Supervised Classification Using Sparse Fisher’s LDA

Irina Gaynanova, James G. Booth and Martin T. Wells

December 11, 2013

Abstract

It is well known that in a supervised classification setting when the number of features, $p$, is smaller than the number of observations, $n$, Fisher’s linear discriminant rule asymptotically achieves Bayes rule. However, there are numerous applications where classification is needed in the $p >> n$ setting. Naive implementation of Fisher’s rule in this setting fails to provide good results because the sample covariance matrix is singular. Moreover, by constructing a classifier that relies on all $p$ features the interpretation of the results is challenging. Our goal is to provide robust classification that relies only on a small subset of features and accounts for the complex correlation structure. We apply an $L_1$ penalty to the discriminant vector to insure sparsity of the solution and use a shrinkage type estimator for the covariance matrix. The resulting optimization problem is solved using an iterative coordinate ascent algorithm. Furthermore, we highlight the difference between the $L_1$ penalized and the $L_1$ constrained versions of the problem and provide justification for feature clustering. The simulation results show that the proposed method performs favorably in comparison to alternatives. The method is used to classify 344 leukemia patients based on 19000 DNA methylation features.

Keywords: Clustering; Coordinate ascent; Discriminant analysis; Duality gap; Methylation; Penalization.

1 Introduction

Linear discriminant analysis (LDA) is a standard method for classification when the number of observations $n$ is much bigger than the number of features $p$. If data follows $p$-variate normal distribution with the same covariance structure across the groups, LDA provides an asymptotically optimal classification rule, meaning that it converges to Bayes rule [Anderson 1984, Chapter 6]. However, it was noted by Dudoit et al. (2002) that naive implementation of LDA for high-dimensional data provides poor classification results in comparison to alternative methods. A rigorous proof of this phenomenon in the case $p >> n$ is given by Bickel and Levina (2004). There are two main reasons for this. First, standard LDA uses the sample covariance matrix to estimate the covariance structure. In high dimensional settings this results in a singular estimator. Secondly, by using all $p$ features in classification, interpretation of the results becomes challenging. Often, only a small
subset of features is relevant and it is of great interest to perform classification and variable selection at the same time.

Several methods have been proposed in the literature to account for these problems. One of the popular approaches is to use the independence rule which overcomes the singularity problem of the sample covariance but ignores the dependency structure. This approach is very appealing because of its simplicity and was encouraged by the work of Bickel and Levina (2004) who showed that it performs better than the standard LDA in the $p >> n$ setting. Examples of diagonal classification methods include Tibshirani et al. (2003); Fan and Fan (2008); Pang et al. (2009); Huang et al. (2010); Witten and Tibshirani (2011).

Unfortunately, independence is only an approximation and it is unrealistic in most applications. Gene interactions, for example, are crucial for the understanding of biological processes and it is important to use this information in classification. Therefore we should aim for better estimators of the covariance matrix instead of using the independence structure. We are not the first to notice that there are drawbacks to the independence approach. Huang et al. (2010) note that the discriminant scores resulting from the diagonal rule are biased. However, instead of changing the rule, they propose adjusting for the bias directly. Fan et al. (2010) argue that it is crucial to take into account the covariance structure. However, they avoid estimating it directly in the $p >> n$ setting by doing preselection of features. Shao et al. (2011) propose thresholding of sample covariance matrix. Cai (2011) estimate the covariance matrix and mean differences directly, and Mai et al. (2012) reformulate the LDA problem using penalized least squares. These approaches, however, are limited to the case of the two groups and the standard LDA formulation instead of Fisher’s version is considered.

The purpose of this work is to extend current methodology in a way that will enable automatic variable selection and account for the complex dependency structure. A motivating example is the analysis of DNA methylation data from patients with Acute Myeloid Leukemia (AML). The ERASMUS data was collected at Erasmus University Medical Center (Rotterdam) and consists of DNA methylation profiles of 344 patients. The ECOG data is obtained from a clinical trial performed by Eastern Cooperative Oncology Group and consists of DNA methylation profiles of 383 patients. Samples from both cohorts were processed according to Thompson et al. (2008). Cluster analysis performed on the ERASMUS data corresponded well to the available biomarker information, providing new insights into the leukemia subtypes based on the methylation patterns (see Figueroa et al. (2010) and Kormaksson et al. (2012) for details). Since only limited biomarker information is available for the ECOG data, it is of great interest to use the clustering information from the ERASMUS data to determine the leukemia subtypes in the ECOG data.

In this work we consider Fisher’s formulation of LDA (Mardia et al. 1979), since it is derived without explicit normality assumption on the data. The sparsity of the solution is enforced by adding an $L_1$ constraint to the objective function and therefore, the optimization problem considered in this paper is similar to the one considered by Witten and Tibshirani (2011). However, we gained additional insights in the proposed method. First, we show that using a diagonal estimate of the covariance structure in Fisher’s LDA (FLDA) with two groups is equivalent to selecting features by using t-statistic with an appropriate threshold and therefore we extend the algorithm to the non-diagonal case. Secondly, we investigate in more detail the optimization aspects of the $L_1$ penalized FLDA problem and show that it is not equivalent to the $L_1$ constrained problem. This has important consequences in terms of solution sparsity and tuning parameter selection.

We propose to evaluate covariance structure through the estimator proposed by Schäfer and
Strimmer (2005), however our algorithm can be used with any other estimator of a general form.

The required algorithmic modification provides additional computational challenges and we use the coordinate ascent algorithm to solve the optimization criteria. Simulation results show significant improvement in misclassification rates over the method proposed by Witten and Tibshirani (2011), especially when the true covariance structure is far from the diagonal.

In our simulation study, we noticed that in certain scenarios it is impossible to achieve a very sparse solution regardless of the choice of the tuning parameter. We show that this phenomenon is due to the duality gap between the \( L_1 \)-penalized and the \( L_1 \)-constrained formulations of the problem. With the success of LASSO (Tibshirani [1996]), it is very common to use the \( L_1 \) penalty as a method to achieve sparsity in the solution (for example Zou et al. [2006] in PCA, Bradley and Mangasarian [1998] in SVM and Bien and Tibshirani [2011] in covariance matrix estimation). The \( L_1 \) penalty in LASSO is motivated by the dual problem where using the \( L_1 \) constraint geometrically means projecting the solution vector onto the subspace that forces certain components to be exactly zero. Since in LASSO the \( L_1 \)-constrained and \( L_1 \)-penalized problems are equivalent, it is natural to expect that adding \( L_1 \) penalty to other objective functions provides the same effect. Unfortunately this is not always the case and we show that this is indeed not true in the FLDA context. We provide an intuitive explanation for this phenomenon and propose using feature clustering as a partial solution.

The rest of the paper is organized as follows. Section 2 provides the formulation of the problem and optimization details. Section 3 describes implementation aspects and justifies clustering. In Section 4 we compare the proposed algorithm with competing methods in simulation studies. Application to DNA methylation data is presented in Section 5. We conclude in Section 6 with a discussion.

2 Methodology

We start by considering the following optimization problem, which corresponds to unconstrained FLDA (Mardia et al. [1979]) when \( p < n \),

\[
\max_v v^t B v \quad \text{subject to} \quad v^t W v = 1.
\]

\( W \) is the \( p \times p \) within-group sum of squares matrix and \( B \) is the \( p \times p \) between-group sum of squares matrix; i.e. \( W = \sum_{i=1}^{g} n_i S_i \), \( B = T - W \), where \( T \) is the total sum of squares matrix, \( n_i \) is the number of observations in the \( i \)th group and \( S_i \) is the sample covariance matrix for the \( i \)th group.

The vector \( v \) that solves this problem is the eigenvector corresponding to the maximum eigenvalue of the rank min \( \min(p, g-1) \) matrix \( W^{-1}B \). Moreover, since \( W \) is a positive definite matrix, the original problem can be rewritten as (Boyd and Vandenberghe [2004])

\[
\max_v v^t B v \quad \text{subject to} \quad v^t W v \leq 1.
\]

Note that \( \frac{1}{n-g} W \) is a sample estimate of the population within-class covariance matrix \( \Sigma_w \) and \( \frac{1}{g-1} B \) is a sample estimate of the population between-class covariance matrix \( \Sigma_b \). Therefore the aim of FLDA is to find a linear combination of features that maximizes the between-group variability with respect to the within-group variability.

One of the problems with the discriminant analysis in the \( p >> n \) setting is that only a relatively small subset of features is relevant. Elimination of features in the FLDA is equivalent to estimating
some of the components of vector \( v \) exactly as zero. As in LASSO [Tibshirani 1996], this can be achieved by introducing an additional \( L_1 \) penalty. In the context of FLDA, this approach was addressed by Witten and Tibshirani [2011] who consider the following optimization problem:

\[
v_{0\lambda} = \arg \max_{v \in \mathbb{R}^p} \left\{ v'^{T} B v - \lambda \sum_{j=1}^{p} |s_j v_j| \right\} \text{ subject to } v'^{T} \text{diag}(W)v \leq 1.
\]

Here \( \lambda \geq 0 \) is a tuning parameter where \( \lambda = 0 \) corresponds to unpenalized version of the estimation procedure. The larger the value of \( \lambda \), the larger the penalty on \( v \) and therefore the smaller the number of non-zero components. Note that each feature of the vector \( v \) is additionally penalized by its sample within-group standard deviation \( s_j \). The solution of Witten and Tibshirani [2011] is restricted to the case in which estimator of \( \Sigma_w \) has a diagonal form. Using a nondiagonal estimator of \( \Sigma_w \) makes the optimization problem more challenging. This drawback is outweighed by the superior performance, which is demonstrated via simulations and a real data application in Sections 4 and 5 respectively.

### 2.1 Estimation of \( \Sigma_w \) and importance of correlation

Since the discovery of the well known Stein phenomenon [Stein 1975], the area of improved estimation for the covariance matrices has been an active research area. A common approach to improve the estimation of the covariance matrix is that of shrinkage estimation, since it is well known that in high dimensions the eigenvalues of the sample covariance matrix are poor estimates for the true eigenvalues. The sample eigenvalues spread over the positive real numbers so that the smallest eigenvalues tend to zero, while the largest tend towards infinity. This causes the sample covariance matrix to become ill-conditioned or even singular. A number of authors have studied the properties of invertible estimates for the covariance matrix, see Bai and Yin [1973], Bickel and Levina [2004], Kubokawa and Srivastava [2008], and Ledoit and Wolf [2003] to name a few. In particular, Kubokawa and Srivastava [2008] develop several estimators that theoretically improve on the Moore-Penrose inverse estimator of the precision matrix, the covariance quantity needed in discriminant analysis.

Some of the main results in Bickel and Levina [2004] rely critically on low rank structure of Moore-Penrose inverse of sample covariance matrix, their argument can not be applied to a positive definite estimators of \( \Sigma_w \). Bickel and Levina [2004, Theorem 1] showed that if \( p/n \to \infty \), then the misclassification rate of LDA that uses the sample covariance matrix goes to 1/2. Their proof is based on using Moore-Penrose inverse of sample covariance matrix which has low rank. Since the diagonal estimator of \( \Sigma_w \) is always positive definite, they show that independence approach is preferable in high-dimensional settings. Indeed, in the case \( \Sigma_w \) is known, the misclassification rate of independence rule is always greater than or equal to the misclassification rate of LDA that uses \( \Sigma_w \). The following example illustrates that correlations are beneficial for classification if \( p \) is large and the number of relevant features (\( \equiv r \)) is small, in contrast to Bickel and Levina [2004].

Consider classification between two groups with the mean of the first group \( \mu_1 = 0 \) and the mean of the second group \( \mu_2 = \delta \). Also assume all the variables \( x_j \) are scaled to have standard deviation one within each population. If the variables are jointly normal and equally correlated, i.e. \( \text{cor}(x_i, x_j) = \rho \) for all \( i \neq j \), then the correlation is beneficial if \( \rho \) is negative or if

\[
\rho > \frac{(\sum_{j=1}^{p} \delta_j)^2 - \sum_{j=1}^{p} \delta_j^2}{(p-1) \sum_{j=1}^{p} \delta_j^2}
\]
Assume now that $\delta_j = 0$ for all $j > r$ and $\delta_j = C$ for all $j \leq r$, where $r \leq p$. Then (2) can be rewritten as

$$\rho > \frac{(\sum_{j=1}^{r} C)^2 - \sum_{j=1}^{r} C^2}{(p-1) \sum_{j=1}^{r} C^2} = \frac{r-1}{p-1}.$$  (3)

In modern high-dimensional applications it is common to assume that only small subset of features is relevant for the classification, i.e. $r \ll p$. As a consequence $\frac{r}{p} = c_\epsilon$, where $c_\epsilon$ is a very small constant. It follows from (3) that even small values of $\rho$ lead to improvements in squared distance.

Therefore, instead of applying discriminant analysis with $W = \sum_{i=1}^{g} n_i S_i$ or assuming independence, we use the following estimator:

$$\tilde{W} = \sum_{i=1}^{g} n_i \tilde{S}_i$$

where $\tilde{S}_i = \tau_i T_i + (1 - \tau_i) S_i$ (Schäfer and Strimmer 2005), $T_i$ is a diagonal, unequal variance estimate and $S_i$ is the sample covariance matrix for the $i$th group. There are several advantages to using this estimator. First, the resulting matrix $\tilde{W}$ is always positive definite and it preserves the correlation structure of the data. Secondly, an optimal $\tau_i$ can be chosen according to Ledoit and Wolf 2004 that guarantees minimal MSE under the existence of the first two moments. Therefore there are no strong distributional assumptions on the data. Finally, given the simple form of the estimator, $\tilde{W}$ can be computed easily and quickly for the large values of $p$. This approach is similar to the one used in regularized discriminant analysis (Friedman 1989; Guo et al. 2007). Aside from the feature selection procedure, our approach is different in that each within-group covariance matrix is shrunk towards a diagonal estimate instead of the identity matrix, and the shrinkage parameter $\tau_i$ is selected automatically.

Alternative approaches can be taken to estimate the covariance structure $\Sigma_w$. Moreover, it is desirable for the estimate to be both positive definite and sparse. Here by sparse we mean that certain entries of the covariance matrix are estimated exactly as zero. There is an extensive literature on covariance matrix estimators that achieve one of these goals. However, to our knowledge only limited methodology is available for achieving both. Among recent advances are methods proposed by Bien and Tibshirani 2011 and Rothman 2012. Unfortunately, the estimators considered by these authors are very computationally intensive and do not scale well to high-dimensional data sets.

### 2.2 Algorithm description of $L_1$-penalized version

After we have defined the appropriate covariance estimator, $\tilde{W}$, the penalized Fisher’s linear discriminant problem can be expressed in the following form:

$$\max_v \left\{ v^t B v - \lambda \sum_{j=1}^{p} |s_j v_j| \right\} \text{ subject to } v^t \tilde{W} v \leq 1.$$  (4)

Remark: Though we use $\tilde{W}$ to find the discriminant vector $v$, the proposed algorithm can be applied to any positive definite estimator of $\Sigma_w$. 

Before describing the computational aspects, we note that there are problems with solving the $L_1$-penalized version of the algorithm instead of the $L_1$-constrained version. In our simulations, we noticed that there seemed to be no value of $\lambda$ that resulted in a small number of non-zero components. That is, the solution path is not smooth in $\lambda$. Specifically, there exists a $\lambda_0$ such that for all $\lambda < \lambda_0$ the solution $v_0$ has at least $m$ non-zero components and, for all $\lambda \geq \lambda_0$, $v_0$ is exactly zero. This phenomenon is illustrated in Figure 4 of Section 4, where more details about the simulations are given. This behavior is problematic if our goal is to perform a very sparse subset selection. A theoretical explanation for the behavior is provided in Sections 2.4 and 2.3. Interestingly, although this phenomenon creates challenges in terms of variable selection, the simulation studies suggest that the $L_1$-penalized version of the algorithm still has good prediction properties.

The optimization problem (4) is non-convex and therefore we can not apply standard convex optimization theory. Following Witten and Tibshirani (2011), we note that (4) can be recast as a biconvex optimization problem

$$\max_{u,v} \left\{ 2u^t B^{1/2} v - \lambda \sum_{j=1}^p |s_j v_j| - u^t u \right\} \text{ subject to } v^t W v \leq 1,$$

since maximizing with respect to $u$ gives $u = B^{1/2} v$.

The problem (5) is convex with respect to $u$ when $v$ is fixed and is convex with respect to $v$ when $u$ is fixed. This property allows the use of Alternate Convex Search (ACS) to find the solution (Gorski et al., 2007, Section 4.2.1). Although in our context ACS ensures that all accumulation points are partial optima and have the same function value (Gorski et al., 2007, Theorem 4.9), the convergence to the global optima is not guaranteed due to non-convexity of (5).

Starting with an initial value $v^{(0)}$ the algorithm proceeds by iterating the following two steps:

**Step 1**

$$u^{(m)} = \arg \max_u \left\{ 2u^t B^{1/2} v^{(m)} - u^t u \right\} = B^{1/2} v^{(m)}$$

**Step 2**

$$v^{(m+1)} = \arg \max_v \left\{ 2(u^{(m)})^t B^{1/2} v - \lambda \sum_{j=1}^p |s_j v_j| \right\} \text{ subject to } v^t W v \leq 1.$$  

The main challenge is to solve Step 2. Following Witten and Tibshirani (2011, Proposition 2), it is useful to reformulate the problem as

$$q^{(m+1)} = \arg \max_q \left\{ 2(u^{(m)})^t B^{1/2} q - \lambda \sum_{j=1}^p |s_j q_j| - q^t W q \right\}$$  

where, if $q^{(m+1)} = 0$, then $v^{(m+1)} = 0$, else $v^{(m+1)} = \frac{q^{(m+1)}}{\sqrt{q^{(m+1)^t} W q^{(m+1)}}}$. Since problem (6) is convex with respect to $q$, the solution $q^{(m+1)}$ satisfies (Boyd and Vandenberghe, 2004)

$$2B^{1/2} u^{(m)} - 2W q^{(m+1)} - \lambda \Gamma = 0,$$  

(7)
where $\Gamma$ is a $p$-vector and each $\Gamma_j$ is a subgradient of $|s_j q_j|$, i.e. $\Gamma_j = s_j$ if $q_j^{(m+1)} > 0$, $\Gamma_j = -s_j$ if $q_j^{(m+1)} < 0$ and $\Gamma_j$ is between $-s_j$ and $s_j$ if $q_j^{(m+1)} = 0$.

If $W$ is a diagonal matrix, then $w_{ji} = 0$ for $i \neq j$ and therefore the solution to (7) (Witten and Tibshirani 2011) is

$$q_j^{(m+1)} = \frac{S \left( \left( B^{1/2} u^{(m)} \right)_j, \frac{\lambda s_j}{2} \right)}{w_{jj}},$$

where $S$ is a soft-thresholding operator, i.e. $S(x, a) = \text{sign}(x)(|x| - a)_+$. Note that each component of vector $q^{(m+1)}$ does not depend on the others and therefore implementation of this step is straightforward. If, however, $W$ has a general form, solving (7) with respect to individual components of the vector $q^{(m+1)}$ gives

$$q_j^{(m+1)} = \frac{S \left( \left( B^{1/2} u^{(m)} \right)_j - \sum_{i \neq j} w_{ji} q_i^{(m+1)}, \frac{\lambda s_j}{2} \right)}{w_{jj}}.$$  

(8)

Because of the additional term, $\sum_{i \neq j} w_{ji} q_i^{(m+1)}$, each component of vector $q^{(m+1)}$ depends on the other components and therefore the solution is not available in closed form. It is proposed to use coordinate update to overcome this challenge.

Let $f(q) = 2u^t B^{1/2} q - \lambda \sum_{j=1}^p |s_j q_j| - q^t W q$ be the objective function and note that $f$ can be written as

$$f(q_1, ..., q_p) = f_0(q_1, ..., q_p) + \sum_{j=1}^p f_j(q_j),$$

where $f_0(q_1, ..., q_p) = 2u^t B^{1/2} q - q^t W q$ is concave and differentiable in $q$ and $f_j(q_j) = -\lambda |s_j q_j|$, $j = 1, ..., p$ are concave in $q$. It was established by Tseng (1988) that coordinate ascent methods converge for such functions.

The resulting algorithm to solve (1) is the following:

1. Find $W = \tilde{W}$ as described in Section 2.1.
2. Choose $\lambda$ as described in Section 3.1.
3. Initialize $v$ as the eigenvector corresponding to the largest eigenvalue of $W^{-1} B$ and normalize it so that $v^t W v = 1$. Set $q^{\text{old}} = v$. These are the starting values for the optimization algorithm.
4. Iterate until convergence

   (a) $u \rightarrow B^{1/2} u$

   (b) Iterate until convergence

   for $l = 1 : p$

   $j = \text{sample}(1 : p)$, sample without replacement

   $$q_j^{\text{new}} = \frac{S \left( \left( B^{1/2} u \right)_j - \sum_{i \neq j} w_{ji} q_i^{\text{old}}, \frac{\lambda s_j}{2} \right)}{w_{jj}}$$

   $q_j^{\text{old}} = q_j^{\text{new}}$  

(9)
Note that instead of doing a linear update of all \( p \) features, a random update is performed in (9). This guarantees a faster convergence rate of the coordinate ascent method (Shalev-Shwartz and Tewari 2009). After each iteration convergence is determined by looking at the sum of absolute deviations from the previous solution, \( D = \sum_{j=1}^{p} |q_j^{\text{new}} - q_j^{\text{old}}| \). Convergence is assumed to be achieved when \( D \) is smaller than the predefined value \( \epsilon > 0 \).

2.3 Discriminating between the two groups using diagonal estimator of \( \Sigma_w \)

In this section we show that when there are two groups, selecting features using penalized FLDA with \( \text{diag}(W) \) is equivalent to selecting features with largest absolute value of t-statistic.

Since \( g = 2 \), the rank of \( B \) is equal to one and therefore \( B = \gamma l^T \), where \( \gamma \) is the only positive eigenvalue of \( B \), and \( l \) is the corresponding eigenvector, so \( l^T l = 1 \). Without loss of generality the features of the eigenvector \( l \) are ordered, \( |l_1| \geq |l_2| \geq \ldots \geq |l_p| > 0 \), and all of the components of vector \( l \) are non-zero. Indeed, if \( |l_j| = 0 \) for some \( j \) then from (1) it automatically follows that \( v_0^j = 0 \).

Write \( W = \text{diag}(s_j^2) \). Consider new variable \( z \) such that \( z_j = v_j s_j \) or equivalently \( z = W^{1/2}v \). Then problem (11) can be rewritten as \( v_{0\lambda} = W^{-1/2}z_{0\lambda} \), where

\[
\begin{align*}
    z_{0\lambda} &= \arg \max_{z \in \mathbb{R}^p} z^T W^{-1/2} B W^{-1/2} z - \lambda \sum_{j=1}^{p} |z_j| \text{ subject to } z^T z \leq 1. 
\end{align*}
\]

Note that vectors \( v_{0\lambda} \) and \( z_{0\lambda} \) have the same support. Moreover, setting \( B' = W^{-1/2} B W^{-1/2} \) it follows that \( \text{rank}(B') = \text{rank}(B) = 1 \), and therefore without loss of generality we can restrict attention to the case when \( W = I \). Therefore the optimization problem has the following form:

\[
\begin{align*}
    v_{0\lambda} &= \arg \max_{v \in \mathbb{R}^p} v^T B v - \lambda ||v||_1 \text{ subject to } v^T v \leq 1. 
\end{align*}
\]

Let \( f(v) = v^T B v - \lambda ||v||_1 \) be the objective function.

**Definition 1.** For all \( j, 1 \leq j \leq p \), define the sets

\[
C_j = \{ v \in \mathbb{R}^p : v \text{ has exactly } j \text{ non-zero components} \}
\]

and
\[
A_j = \{ v \in \mathbb{R}^p : \text{first } j \text{ components of vector } v \text{ are non-zero, all other components are zero} \}.
\]

**Remark:** \( A_j \subset C_j \) for \( 1 \leq j \leq p - 1 \) and \( A_j = C_j \) for \( j = p \).

Also define \( A_0 = \{ 0 \} \), the set that consists of the \( p \)-dimensional zero-vector. Consider

\[
F_j = \max_{v \in A_j \cup A_0} f(v) \text{ subject to } v^T v \leq 1.
\]

Note that \( F_0 = 0 \) corresponds to \( f(v) \) evaluated at \( v = 0 \) and \( F = \max_{v^T v \leq 1} f(v) \) corresponds to the original problem (11).
Proposition 1. Under the constraint $v^T v \leq 1$

$$\max_{v \in C_j \cup A_0} f(v) = \max_{v \in A_j \cup A_0} f(v) = F_j$$

This proposition shows that in case $W = I$ selecting features with an $L_1$ penalty is equivalent to selecting features with largest values of $|l_i|$, i.e. with largest absolute differences between the sample means of the two groups. In case $W = \text{diag}(s_j^2)$ notice that $l' = W^{-1/2}l$ is the eigenvector of $B' = W^{-1/2}BW^{-1/2}$. From (10) it follows that selecting features with an $L_1$ penalty is equivalent to selecting features with largest values of $|l'_i| = \frac{|l_i|}{s_i}$, i.e. with largest absolute values of the t-statistic.

To summarize, features selected by solving problem (1) are the same as features selected by the t-test with an appropriate threshold.

2.4 $L_1$ constraint versus $L_1$ penalization

So far we have considered $L_1$-penalized Fisher’s LDA, i.e. the additional $L_1$ norm was incorporated into the objective function:

$$v_{0\lambda} = \arg \max_{v \in \mathbb{R}^p} \left\{ v^T Bv - \lambda ||v||_1 \right\} \text{ subject to } v^T \tilde{W} v \leq 1,$$

where we omit the additional penalty weights for simplicity of illustration.

A closely related problem can be formulated by incorporating an $L_1$ norm directly in the constraint set:

$$v_{0t} = \arg \max_{v \in \mathbb{R}^p} v^T Bv \text{ subject to } v^T \tilde{W} v \leq 1, ||v||_1 \leq t,$$

where $t \geq 0$ is a tuning parameter. Note that we only need to consider $t \geq 1$ since smaller values of $t$ result in $v_{0t}$ such that $v_{0t}^T \tilde{W} v_{0t} < 1$. Solving problem (13) is computationally more challenging than solving problem (12) since it additionally involves a binary search before performing update (9).

Unlike LASSO, there exists a duality gap between problems (12) and (13). For every $\lambda \geq 0$ we can find $t \geq 1$ such that $v_{0\lambda} = v_{0t}$. The contrary, however, is not generally true.

First, we will show that every solution to (12) is the solution to (13). Indeed, fix any $\lambda \geq 0$ and let $v_{0\lambda}$ be the corresponding solution of (12). Then it follows that for any $v$ such that $v^T \tilde{W} v \leq 1$

$$v_{0\lambda}^T B v_{0\lambda} - \lambda ||v_{0\lambda}||_1 \geq v^T Bv - \lambda ||v||_1.$$  

(14)

Now set $t = ||v_{0\lambda}||_1$ in (13). Then for each $v$ such that $v^T \tilde{W} v \leq 1$ and $||v||_1 \leq t$ we have from (14)

$$v_{0\lambda}^T B v_{0\lambda} \geq v^T Bv + \lambda (||v_{0\lambda}||_1 - ||v||_1) = v^T Bv + \lambda (t - ||v||_1) \geq v^T Bv.$$

This means $v_{0t} = v_{0\lambda}$ and $v_{0\lambda}$ is the solution to (13).

We now explain why not every solution to (13) is the solution to (12) and discuss the implications for the algorithm in terms of performance and sparsity level of the solution. In LASSO case, the optimization problem is convex and therefore there is a guarantee that there is no duality gap between $L_1$-penalized and $L_1$-constrained versions of the algorithm (Bertsekas 1999 Proposition 5.2.1). In Fisher’s LDA the problem is non-convex and therefore this guarantee no longer applies.

To illustrate when and why the duality gap appears, consider the case $p = 2, g = 2$ and $\Sigma_w = I$. In this case $B$ is a $2 \times 2$ matrix with its rank equal to one. Therefore it is completely determined
by its positive eigenvalue and the corresponding eigenvector. Set the eigenvalue equal to one and consider two scenarios. In the first scenario, initialize the eigenvector \( l \) of matrix \( B \) as \((0.2, 0.8)\) and in the second scenario as \((0.5, 0.6)\). In both cases the vectors are normalized to have \( L_2 \) norm exactly equal to one. Note that the features of the eigenvectors are not equal. However, the absolute difference between them is smaller in the second scenario \((0.1 \text{ versus } 0.6)\).

Following Bertsekas (1999, Chapter 5), we use a geometry-based approach to visualize the relationship between solutions to \( L_1 \)-constrained and \( L_1 \)-penalized problems. Consider set \( S \) of constrained pairs \( (f(v) = -v^T B v, h(v) = ||v||_1) \) as \( v \) ranges over \( \mathbb{R}^p \) with \( v^T v \leq 1 \). Problem (13) can be formulated as a minimal common point problem: finding a point with a minimal \( f \)th coordinate among all points common to both \( S \) and halfspace \( h \leq t \). Furthermore, if it is possible to construct a hyperplane supporting \( S \) from below at this point, then there exists \( \lambda > 0 \) such that solutions to (12) and (13) are the same. For both scenarios, the corresponding sets \( S \) are plotted in Figure 1. In the first case, for each \( t \geq 1 \) a supporting hyperplane can be constructed at each minimal point of the set \( S \) that has coordinate \( h \leq t \). In the second case, no supporting hyperplane can be constructed at the minimal point of the set \( S \) with \( h \leq 1 \) because it has to lie below the point \((0, 0)\) and the minimal point corresponding to \( h = 1.4 \). Moreover, such a supporting hyperplane does not exist for the values of \( t \) ranging from 1 to 1.4 and therefore there exists no \( \lambda \) such that \( v_{0\lambda} = v_0 \) for those \( t \). The difference between the two scenarios is only in the matrix \( B \), or more precisely, in how close the features of the eigenvector of the matrix \( B \) are to each other.

![Figure 1: Visualization of the set \( S \), \( l \) is the eigenvector of matrix \( B \).](image)

There is another interesting implication that can be observed from the shape of the sets \( S \). If \( t \) takes a value that is smaller than a certain threshold (like 1.4 in the second scenario), the only supporting hyperplane subject to the constraints intersects the set \( S \) at the point corresponding to \( f = 0 \) and \( h = 0 \). Therefore, for such \( t \) the corresponding solution to the dual problem (12) is exactly \( v = 0 \). In other words, the solution path of (12) is not smooth in \( \lambda \) and a sudden drop in the number of non-zero features can be observed. This is consistent with the results seen in simulations.

The duality gap between the problems (12) and (13) is closely related to undirect restriction on the range of values of \( t \), which leads to the restriction on the sparsity level we can achieve by varying the tuning parameter \( \lambda \). We have illustrated that the eigenstructure of \( B \) plays an important role
in this phenomenon. Further we derive sufficient conditions for observing the drop in the number of features for the simple case of two groups and diagonal estimator of $\Sigma_w$ (Proposition 4). The two scenarios described above are consistent with this result: the conditions of Proposition 4 are satisfied for the second scenario but not for the first.

Consider the simple case of discriminating between the two groups and diagonal estimator of $\Sigma_w$ (without loss of generality we consider $\Sigma_w = I$). Following the notation of Section 2.3, we derive an upper and a lower bound for the maximum $F_j$ of the objective function $f(v) = v^T B v - \lambda ||v||_1$ under the constraints $v^T v \leq 1$ and $v \in A \cup A_0$ ($v$ is either identically 0 or has $j$ non-zero components).

**Proposition 2.** $F_j \leq F_j \leq F_j^*$ where

$$F_j = \gamma ||v||_2^2 - \lambda \frac{||v||_1}{||v||_2}$$

and

$$F_j^* = \max \left( 0, \gamma ||v||_2^2 - \lambda \frac{||v||_2}{||v||_1} \right).$$

Here $l_i^j = l_i I\{1 \leq i \leq j\}$ and $l$ is the eigenvector of $B$.

It follows from Proposition 1 that $F = \max(F_0, F_1, \ldots, F_p) = \max(0, F_1, \ldots, F_p)$ and $v_0 \in \bigcup_{i=0}^p A_i$. Since $A_i$ are disjoint by construction, this means that there exists $l$ such that $v_0 \in A_l$ and $v_0 \notin A_j$ for all $j \neq l$. Therefore to quantify the drop in number of features, we need to show that there exists $m$ such that $v_0 \in A_0 \cup \bigcup_{i=m}^p A_i$ regardless of the choice of the tuning parameter $\lambda$. In other words, the solution to (11) has at least $m$ non-zero features or is exactly 0.

We first show that, depending on the value of $\lambda$, there exists $m_\lambda \geq 1$ such that $v_0 \in A_0 \cup \bigcup_{i=m_\lambda}^p A_i$.

**Proposition 3.** Define $m_\lambda = \min\{j : 1 \leq j \leq p\}$ such that $||v||_2 > \frac{\lambda}{\gamma |l_1|}$. Then $v_{0\lambda} \in A_0 \cup \bigcup_{i=m_\lambda}^p A_i$. If there is no $j$ for which $||v||_2 > \frac{\lambda}{\gamma |l_1|}$ holds, then $v_{0\lambda} = 0$.

**Remark:** For $\lambda < \gamma |l_1|$, the value of $m_\lambda$ increases with $\lambda$. Moreover, if $\lambda \geq \gamma |l_1|$ then $v_{0\lambda} = 0$.

**Proposition 4.** Define $m'$ as $\max\{j : 1 \leq j \leq p - 1\}$ for which there exists $r > j$ such that $||v||_2 \leq \frac{||v||_2}{||v||_1}$ and $m' = 0$ otherwise. Then $v_{0\lambda} \in A_0 \cup \bigcup_{i=m'}^p A_i$ regardless of the choice of the tuning parameter $\lambda$.

Combining Propositions 3 and 4 we have that for any $\lambda < \gamma |l_1|$, the solution $v_{0\lambda} \in A_0 \cup \bigcup_{i=\max(m'+1, m_\lambda)}^p A_i$ and if $\lambda \geq \gamma |l_1|$ then $v_{0\lambda} = 0$. Therefore, for the two groups case with $\Sigma_w = I$, there is a theoretical explanation for the observed behavior that the solution $v_{0\lambda}$ to problem (11) is either $v_{0\lambda} = 0$ or has at least $m$ non-zero components, where $m \geq \max(m'+1, m_\lambda)$.

We performed several simulations to investigate how the value of $\max(m'+1, m_\lambda)$ varies with $p$ and how close the bound we obtained is to the value of $m$ observed empirically. For this purpose we generated $l$ as a random vector with each component $l_i$ coming from the uniform distribution on $[0, 1]$. Afterwards, we standardized $l$ to have $L_2$ norm equal to one and ordered the features so that $|l_1| \geq |l_2| \geq \ldots \geq |l_p| > 0$. The results are shown in Figure 2. Dots correspond to the number of features selected by the algorithm and the dashed line corresponds to the bound $\max(m'+1, m_\lambda)$.

The simulations indicate that the value of $m$ increases with the value of $p$. Since in both cases the features $l_i$ were generated uniformly at random, the observed relationship between $m$ and $p$
can not be explained by the structural differences in $l$. However, $l$ is restricted to have $L_2$ norm equal to one, which means that the increase in $p$ leads on average to a decrease in the differences between the individual features of $l$. Therefore we argue that the larger are the differences between the features of $l$, the higher level of sparsity can be achieved by solving $L_1$-penalized problem \( (12) \). Further in Section 3.3 we show that the theoretical bound $m'$ from Proposition 4 is consistent with this intuition.

3 Implementation

3.1 Selection of the tuning parameter $\lambda$

It is traditional to choose the tuning parameter $\lambda$ by cross-validation. However it is not clear in this context what value should be considered as the upper bound that sets all the components of the vector $v$ to 0. Here we provide some intuition on how this value is chosen.

From the form of the update on $q$ given in (8), note that all components are set to zero if and only if

$$\max_j \left| (2B^{1/2}u^{(m)})_j / s_j \right| \leq \lambda. \quad (15)$$

Since $u = B^{1/2}v$, (15) can be rewritten as

$$\max_j \left| (2Bv^{(0)})_j / s_j \right| \leq \lambda,$$

where $v^{(0)}$ is the initial value of $v$. This leads to an upper bound on $\lambda$:

$$\lambda_{\text{max}} = 2 \max_j \left| (Bv^{(0)})_j / s_j \right|.$$

Note that this bound depends on the $v^{(0)}$, which is natural since $v^{(0)}$ corresponds to the solution of Fisher’s discriminant problem without the $L_1$ constraint. Cross-validation is then performed on a grid over the set $[0, \lambda_{\text{max}}]$. 

Figure 2: Features selected by the algorithm vs theoretical threshold.
3.2 Producing several discriminant vectors

Previously we have only considered the problem of finding the first discriminant vector. If there are more than two groups it is often the case that this is not enough for classification, and additional discriminant vectors should be evaluated. Witten and Tibshirani (2011) address this problem by adjusting the matrix $B$. Let $X$ be an $n \times p$ standardized data matrix and let $Y$ be an $n \times g$ group indicator matrix where $Y_{ig} = 1$ if observation $i$ is in group $g$ and 0 otherwise. They consider

$$B_l = X^T Y (Y^T Y)^{-1/2} P_{l+} (Y^T Y)^{-1/2} Y^T X,$$

where the subscript $l$ corresponds to the $l$th discriminant vector, $P_1^\perp = P_1 = I$ and $P_1^\perp$ is an orthogonal projection matrix into the space orthogonal to $(Y^T Y)^{-1/2} Y^T v_i$ for all $i < l$. Indeed, $Y^T Y$ is a $g \times g$ diagonal matrix where each diagonal element is the number of observations in the corresponding group and $Y^T X$ sums the features of $X$ by group.

Though this approach guarantees orthogonality of discriminant vectors, it also provides additional challenge in interpretation since the same features may be involved in more than one vector. A simpler approach is to find discriminant vectors by reducing the feature space sequentially. After the first discriminant vector is produced, the features corresponding to its non-zero components are eliminated and then the algorithm is repeated on the remaining features.

3.3 Clustering

Let us take a closer look at the conditions of Proposition 4. The larger the value of $m'$, the less sparse are the solutions obtained by varying $\lambda$. To have a large value for $m'$, we need to be able to find a large value of $j$ such that there exists $r > j$ with $\|l^r\|_2 \leq \frac{\|l^r\|_2}{\|l^r\|_1}$, where $l$ is the eigenvector of $B$. This inequality can be rewritten as $\|l^r\|_2|l_1|\|l^r\|_1 \leq \|l^r\|_2|l^r|_2$. Note that, given the ordering of the features of the vector $l$ and the fact that $r > j$, we always have $\|l^r\|_2 \leq \|l^r\|_2$ and $|l_1|\|l^r\|_1 \geq \|l^r\|_2$. Hence the difference between $|l_1|\|l^r\|$ and $\|l^r\|_2$ should be relatively small in comparison to the difference between $\|l^r\|_2$ and $\|l^r\|_2$. It follows that the closer the features of the vector $l$ are to each other, the more likely this inequality holds leading to a larger value of $m'$.

The fact that the eigenvector $l$ corresponds to the differences between the group means leads to a more intuitive interpretation of Proposition 4. If a group of features provides similar discriminant power in terms of having similar values for the differences between the two groups, the penalized criterion either selects the whole group or eliminates the whole group.

This observation suggests the following strategy. Use distances between the groups to cluster the features and create new meta-features as the averages over features belonging to each cluster. Then apply the algorithm to the meta-features. There are several advantages to this procedure. First, the dimensionality of the original problem is reduced from $p$ to the number of clusters $k$. Secondly, the procedure is justifiable from a biological perspective, since it reflects a reluctance to choose only one features among the group of features that have very similar behavior. Finally, the differences between the meta-features are larger than the differences between the original features, which helps to overcome the problem of finding a sparse solution discussed above.

4 Simulation results

The following classification methods are considered for comparison: penalized FLDA proposed by Witten and Tibshirani (2011); its modification described in this work; and support vector machines
with an $L_1$ penalty. In what follows, we refer to these methods as FLDAdiag, FLDA and SVM respectively. For the simplicity of illustration, the comparison is performed for the case of the two groups assuming the data comes from the multivariate normal distribution. We use the R package "penalizedLDA" for FLDAdiag [Witten 2011] and "penalizedSVM" for SVM [Becker et al. 2010]. In all cases, the tuning parameter is selected by cross-validation.

The following covariance structures are considered:

1. Diagonal unit variance, $\Sigma_w = I$

2. Blockdiagonal with a network structure. We randomly choose positions for 40 blocks each of size 4. Within each block the offdiagonal elements are set to 0.75. Five pairs of blocks were chosen randomly to have a correlation of 0.7 between their elements. This initialization is an attempt to mimic a network structure in which there is strong correlation between elements that are close to one another in addition to strong correlation between certain groups of elements that are not necessarily close. Our motivation for considering this structure is based on Xiao [2011], who consider spatial correlations between genes linearly separate on the chromosome but spatially close with respect to its three-dimensional structure.

3. Simulated from DNA methylation data. The covariance matrix $\Sigma_w$ is estimated according to Schäfer and Strimmer [2005] based on the features selected from the ERASMUS DNA methylation data.

The clustering idea described in Section 3.3 is not implemented in the simulation studies to make the comparisons between different methods easier. For each covariance structure, simulations are performed over 25 iterations. The training set has 100 observations per group and the test set has 500 observations per group. The dimension $p$ is set to 800. At each iteration, the training set and the test set are generated from a multivariate normal distribution with means $\mu_1$ and $\mu_2$ and covariance structure $\Sigma_w$. The mean of the first group is set to $\mu_1 = 0$, while the mean of the second group is non-zero for the first $r = 80$ features (with values ranging from 0.2 to 0.6) and 0 everywhere else. This configuration reflects the $p >> n$ framework with only 10% relevant features.

**Remark:** The choice of $\mu_2$ is influenced by the Example 11.8.1 from Mardia et al. [1979], where it is shown that the correlation is especially influential in classification when there is a difference in the components of the mean vector.

The simulation results are summarized in Table 1. Errors correspond to average percent of misclassified observations, features show how many variables on average were selected by each algorithm and correct features indicates how many features were selected out of original $r = 80$. Mean values are reported and empirical standard deviations over 25 iterations are shown in brackets. The misclassification errors are given as percentages.

Since FLDAdiag explicitly uses the diagonal estimate of the covariance structure, we expected FLDA to perform worse for the case $\Sigma_w = I$. However, the performance of the two methods is the same. SVM performs much worse than FLDA with this covariance structure.

FLDA performs slightly better for the blockdiagonal covariance matrix with network structure. However, the most significant difference is observed for the covariance structure simulated from the real data. In this case both FLDAdiag and SVM have much higher misclassification rates than FLDA.

To make sure that the poor performance of FLDAdiag is not affected by the particular choice of 800 features, we independently generated additional 6 covariance structures: 3 from ERASMUS
Table 1: Comparison of FLDA, FLDAdiag and SVM.

|                | Covariance matrix | FLDA | FLDAdiag | SVM  |
|----------------|-------------------|------|----------|------|
| **Errors**     | Diagonal          | 7.1(1)| 7.1(1)   | 17.3(2.1) |
|                | Blockdiagonal     | 17.9(1.7) | 19.5(2.3) | 25(2.5) |
|                | Simulated         | 3.4(0.6) | 32.6(4.6) | 13.4(2.5) |
| **Features**   | Diagonal          | 110(19) | 127(32)  | 94(36) |
|                | Blockdiagonal     | 138(47) | 116(36)  | 133(29) |
|                | Simulated         | 365(52) | 114(55)  | 66(12) |
| **Correct**    | Diagonal          | 60(4)  | 63(4)    | 35(9) |
|                | Blockdiagonal     | 63(8)  | 61(7)    | 30(5) |
|                | Simulated         | 69(4)  | 61(9)    | 23(4) |

DNA methylation data and 3 from ERASMUS gene expression data. Each structure was generated using a different set of 800 features and the prediction error for FLDA and FLDAdiag was calculated based on the 10 iterations. The results are presented in Figure 3. Note that FLDA performs consistently better than FLDAdiag with approximately a 3-fold reduction in classification errors.

Though our method results in good prediction performance, the number of selected features is still relatively high in comparison to the number of truly different features. This is not a drawback of cross-validation, but is an intrinsic property of the $L_1$ penalized problem we are solving. As we discuss in Section 2, very sparse solutions are not always obtainable. Indeed, we performed a separate study using one of the simulated covariance matrices to evaluate the number of non-zero components selected by the algorithm versus the value of the tuning parameter. The results are shown in Figure 4. The drop is observed at around 300 features which is only somewhat less that the mean value for selected features in Table 1.

This property, however, does not significantly affect the performance of the method. Figure 5 illustrates how the misclassification rate varies with the value of the tuning parameter $\lambda$.
**Figure 4:** Number of non-zero features vs tuning parameter.

**Figure 5:** Percent of misclassified observations in the test set of size 1000.
5 Application to DNA Methylation Data

We applied our approach to the ERASMUS methylation data, where \( p = 18954 \). The ECOG data was used as a test set. Since t.8.21 (AML1/ETO) and inv.16 (CBFb/MYH11) biomarkers were available for both datasets, they were used to assess the performance of the algorithm. Clustering was performed using the k-means algorithm [Hartigan and Wong, 1979]. The number of clusters was chosen to be \( k = 200 \) (approximately 1% of the number of original features) and the number of random starts was set to 1000. Note that even with such strong dimension reduction, \( k \) is still much bigger than \( n \) since there are 52 patients in both groups in the ERASMUS data. After transforming the features, the discriminant algorithm was applied using different values of the tuning parameter \( \lambda \). From Figure 6 we can see that the drop in number of selected features is no longer observed. This is consistent with results of Section 2 where we explained why separating the features helps to achieve sparser solutions.

Based on the cross-validation results, the tuning parameter \( \lambda = 0.24 \) was chosen, which corresponds to selecting 5 clusters with centers at 1.8, -2.2, -1.7, 2.4 and 1.5. These 5 selected clusters contain only 79 out of 18954 original features. Recall that it is not possible to select fewer than 300 features if no clustering is performed. The classification error on the ECOG data using these 5 clusters is 0. In contrast, FLDAdiag applied to the same data chooses the tuning parameter that selects 32 clusters (1663 original features). To investigate the effect of the number of clusters, \( k \), the same data was analyzed with \( k = 500 \). In this case the drop in features is again not observed, after cross-validation the algorithm selects the tuning parameter \( \lambda = 0.133 \), which corresponds to selecting 7 clusters containing 73 original features. The prediction error is again 0. Among these 73 features 51 were among the ones selected with \( k = 200 \).

Given the positive effect of clustering, we applied the same strategy to perform discriminant analysis on four methylation subtypes (inv.16, t.8.21, t.15.17 and double mutants). Since there are 4 subtypes, 3 discriminant vectors can be considered, obtained sequentially as described in Section 3.2. To summarize, the following steps are performed:
1. Choose the first two subtypes.

2. Cluster the features based on the differences between the two subtypes.

3. Perform discriminant analysis on clusters using only two selected subtypes.

4. Determine which original features correspond to selected clusters and eliminate them from the feature set.

5. Merge two subtypes into one and add one subtype from the remaining ones.

6. Repeat from Step 2 until the final subtype is added.

The resulting scores for the ERASMUS data are displayed in Figure 7. The resulting scores for the ECOG data plotted on the same coordinates are in Figure 8. The scores are based on 5, 5 and 7 clusters correspondingly (72, 61 and 104 original features). Symbols ”1” and ”2” correspond to subtypes known in both datasets (t.8.21 and inv.16). The figure indicates strong agreement between the ERASMUS and the ECOG datasets in terms of subtype classification.

DNA methylation data sets from the ERASMUS and ECOG studies are available in GEO, http://www.ncbi.nlm.nih.gov/geo/, with access number GSE18700 for ERASMUS. The number for ECOG is pending.

6 Discussion

In this work we consider an extension of Fisher’s discriminant analysis to the case where $p >> n$. The core of the method is optimization problem (4) which poses several challenges. First, the problem is non-convex meaning that convergence to the global maximum is not guaranteed. Secondly, the matrices $\Sigma_w$ and $\Sigma_b$ have to be reliably estimated. Our use of a non-diagonal shrinkage estimator for $\Sigma_w$ leads to improved misclassification rates over the algorithm proposed by Witten and Tibshirani (2011) at the expense of some reduction in computational speed. We also identify problems with existing algorithms in selecting a sparse subset of features for classification and propose a solution involving pre-clustering of features. Software for applying the method described in this paper is under development and will be soon available on CRAN as sparseFLDA.

The standard FLDA framework involves finding the dominant eigenvector of $W^{-1}B$. In this paper we use a shrinkage estimator $\tilde{W}$ in place of $W$, but an alternative when $p >> n$ is to estimate the product $\Sigma^{-1}_w \Sigma_b$ directly. In the case of two groups, this approach is considered by Cai (2011) who proposes to directly estimate $\Sigma^{-1}_w(\mu_1 - \mu_2)$. In future research we are planning to extend this idea to the general, multiple group, setting.

Acknowledgement

We thank Michael Todd for a useful discussion of duality theory. This research was partially supported by NSF grant DMS-1208488.
Appendix

Proof of Proposition 1

Proof. Follows from the fact that the eigenvector $l$ of matrix $B$ has sorted features and if $|l_i| \geq |l_j|$ then $|v_i| \geq |v_j|$.

For any vector $v \in \mathbb{R}^p$, $Bv = \gamma l^Tv$, so $(Bv)_i = \gamma (l^Tv)_i$ and $\text{sign}((Bv)_i) = \text{sign}((l^Tv)_i)$. Therefore $|(Bv)_i| \geq |(Bv)_j|$ iff $|l_i| \geq |l_j|$. Since the algorithm update for solving (11) has the form

$$v_{j+1}^m = \frac{\text{sign}((Bv^m)_j) \max(0, |(Bv^m)_j| - \lambda/2)}{\delta},$$

it follows that if $|v_{k+1}^m| > 0$ then $|v_{i+1}^m| > 0$ for all $i < k$ since $|l_1| \geq |l_2| \geq \ldots \geq |l_p| > 0$. 

Figure 7: Discriminant scores for 4 subtypes projected onto 2 dimensions, ERASMUS.

Figure 8: Discriminant scores for 4 subtypes projected onto 2 dimensions, ERASMUS+ECOG.
Proof of Proposition 2

Proof. 1. Suppose $F_k \leq F_k$. By definition $F_k = \max_{v \in A_k \cup A_0} \langle v, v' \rangle \leq 1$ for all $v' \in A_k \cup A_0$ such that $v^Tv' \leq 1$. Take $v' = \frac{v}{\|v\|_2}$. Therefore it follows

$$f(v') = v'^T \gamma l^T v' - \lambda \sum_{i=1}^p |v_i'| = \gamma (l^T v')^2 - \lambda \sum_{i=1}^k |v_i'| = F_k.$$  

2. Suppose $F_k \leq F_k$. For all $v \in A_k \cup A_0$ such that $v^Tv \leq 1$ we have

$$f(v) = \gamma (l^T v)^2 - \lambda \sum_{i=1}^k |v_i| = \gamma \sum_{i=1}^k |l_i||v_i| \left( \sum_{i=1}^k |l_i||v_i| - \frac{\lambda}{\gamma |l_i|} \right) \leq$$

$$\leq \gamma \sum_{i=1}^k |l_i||v_i| \left( ||l^k||_2 - \frac{\lambda}{\gamma |l_i|} \right) \leq \max \left( 0, \gamma ||l^k||_2 \left( ||l^k||_2 - \frac{\lambda}{\gamma |l_i|} \right) \right) = F_k.$$

Proof of Proposition 3

Proof. If $F_k \leq 0$ then $v_0 \notin A_k$ since $F_0 = 0 \geq F_k$ and $F = \max(F_0, F_1, ..., F_p)$. From here and Proposition 2 it follows that a necessary condition for $v_0 \in A_k$ is that $F_k > 0$. Indeed, if $F_k = 0$ then $F_k \leq F_k = 0$.

$F_k > 0$ by definition is equivalent to

$$\gamma ||l^k||_2^2 - \lambda \frac{||l^k||_2}{|l_1|} > 0.$$  

This can be rewritten as $||l^k||_2(\gamma ||l^k||_2 - \frac{\lambda}{|l_1|}) > 0$, which is equivalent to $||l^k||_2 > \frac{\lambda}{\gamma |l_1|}$.

Since $||l^k||_2 \leq ||l^m||_2$ for all $m \geq k$, this means that if $F_k > 0$ then $F_m > 0$ for all $m \geq k$. On the other hand, if $F_m = 0$ then $||l^m||_2 \leq \frac{\lambda}{|l_1|}$, which means that $F_k = 0$ for all $k \leq m$. From here it follows that if $F_m = 0$ then $v_0 \notin \bigcup_{i=1}^m A_i$ which is equivalent to $v_0 \in A_0 \cup \bigcup_{i=m+1}^p A_i$.

Proof of Proposition 4

Proof. Note that a sufficient condition for $F_r > F_k$ for $r > k$ is $F_r > F_k$. From Proposition 2 this is equivalent to

$$\gamma ||l^r||_2^2 - \lambda \frac{||l^r||_2}{||l^r||_1} > \max \left( 0, \gamma ||l^k||_2^2 - \lambda \frac{||l^k||_2}{|l_1|} \right).$$  

Now since $F_k = 0$ then $v_0 \notin \bigcup_{i=1}^k A_i$, we only need to consider the case $F_k > 0$. So we can consider

$$\gamma ||l^r||_2^2 - \lambda \frac{||l^r||_2}{||l^r||_1} > \gamma ||l^k||_2^2 - \lambda \frac{||l^k||_2}{|l_1|}$$  

which is equivalent to

$$\frac{\lambda}{\gamma} < \frac{||l^r||_2^2 - ||l^k||_2^2}{||l^r||_2^2 - ||l^k||_2^2}.$$
We know that $F_k > 0$ iff $\frac{x}{\gamma} < |l_1||l^k||_2$. It follows that $F^r > F_k$ when $F_k > 0$ is equivalent to

$$|l_1||l^k||_2 \leq \frac{|l^r||^2}{|l^r||^2} - \frac{|l^k||^2}{|l^k||^2}.$$

This can be rewritten as

$$|l^k||_2 \leq \frac{|l^r||^{3/2}}{|l_1||l^r||^1}.$$

Since $|l^m||_2 \geq |l^k||_2$ for all $m \geq k$, the result follows. \qed

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis, 2nd ed.* New York: John Wiley & Sons.

Bai, Z. and Y. Yin (1973). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Probab. 21*, 1275–1294.

Becker, N., W. Werft, and A. Benner (2010). *penalizedSVM: Feature Selection SVM using penalty functions.* R package version 1.1.

Bertsekas, D. P. (1999). *Nonlinear Programming.* Belmont: Athena Scientific.

Bickel, P. J. and E. Levina (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.

Bien, J. and R. Tibshirani (2011). Sparse estimation of a covariance matrix. *Biometrika 98*(4), 807–820.

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

Bradley, P. S. and O. L. Mangasarian (1998). Feature selection via concave minimization and support vector machines. In *Proceedings of the 15th International Conference on Machine Learning - ICML’98*, Madison, pp. 82–90.

Cai, T. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Am. Statist. Ass. 106*(496), 1566–1577.

Cochran, W. (1964). On the Performance of the Linear Discriminant Function. *Technometrics 6*(2), 179–190.

Dudoit, S., J. Fridlyand, and T. P. Speed (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Statist. Ass. 97*(457), 77–87.

Fan, J. and Y. Fan (2008). High dimensional classification using features annealed independence rules. *Ann. Statist. 36*(6), 2605.
Fan, J., Y. Feng, and X. Tong (2010). A road to classification in high dimensional space. *arXiv preprint arXiv:1011.6095*.

Figueroa, M. E., S. Lugthart, Y. Li, C. Erpelinck-Verschueren, X. Deng, P. J. Christos, E. Schifano, J. Booth, W. van Putten, L. Skrabanek, F. Campagne, M. Mazumdar, J. M. Greally, P. J. M. Valk, B. Loewenberg, R. Delwel, and A. Melnick (2010). DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell 17*(1), 13–27.

Friedman, J. H. (1989). Regularized discriminant analysis. *J. Am. Statist. Ass. 84*(405), 165–175.

Gorski, J., F. Pfeuffer, and K. Klamroth (2007). Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research 66*(3), 373–407.

Guo, Y., T. Hastie, and R. Tibshirani (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics 8*(1), 86–100.

Hartigan, J. A. and M. A. Wong (1979). Algorithm AS 136: A k-means clustering algorithm. *J. R. Statist. Soc. C 28*, 100–108.

Huang, S., T. Tong, and H. Zhao (2010). Bias-corrected diagonal discriminant rules for high-dimensional classification. *Biometrics 66*(4), 1096–1106.

Kormaksson, M., J. Booth, M. E. Figueroa, and A. Melnick (2012). Integrative model-based clustering of microarray methylation and expression data. *Appl. Statist. 6*(3), 1327–1347.

Kubokawa, T. and M. S. Srivastava (2008). Estimation of the precision matrix of a singular wishart distribution and its application in high-dimensional data. *Journal of Multivariate Analysis 99*(9), 1906–1928.

Ledoit, O. and M. Wolf (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance 10*(5), 603–621.

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88*(2), 365–411.

Mai, Q., H. Zou, and M. Yuan (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika 99*(1), 29–42.

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

Pang, H., T. Tong, and H. Zhao (2009). Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. *Biometrics 65*(4), 1021–1029.

Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika 99*(3), 733–740.

Schäfer, J. and K. Strimmer (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology 4*(1), 32.
Shalev-Shwartz, S. and A. Tewari (2009). Stochastic methods for l1 regularized loss minimization. In *Proceedings of the 26th International Conference on Machine Learning - ICML'09*, New York, pp. 1–8. ACM Press.

Shao, J., Y. Wang, X. Deng, and S. Wang (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist. 39*(2), 1241–1265.

Stein, C. (1975). Estimation of a covariance matrix. Rietz Lecture.

Thompson, R. F., M. Reimers, B. Khulan, M. Gissot, T. A. Richmond, Q. Chen, X. Zheng, K. Kim, and J. M. Greally (2008). An analytical pipeline for genomic representations used for cytosine methylation studies. *Bioinformatics 24*(9), 1161–1167.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J.R. Statist. Soc. B 58*(1), 267–288.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science 18*(1), 104–117.

Tseng, P. (1988). Coordinate Ascent for Maximizing Nondifferentiable Concave Functions. Massachusetts Institute of Technology, Laboratory for Information and Decision Systems.

Witten, D. (2011). *penalizedLDA: Penalized classification using Fisher's linear discriminant*. R package version 1.0.

Witten, D. M. and R. Tibshirani (2011). Penalized classification using Fisher’s linear discriminant. *J. R. Statist. Soc. B 73*(5), 753–772.

Xiao, G. (2011). Modeling three-dimensional chromosome structures using gene expression data. *J. Am. Statist. Ass. 106*(493), 61–72.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics 15*(2), 265–286.