k} \{ \tilde{\boldsymbol{\beta}}_k^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}^{-1/2} \mathbf{X}^{\mathrm{T}} \mathbf{Y} (\mathbf{Y}^{\mathrm{T}} \mathbf{Y})^{-1/2} \mathbf{P}_k^{\perp} (\mathbf{Y}^{\mathrm{T}} \mathbf{Y})^{-1/2} \mathbf{Y}^{\mathrm{T}} \mathbf{X} \tilde{\Sigma}_{\mathrm{w}}^{-1/2} \tilde{\boldsymbol{\beta}}_k \} \text{ subject to } \|\tilde{\boldsymbol{\beta}}_k\|^2 \leqslant 1, \tag{39}$$

which is equivalent to

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

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

We claim that $\tilde{\boldsymbol{\beta}}_k$ and $\mathbf{u}_k$ that solve problem (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 problem (40) with respect to $\tilde{\boldsymbol{\beta}}_k$ yields

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

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

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

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, $\mathbf{P}_k^{\perp} = \mathbf{I} - \Sigma_{i=1}^{k-1} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}$. Therefore, by problem (41), $\mathbf{u}_k$ is the kth eigenvector of $\mathbf{A}^{\mathrm{T}} \mathbf{A}$, or equivalently the kth right singular vector of $\mathbf{A}$. So, by problem (40), $\tilde{\boldsymbol{\beta}}_k$ is the kth left singular vector of $\mathbf{A}$, or equivalently the kth eigenvector of

$$\mathbf{A} \mathbf{A}^{\mathrm{T}} = n \tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}^{-1/2} \hat{\boldsymbol{\Sigma}}_{\mathrm{b}} \tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}^{-1/2}.$$

Therefore, the solution to problem (6) is the kth discriminant vector.

A.3. Proof of proposition 2

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

$$2 \hat{\boldsymbol{\Sigma}}_{\mathrm{b}} \boldsymbol{\beta}^{(m-1)} - \lambda \boldsymbol{\Gamma} - 2\delta \tilde{\boldsymbol{\Sigma}}_{\mathrm{w}} \boldsymbol{\beta} = 0, \qquad \delta \geqslant 0, \qquad \delta(\boldsymbol{\beta}^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}_{\mathrm{w}} \boldsymbol{\beta} - 1) = 0, \qquad \boldsymbol{\beta}^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}_{\mathrm{w}} \boldsymbol{\beta} \leqslant 1, \tag{43}$$

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 $\Sigma_{j=1}^{p} |\hat{\sigma}_j \beta_j|$ 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, $|(2 \hat{\boldsymbol{\Sigma}}_{\mathrm{b}} \boldsymbol{\beta}^{(m-1)})_j| > \lambda \hat{\sigma}_j$. Then it must be the case that $2\delta \tilde{\boldsymbol{\Sigma}}_{\mathrm{w}} \boldsymbol{\beta} \neq 0$. So $\delta > 0$ and $\boldsymbol{\beta}^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}_{\mathrm{w}} \boldsymbol{\beta} = 1$. Then the Karush–Kuhn–Tucker conditions simplify to

$$2 \hat{\boldsymbol{\Sigma}}_{\mathrm{b}} \boldsymbol{\beta}^{(m-1)} - \lambda \boldsymbol{\Gamma} - 2\delta \tilde{\boldsymbol{\Sigma}}_{\mathrm{w}} \boldsymbol{\beta} = 0, \qquad \boldsymbol{\beta}^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}_{\mathrm{w}} \boldsymbol{\beta} = 1, \qquad \delta > 0. \tag{44}$$

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

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

The same set of arguments applied to problem (20) lead to proposition 2, part (b).

A.4. Proof of proposition 3

Consider problem (17) with tuning parameter λ_1 and $k=1$. Then by theorem 6.1.1 of Clarke (1990), if there is a non-zero solution $\boldsymbol{\beta}^*$, then there exists $\mu \geqslant 0$ such that

$$0 \in 2\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\boldsymbol{\beta}^* - \lambda_1\, \boldsymbol{\Gamma}(\boldsymbol{\beta}^*) - 2\mu\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}\boldsymbol{\beta}^*, \tag{45}$$

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 $\mathrm{sgn}(\beta_j)$ if $\beta_j \neq 0$ and is between -1 and 1 if $\beta_j = 0$. Left multiplying expression (45) by $\boldsymbol{\beta}^*$ yields $0 = 2\boldsymbol{\beta}^{*\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\boldsymbol{\beta}^* - \lambda_1\|\boldsymbol{\beta}^*\|_1 - 2\mu\boldsymbol{\beta}^{*\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}\boldsymbol{\beta}^*$. Since the sum of the first two terms is positive (since $\boldsymbol{\beta}^*$ is a non-zero 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 $a = \surd(n\boldsymbol{\beta}^{*\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\boldsymbol{\beta}^*)$. By inspection, $a \neq 0$, since otherwise $\boldsymbol{\beta}^*$ would not be a non-zero solution. Also, let $\lambda_2 = \lambda_1\{(1-ca)/a\}$. Note that $1 - ca = 1/(1+\mu) > 0$, so $\lambda_2 > 0$.

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

$$2\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\hat{\boldsymbol{\beta}} - \lambda_2\, \boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}})\frac{\surd(n\hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\hat{\boldsymbol{\beta}})}{1-\surd(n\hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\hat{\boldsymbol{\beta}})} - 2\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}\hat{\boldsymbol{\beta}}\frac{\surd(n\hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\hat{\boldsymbol{\beta}})}{1-\surd(n\hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\hat{\boldsymbol{\beta}})}, \tag{47}$$

or, equivalently,

$$\begin{aligned} 2c\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\boldsymbol{\beta}^* - \lambda_2\, \boldsymbol{\Gamma}(\boldsymbol{\beta}^*)\frac{ac}{1-ac} - 2c\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}\boldsymbol{\beta}^*\frac{ac}{1-ac} &= 2c\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\boldsymbol{\beta}^* - \lambda_1 c\, \boldsymbol{\Gamma}(\boldsymbol{\beta}^*) - 2c\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}\boldsymbol{\beta}^*\frac{ac}{1-ac} \\ &= 2c\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\boldsymbol{\beta}^* - \lambda_1 c\, \boldsymbol{\Gamma}(\boldsymbol{\beta}^*) - 2c\mu\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}\boldsymbol{\beta}^* \\ &= c\{2\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\boldsymbol{\beta}^* - \lambda_1\, \boldsymbol{\Gamma}(\boldsymbol{\beta}^*) - 2\mu\tilde{\boldsymbol{\Sigma}}_{\mathrm{w}}\boldsymbol{\beta}^*\}. \end{aligned} \tag{48}$$

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

A.5. Proof of proposition 4

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

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

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

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

This follows from the fact that problem (37) reduces to

$$\mathrm{maximize}_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}^{\mathrm{T}}\bar{\mathbf{X}}_1 - \lambda\sum_{j=1}^{p}|\beta_j\hat{\sigma}_j|\right) \text{ subject to } \sum_{j=1}^{p}\beta_j^2\hat{\sigma}_j^2 \leqslant 1, \tag{51}$$

since $(1/\surd n)\mathbf{X}^{\mathrm{T}}\mathbf{Y}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1/2} = \bar{\mathbf{X}}_1(1/\surd 2 - 1/\surd 2)$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{b}} = (1/n)\mathbf{X}^{\mathrm{T}}\mathbf{Y}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X}$.

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

A.6. Proof of equivalence of Fisher's linear discriminant analysis and optimal scoring

Consider the following two problems:

$$\mathrm{maximize}_{\boldsymbol{\beta}\in\mathbb{R}^p}(\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\boldsymbol{\beta}) \text{ subject to } \boldsymbol{\beta}^{\mathrm{T}}(\hat{\boldsymbol{\Sigma}}_{\mathrm{w}} + \Omega)\boldsymbol{\beta} = 1 \tag{52}$$

and

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

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 problems (52) and (53), $\boldsymbol{\Omega}$ is a matrix such that $\hat{\boldsymbol{\Sigma}}_{\mathrm{w}}+\boldsymbol{\Omega}$ is positive definite; if $\boldsymbol{\Omega}=0$ then these two problems reduce to Fisher's LDA and optimal scoring. Optimizing problem (53) with respect to $\boldsymbol{\theta}$, we see that the $\boldsymbol{\beta}$ that solves problem (53) also solves

$$\text{minimize}_{\beta}\left\{-\frac{2}{\surd n}\surd(\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\boldsymbol{\beta})+\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\boldsymbol{\beta}+\boldsymbol{\beta}^{\mathrm{T}}(\hat{\boldsymbol{\Sigma}}_{\mathrm{w}}+\boldsymbol{\Omega})\boldsymbol{\beta}\right\}. \tag{54}$$

For notational convenience, let $\tilde{\boldsymbol{\beta}}=(\hat{\boldsymbol{\Sigma}}_{\mathrm{w}}+\boldsymbol{\Omega})^{1/2}\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\Sigma}}_{\mathrm{b}}=(\hat{\boldsymbol{\Sigma}}_{\mathrm{w}}+\boldsymbol{\Omega})^{-1/2}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}(\hat{\boldsymbol{\Sigma}}_{\mathrm{w}}+\boldsymbol{\Omega})^{-1/2}$. Then, the problems become

$$\text{maximize}_{\tilde{\beta}}(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\mathrm{b}}\tilde{\boldsymbol{\beta}})\text{ subject to } \tilde{\boldsymbol{\beta}}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}=1 \tag{55}$$

and

$$\text{minimize}_{\tilde{\beta}}\left\{-\frac{2}{\surd n}\surd(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\mathrm{b}}\tilde{\boldsymbol{\beta}})+\tilde{\boldsymbol{\beta}}^{\mathrm{T}}(\tilde{\boldsymbol{\Sigma}}_{\mathrm{b}}+\mathbf{I})\tilde{\boldsymbol{\beta}}\right\}. \tag{56}$$

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

$$\tilde{\boldsymbol{\Sigma}}_{\mathrm{b}}\hat{\boldsymbol{\beta}}\left\{1-\frac{1}{\surd(n\hat{\boldsymbol{\beta}}^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_{\mathrm{b}}\hat{\boldsymbol{\beta}})}\right\}+\hat{\boldsymbol{\beta}}=0, \tag{57}$$

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

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

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

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 Res.*, **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**, 989–1010.

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

Breiman, L. and Ihaka, R. (1984) Nonlinear discriminant analysis via scaling and ACE. *Technical Report*. University of California at Berkeley, Berkeley.

Clarke, F. (1990) *Optimization and Nonsmooth Analysis*. Troy: Society for Industrial and Applied Mathematics.

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. Am. Statist. Ass.*, **96**, 1115–1160.

Fan, J. and Fan, Y. (2008) High-dimensional classification using features annealed independence rules. *Ann. Statist.*, **36**, 2605–2637.

Friedman, J. (1989) Regularized discriminant analysis. *J. Am. Statist. Ass.*, **84**, 165–175.

Friedman, J., Hastie, T., Hoefling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. *Ann. Appl. Statist.*, **1**, 302–332.

Gorski, J., Pfeuffer, F. and Klamroth, K. (2007) Biconvex sets and optimization with biconvex functions: a survey and extensions. *Math. Meth. Oper. Res.*, **66**, 373–407.

Grosenick, L., Greer, S. and Knutson, B. (2008) Interpretable classifiers for fMRI improve prediction of purchases. *IEEE Trans. Neur. Syst. Rehabiltn Engng*, **16**, 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. *Ann. Statist.*, **23**, 73–102.

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

Hoefling, H. (2010) A path algorithm for the fused lasso signal approximator. *J. Computnl Graph. Statist.*, **19**, 984–1006.

Hunter, D. and Lange, K. (2004) A tutorial on MM algorithms. *Am. Statistn*, **58**, 30–37.

Johnson, N. (2010) A dynamic programming algorithm for the fused lasso and 10-segmentation.

Jolliffe, I., Trendafilov, N. and Uddin, M. (2003) A modified principal component technique based on the lasso. *J. Computnl Graph. Statist.*, **12**, 531–547.

Krzanowski, W. J., Jonathan, P., McCarthy, W. V. and Thomas, M. R. (1995) Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Appl. Statist.*, **44**, 101–115.

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

Lange, K., Hunter, D. and Yang, I. (2000) Optimization transfer using surrogate objective functions. *J. Computnl Graph. Statist.*, **9**, 1–20.

Leng, C. (2008) Sparse optimal scoring for multiclass cancer diagnosis and biomarker detection using microarray data. *Computnl Biol. Chem.*, **32**, 417–425.

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

Nakayama, R., Nemoto, T., Takahashi, H., Ohta, T., Kawai, A., Yoshida, T., Toyama, Y., Iehikawa, H. and Hasegama, T. (2007) Gene expression analysis of soft tissue sarcomas: characterization and reclassification of malignant fibrous histiocytoma. *Mod. Pathol.*, **20**, 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. *Proc. Natn. Acad. Sci. USA*, **98**, 15149–15154.

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011) Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.*, **39**, 1241–1265.

Shen, H. and Huang, J. Z. (2008) Sparse principal component analysis via regularized low rank matrix approximation. *J. Multiv. Anal.*, **101**, 1015–1034.

Sun, L., Hui, A., Su, Q., Vortmeyer, A., Kotlivarov, 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. *Computnl Statist. Data Anal.*, **52**, 423–437.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. 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. Natn. Acad. Sci. USA*, **99**, 6567–6572.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003) Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.*, **18**, 104–117.

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

Trendafilov, N. and Jolliffe, I. (2007) DALASS: variable selection in discriminant analysis via the LASSO. *Computnl Statist. Data Anal.*, **51**, 3718–3736.

Witten, D. M. (2011) penalized LDA: penalized classification using Fisher's linear discriminant. *R Package Version 1.0*. (Available from `http://cran.r-project.org/web/packages/penalized LDA/index.htm`.)

Witten, D. M. and Tibshirani, R. (2009) Covariance-regularized regression and classification for high dimensional problems. *J. R. Statist. Soc.* B, **71**, 615–636.

Witten, D., Tibshirani, R. and Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.

Xu, P., Brock, G. and Parrish, R. (2009) Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computnl Statist. Data Anal.*, **53**, 1674–1687.

Zhu, J. and Hastie, T. (2004) Classification of gene microarrays by penalized logistic regression. *Biostatistics*, **5**, 427–443.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B, **67**, 301–320.

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal component analysis. *J. Computnl Graph. Statist.*, **15**, 265–286.