Alternating direction method of multipliers for sparse zero-variance discriminant analysis and principal component analysis

Brendan P.W. Ames [*] Mingyi Hong [†]

January 21, 2014

Abstract

We consider the task of classification in the high-dimensional setting where the number of features of the given data is significantly greater than the number of observations. To accomplish this task, we propose sparse zero-variance discriminant analysis (SZVD) as a method for simultaneously performing linear discriminant analysis and feature selection on high-dimensional data. This method combines classical zero-variance discriminant analysis, where discriminant vectors are identified in the null space of the sample within-class covariance matrix, with penalization applied to the discriminant vectors to induce sparse solutions. We propose a simple algorithm based on the alternating direction method of multipliers for approximately solving the resulting nonconvex optimization problem. Further, we show that this algorithm is applicable to a larger class of penalized generalized eigenvalue problems, including a particular relaxation of the sparse principal component analysis problem. Theoretical guarantees for convergence of our algorithm to stationary points of the original nonconvex problem and the results of numerical experiments evaluating performance of our classification heuristic on simulated data and data drawn from applications in time-series classification are also provided.

[*]Department of Computing + Mathematical Sciences, California Institute of Technology, 1200 E. California Blvd., Mail Code 305-16, Pasadena, CA, 91125, bpames@caltech.edu

[†]Department of Electrical and Computer Engineering, University of Minnesota, 117 Pleasant Street SE, Walter Library RM 455, Minneapolis, MN, 55406, mhong@umn.edu

1 Introduction

In this paper, we consider penalized generalized eigenproblems of the form

$$\max_{\mathbf{x}\in\mathbf{R}^p} \left\{ -\tfrac{1}{2}\mathbf{x}^T B\mathbf{x} + \gamma\|D\mathbf{x}\|_1 : W\mathbf{x} = 0,\ \mathbf{x}^T\mathbf{x} \leq 1 \right\}, \tag{1.1}$$

where B and W are $p \times p$ positive semidefinite matrices, D is a $p \times p$ orthogonal matrix, $\|\cdot\|_1$ is the ℓ_1-norm on $\mathbf{R}^p$ defined by $\|\mathbf{y}\|_1 = |y_1| + |y_2| + \cdots + |y_p|$, and $\gamma > 0$ is a fixed regularization parameter. That is, we seek some sparse unit vector $\mathbf{x} \in \mathbf{R}^p$, with respect to the orthogonal basis defined by the columns of D, maximizing the quadratic form $\mathbf{x}^T B\mathbf{x}$ over the null space of W; here the ℓ_1-norm acts as a convex surrogate of vector cardinality. Such generalized eigenproblems commonly arise when performing dimensionality reduction of high-dimensional data, particularly in linear discriminant analysis (LDA) and its unsupervised analogue, principal component analysis (PCA). When unpenalized, LDA and PCA seek projections of high-dimensional data, i.e., p is much greater than the number of observations and the ranks of B and W, to a significantly lower dimensional space such that variance is maximized within the projected space, typically for the purpose of some statistical task such as classification or model fitting. If the data set is well-approximated by its low-dimensional projection, then significant improvement in computational complexity of the prediction task, as well as quality and interpretability of the predictions can be obtained. The use of an ℓ_1-norm regularization term (or other sparsity inducing penalty) encourages sparse loading vectors for computing the low-dimensional representation, allowing further improvement in computational efficiency and interpretability, although often at a significant increase in cost of computing the loading vectors. This approach in itself is not novel; ℓ_1-regularization and similar techniques have long been used in the statistics, machine learning, and signal processing communities to induce sparse solutions, most notably in the LASSO [37] and compressed sensing [8, 30] regimes. A brief review of ℓ_1-regularization for high-dimensional LDA and PCA can be found in Sect. 2.

The primary contributions of this paper are twofold. First, we propose a new heuristic for penalized classification, based on recasting ℓ_1-penalized zero-variance discriminant analysis as a special case of penalized eigenproblems of the form (1.1); we provide a brief overview of zero-variance discriminant analysis in Sect. 2.2. Second, we propose a new algorithm for obtaining approximate solutions of penalized eigenproblems of the form (1.1), based on the alternating direction method of multipliers. Our algorithm essentially finds an approximate solution of (1.1) by alternately maximizing each term of the objective function until convergence. Although the problem (1.1) is nonconcave in general, we will see that it is easy to maximize each term of the objective, either $\mathbf{x}^T B\mathbf{x}$ or $-\|D\mathbf{x}\|_1$, with the other fixed. We develop

this algorithm in Sect. 3, as motivated by its use as a heuristic for penalized zero-variance discriminant analysis. Further, we show that this algorithm converges to a stationary point of (1.1) under certain assumptions on the matrices W and D in Sect. 3.3 and empirically test performance of the resulting classification heuristic in Sect. 4.

2 Linear Discriminant Analysis

2.1 Fisher Linear Discriminant Analysis

Given a data set with each observation labeled as belonging to one of several classes, *Fisher Linear Discriminant Analysis* (LDA) [18], [39], [23, Chapter 4] seeks a low-dimensional representation of the data where the projected class-means are well separated, relative to the projections of the individual classes. Suppose that the data are given as the rows of the matrix $X \in \mathbf{R}^{n\times p}$; here, each row $\mathbf{x}_i \in \mathbf{R}^p$ of X represents a single observation of a vector containing p features and the data set contains n such observations. Each observation is known to belong to exactly one of K classes, denoted $C_1, C_2, \ldots, C_K$. We assume that the data has been centered and normalized so that each feature has mean equal to 0 and variance equal to 1. Considered as a separate data set, the mean and covariance of each class C_i may be approximated by the sample class-mean $\boldsymbol{\mu}_i$ and covariance matrix Σ_i given by

$$\boldsymbol{\mu}_i = \sum_{j\in C_i} \frac{\mathbf{x}_j}{|C_i|}, \qquad \Sigma_i = \frac{1}{n}\sum_{j\in C_i} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T,$$

respectively. On the other hand, we may approximate the *within-class covariance matrix* W, measuring variability within classes, by the sum of the sample class-covariance matrices

$$W = \frac{1}{n}\sum_{i=1}^{K}\sum_{j\in C_i} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T.$$

On the other hand, we define the sample *between-class covariance matrix*, measuring variability between the class means, as

$$B = \frac{1}{n}\sum_{i=1}^{K} \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T. \tag{2.1}$$

It is important to note that the matrix B has rank bounded above by $K-1$. Indeed, the column space of B is spanned by the K linearly dependent vectors $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K\}$ and, as such, B has rank at most $K-1$.

As mentioned earlier, we would like to identify a projection of the rows of X to a lower dimensional space such that the projected class means are well separated, while observations within the same class are relatively close in the projected space. To do so, LDA yields a set of nontrivial loading vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{K-1}$, obtained by repeatedly maximizing the criterion

$$J(\mathbf{w}) = \frac{\mathbf{w}^T B \mathbf{w}}{\mathbf{w}^T W \mathbf{w}}. \tag{2.2}$$

By the fact that B has rank at most $K-1$, there must exist at most $K-1$ orthogonal directions $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{K-1}$ such that the quadratic form $\mathbf{w}_i^T B \mathbf{w}_i$ has nonzero value.

To motivate the use of this criterion, we consider the case when $K = 2$. When $K = 2$, the sample between-class covariance matrix is often written as $B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$; a straight-forward calculation shows that such a transformation results in a constant scaling of the value of the numerator of (2.2) from that given by (2.1) and, therefore, does not alter the set of maximizers of (2.2). For each $\mathbf{w} \in \mathbf{R}^p$, we have

$$\mathbf{w}^T B \mathbf{w} = \mathbf{w}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{w} = (\mathbf{w}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2 = |\mathbf{w}^T \boldsymbol{\mu}_1 - \mathbf{w}^T \boldsymbol{\mu}_2|^2,$$

and so the numerator of (2.2) is exactly the squared distance separating the projected class means. On the other hand, the total within-class scatter in the one-dimensional space spanned by $\mathbf{w}$ is equal to

$$\mathbf{w}^T W \mathbf{w} = \sum_{i=1}^{K} \sum_{j \in C_i} \mathbf{w}^T (\mathbf{x}_j - \boldsymbol{\mu}_i)^T (\mathbf{x}_j - \boldsymbol{\mu}_i)^T \mathbf{w} = \sum_{i=1}^{K} \sum_{j \in C_i} |\mathbf{w}^T \mathbf{x}_j - \mathbf{w}^T \boldsymbol{\mu}_i|^2.$$

Therefore, (2.2) is exactly maximizing the ratio of the between-class scatter to the within-class scatter in the projected space when $K = 2$. Extending this rationale to all choices of K yields the LDA criterion (2.2).

To perform dimensionality reduction using LDA, we identify the desired loading vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{K-1}$ by sequentially solving the optimization problem

$$\mathbf{w}_i = \arg\max_{\mathbf{w} \in \mathbf{R}^p} \left\{ \frac{\mathbf{w}^T B \mathbf{w}}{\mathbf{w}^T W \mathbf{w}} : \mathbf{w}^T W \mathbf{w}_j = 0 \; \forall \, j = 1, \ldots, i-1 \right\} \tag{2.3}$$

for all $i = 1, 2, 3, \ldots, K-1$. That is, $\mathbf{w}_i$ is the vector W-conjugate to the span of $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{i-1}\}$ maximizing the LDA criterion (2.2). Noting that the criterion $J(\mathbf{w})$ is invariant to scaling, we may assume that $\mathbf{w}^T W \mathbf{w} \leq 1$ and rewrite (2.3) as

$$\mathbf{w}_i = \arg\max_{\mathbf{w} \in \mathbf{R}^p} \left\{ \mathbf{w}^T B \mathbf{w} : \mathbf{w}^T W \mathbf{w} \leq 1, \; \mathbf{w}^T W \mathbf{w}_j = 0 \; \forall \, j = 1, \ldots, i-1 \right\}. \tag{2.4}$$

Thus, finding the $K-1$ discriminant vectors is equivalent to solving the generalized eigenproblem (2.4). When the sample within-class covariance matrix W is nonsingular, we may solve (2.4) by performing the change of variables $\mathbf{z} = W^{1/2}\mathbf{w}$. After this change of variables, we have $\mathbf{w}_i = W^{-1/2}\mathbf{z}_i$, where

$$\mathbf{z}_i = \arg\max_{\mathbf{z}\in\mathbf{R}^p} \left\{ \mathbf{z}^T W^{-1/2} B W^{-1/2} \mathbf{z} : \mathbf{z}^T\mathbf{z} = 1,\, \mathbf{z}^T\mathbf{z}_j = 0,\, j = 1, \ldots, i-1 \right\}.$$

That is, we may find the desired set of discriminant vectors by finding the set of nontrivial unit eigenvectors of $W^{-1/2}BW^{-1/2}$ and multiplying each eigenvector by $W^{1/2}$.

2.2 High-Dimensional Linear Discriminant Analysis

Two significant obstacles arise when performing LDA in a high-dimensional setting, i.e., when $p > n$. First, the sample within-class covariance matrix W is singular. Indeed, W has rank at most n, as it is a linear combination of n rank-one matrices. If there exists some vector $\mathbf{w}$ in the null space of W not belonging to the null space of B, then the objectives of (2.3) and (2.4) can be made arbitrarily large. Therefore, both (2.3) and (2.4) are unbounded if $p > n$ in general. Second, when p is very large, the discriminant vectors typically contain p nonzero entries with no discernible structure and, thus, are often difficult to interpret.

Several solutions for the singularity problem have been proposed in the literature. One such proposed solution is to replace W in (2.4) with a positive definite approximation $\tilde{W}$, e.g., the diagonal estimate $\tilde{W} = \mathrm{Diag}(\mathrm{diag}(W))$; see [19, 29, 15, 3, 43]. Here, $\mathrm{Diag}(\mathbf{x})$ denotes the $p \times p$ diagonal matrix with diagonal entries given by $\mathbf{x} \in \mathbf{R}^p$ and $\mathrm{diag}(X)$ denotes the vector in $\mathbf{R}^p$ with entries equal to those on the diagonal of $X \in \mathbf{R}^{p\times p}$. After replacing the population covariance with a positive definite approximation $\tilde{W}$, we obtain a set of discriminant vectors maximizing the modified LDA criterion by sequentially solving the generalized eigenproblems

$$\mathbf{w}_i = \arg\max_{\mathbf{w}\in\mathbf{R}^p} \left\{ \mathbf{w}^T B\mathbf{w} : \mathbf{w}^T\tilde{W}\mathbf{w} \le 1,\ \mathbf{w}^T\tilde{W}\mathbf{w}_j = 0\ \forall\, j = 1, \ldots, i-1 \right\} \tag{2.5}$$

as before. Unfortunately, it may often be difficult to obtain a good positive definite approximation of the population within-class scatter matrix. For example, the diagonal approximation ignores any correlation between features, while approximations based on perturbation of W may require some training to obtain a suitable choice of $\tilde{W}$.

On the other hand, *zero-variance discriminant analysis* (ZVD), as proposed by Krzanowski et al. [29], embraces the singularity of W by seeking a set of discriminant vectors belonging to the null space of W. If $\mathrm{Null}(W) \not\subseteq \mathrm{Null}(B)$, we may obtain nontrivial discriminant vectors

by solving the generalized eigenproblem

$$\max \left\{ \mathbf{w}^T B \mathbf{w} : W\mathbf{w} = 0,\ \mathbf{w}^T \mathbf{w} = 1 \right\}. \tag{2.6}$$

That is, in ZVD we seek orthogonal directions $\mathbf{w}$ belonging to Null(W) maximizing between-class scatter; because we are restricting our search to Null(W), we seek orthogonal directions, not W-conjugate directions as before. If the columns of $N \in \mathbf{R}^{p \times d}$ form an orthonormal basis for Null(W), then ZVD is equivalent to the eigenproblem

$$\max_{\mathbf{x} \in \mathbf{R}^d} \left\{ \mathbf{x}^T (N^T B N) \mathbf{x} : \mathbf{x}^T \mathbf{x} = 1 \right\}, \tag{2.7}$$

where d denotes the dimension of Null(W). Clearly, the dimension of Null(W)$\setminus$Null(B) may be less than $K-1$. In this case, $N^T B N$ has less than $K-1$ nontrivial eigenvectors; a full set of $K-1$ discriminant vectors can be obtained by searching for the remaining discriminant vectors in the complement of Null(W) (see [16, pp. 8-9]). Alternately, reduced rank LDA could be performed using only the nontrivial discriminant vectors found in Null(W).

2.3 Penalized Linear Discriminant Analysis

While ZVD and the use of a positive definite approximation of the within-class covariance matrix each solves the singularity problem (to varying degrees), neither method addresses the interpretability problem. Indeed, these methods all reduce to generalized eigenvalue problems and there is no reason to expect the solutions of these eigenproblems to contain any meaningful structure. Ideally, one would like to simultaneously perform feature selection by obtaining a set of discriminant vectors containing relatively few nonzero entries (or some other special structure). In this case, one would be able to identify which features are truly important in the dimensionality reduction, while significantly improving computational efficiency through the use of sparse loading vectors.

The problem of identifying sparse solutions to eigenproblems has received significant attention, primarily in relation to sparse principal component analysis. In *principal component analysis* (PCA) [23, Section 14.5], one seeks a dimensionality reduction maximizing variance in the lower dimensional space. Specifically, the first k principal components are the k orthogonal directions $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$ maximizing $\mathbf{w}^T \Sigma \mathbf{w}$, where $\Sigma \in \mathbf{S}_+^p$ is an approximation of the population covariance matrix (typically the sample covariance matrix); here $\mathbf{S}_+^p$ denotes the cone of $p \times p$ positive semidefinite matrices. Thus, principal component analysis reduces to identifying the k leading eigenvectors of the approximation of the covariance matrix Σ. It is known that the sample covariance is a consistent estimator of the population covariance,

i.e., the sample covariance matrix converges to the true population covariance matrix with probability 1 as the sample size n tends to infinity for fixed number of features p. However, when p is larger than n, as it is in the high-dimensional setting, the sample covariance matrix may be a poor approximation of the population covariance; see [25, 2, 33]. One approach to addressing this issue, as well as to increase interpretability of the obtained loading vectors, is to add regularization in the form of the restriction that the principal component vectors be sparse. Many different methods for this task have been proposed, typically involving ℓ_0 or ℓ_1-regularization, convex relaxation, thresholding, or combination of all three; see [26, 47, 14, 12, 42, 27, 31, 44, 13, 1, 34, 45, 32] and the references within.

Fewer approaches have been proposed for performing regularized LDA in the high-dimensional setting. In [41], Witten and Tibshirani propose a penalized version of LDA where the kth discriminant vector is the solution of the optimization problem

$$\mathbf{w}_k = \max_{\mathbf{w}\in\mathbf{R}^p} \left\{ \mathbf{w}^T B \mathbf{w} - \rho(\mathbf{w}) : \mathbf{w}^T \tilde{W} \mathbf{w} \leq 1,\ \mathbf{w}^T \tilde{W} \mathbf{w}_i = 0 \ \forall\, i \leq k-1 \right\}, \tag{2.8}$$

where $\rho : \mathbf{R}^p \to \mathbf{R}_+$ is either an ℓ_1-norm or fused LASSO penalty function, and $\tilde{W}$ is the diagonal estimate of the within-class covariance $\tilde{W} = \mathrm{Diag}(\mathrm{diag}(W))$. The optimization problem (2.8) is nonconvex, because B is positive semidefinite, and cannot be solved as a generalized eigenproblem due to the presence of the regularization term $\rho(\mathbf{w})$. Consequently, it is unclear if it is possible to solve (2.8) efficiently. Witten and Tibshirani propose a minorization algorithm for approximately solving (2.8); the use of the diagonal estimate $\tilde{W}$ is partially motivated by its facilitation of the use of soft thresholding when solving the subproblems arising in this minorization scheme when using the ℓ_1-penalty.

Clemmensen et al. [11] consider an iterative method for penalized regression to obtain sparse discriminant vectors. Specifically, Clemmensen et al. apply an elastic net penalty [46] to the optimal scoring formulation of the LDA classification rule discussed in [22] as follows. Suppose that the first $k-1$ discriminant vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{k-1}$ and scoring vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{k-1}$ have been computed. Then the kth discriminant vector $\mathbf{w}_k$ and scoring vector $\boldsymbol{\theta}_k$ are the optimal solution pair of the problem

$$\begin{array}{ll} \min\limits_{\mathbf{w},\boldsymbol{\theta}} & \|Y\boldsymbol{\theta} - X\mathbf{w}\|^2 + \lambda_1 \mathbf{w}^T \Omega \mathbf{w} + \lambda_2 \|\mathbf{w}\|_1 \\ \text{s.t.} & \boldsymbol{\theta}^T Y^T Y \boldsymbol{\theta} = n,\ \boldsymbol{\theta}^T Y^T Y \boldsymbol{\theta}_\ell = 0 \, \forall\, \ell < k. \end{array} \tag{2.9}$$

Here Y is the $n \times K$ partition matrix of the data set X, i.e., Y_{ij} is the binary indicator variable for membership of the ith observation in the jth class, Ω is a positive definite matrix chosen to ensure that $W + \Omega$ is positive definite and to encourage smoothness of the obtained discriminant vectors, and λ_1 and λ_2 are nonnegative tuning parameters controlling

the ridge regression and ℓ_1-penalties, respectively. Clemmensen et al. propose the following iterative alternating direction method for solving (2.9). Suppose that we have the approximate solutions $\tilde{\mathbf{w}}_i$ and $\tilde{\boldsymbol{\theta}}_i$ of (2.9) at the ith step. These approximations are updated by first solving (2.9) for $\tilde{\mathbf{w}}_{i+1}$ with $\boldsymbol{\theta}$ fixed (and equal to $\tilde{\boldsymbol{\theta}}_i$); $\tilde{\boldsymbol{\theta}}_{i+1}$ is then updated by solving (2.9) with $\mathbf{w}$ fixed (and equal to $\tilde{\mathbf{w}}_{i+1}$). This process is repeated until the sequence of approximate solutions has converged or a maximum number of iterations has been performed. It is unclear if this algorithm is converging to a local minimizer because the criterion (2.9) is nonconvex; however, it can be shown that the solution of (2.8) is a stationary point of (2.9) under mild assumptions (see [41, Sect. 7.1]). In addition to these penalized heuristics, several thresholding methods [38, 21, 35] for sparse LDA have also been proposed; a summary and numerical comparison of several of these cited methods can be found in [9].

3 Penalized Zero-Variance Fisher Linear Discriminant Analysis

In this section, we propose a penalized version of the zero-variance discriminant analysis (ZVD) of Krazanowski et al. [29]. As in [41] and [11], we add ℓ_1-penalization to induce sparse loading vectors; here, to the generalized eigenproblem solved in zero-variance discriminant analysis. Specifically, we solve the problem

$$\max_{\mathbf{w}\in\mathbf{R}^p} \left\{ \tfrac{1}{2}\mathbf{w}^T B\mathbf{w} - \gamma \textstyle\sum_{i=1}^p \sigma_i \left| \left(D\mathbf{w}\right)_i \right| : W\mathbf{w} = 0, \mathbf{w}^T\mathbf{w} \leq 1 \right\} \tag{3.1}$$

to obtain the first discriminant vector; if the discriminant vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{k-1}$ have been identified, $\mathbf{w}_k$ can be found by appending $\{\mathbf{w}_1^T, \ldots, \mathbf{w}_{k-1}^T\}$ to the rows of W and solving (3.1). Here, $D \in \mathbf{O}^p$ is an orthogonal matrix, and the ℓ_1-penalty acts as a surrogate for the cardinality of $\mathbf{w}$ with respect to the basis given by the columns of D. The parameter $\boldsymbol{\sigma} \in \mathbf{R}^p$ is a scaling vector used to control emphasis of penalization; for example, the scaling vector σ may be taken to be the within-class standard deviations of the features $\boldsymbol{\sigma} = \sqrt{\operatorname{diag} W}$ to ensure that a greater penalty is imposed on features that vary the most within each class. As before, letting the columns of $N \in \mathbf{R}^{p\times d}$ form a basis for $\operatorname{Null}(W)$ yields the equivalent formulation

$$\max_{\mathbf{x}\in\mathbf{R}^d} \left\{ \tfrac{1}{2}\mathbf{x}^T N^T B N\mathbf{x} - \gamma \textstyle\sum_{i=1}^p \sigma_i \left| \left(DN\mathbf{x}\right)_i \right| : \mathbf{x}^T\mathbf{x} \leq 1 \right\}. \tag{3.2}$$

As in (2.8), (3.2) is the maximization of the sum of a convex function and a concave function over the unit ball; it is unknown if an efficient algorithm for solving (3.2) exists, although maximizing nonconcave functions is NP-hard in general. We next develop a heuristic based on the *alternating direction method of multipliers* to approximately solve (3.2) and use the

obtained approximate solutions as our set of discriminant vectors.

3.1 Alternating Direction Method of Multipliers

Given problems of the form $\min_{\mathbf{x},\mathbf{y}} \{f(\mathbf{x}) + g(\mathbf{y}) : A\mathbf{x} + B\mathbf{y} = \mathbf{c}\}$ the *alternating direction method of multipliers* (ADMM) attempts to the solve the problem by iteratively minimizing the augmented Lagrangian of the problem with respect to each primal decision variable, and then updating the dual variable using dual ascent; a recent survey on ADMM and related methods can be found in [4]. To transform (3.2) to a form appropriate for ADMM, we define an additional decision variable $\mathbf{y} \in \mathbf{R}^p$ such that $DN\mathbf{x} = \mathbf{y}$. After this splitting of variables and replacing the maximization with an appropriate minimization, (3.2) is equivalent to

$$\min_{\mathbf{x}\in\mathbf{R}^d,\mathbf{y}\in\mathbf{R}^p} \left\{-\tfrac{1}{2}\mathbf{x}^T(N^TBN)\mathbf{x} + \gamma\textstyle\sum_{i=1}^p \sigma_i\,|y_i| : \mathbf{y}^T\mathbf{y} \le 1,\ DN\mathbf{x} = \mathbf{y}\right\}. \tag{3.3}$$

Letting $A = N^TBN$ and $\rho(\mathbf{y}) = \sum_{i=1}^p \sigma_i\,|y_i|$, we see that (3.3) is equivalent to

$$\min_{\mathbf{x}\in\mathbf{R}^d,\mathbf{y}\in\mathbf{R}^p} \left\{-\tfrac{1}{2}\mathbf{x}^TA\mathbf{x} + \gamma\rho(\mathbf{y}) : \mathbf{y}^T\mathbf{y} \le 1,\ DN\mathbf{x} - \mathbf{y} = 0\right\}. \tag{3.4}$$

This transformation has the additional benefit that $\rho(\mathbf{y})$ is separable in $\mathbf{y}$, while $\rho(DN\mathbf{x})$ is not separable in $\mathbf{x}$; this fact will play a significant role in the ADMM algorithm, as we will see shortly. We are now ready to apply ADMM to approximately solve (3.4).

The problem (3.4) has augmented Lagrangian

$$L_\beta(\mathbf{x},\mathbf{y},\mathbf{z}) = -\frac{1}{2}\mathbf{x}^TA\mathbf{x} + \gamma\rho(\mathbf{y}) + \delta_B(\mathbf{y}) + \mathbf{z}^T(DN\mathbf{x} - \mathbf{y}) + \frac{\beta}{2}\|DN\mathbf{x} - \mathbf{y}\|^2,$$

where β is a regularization parameter chosen to make L_β strictly convex in each of $\mathbf{x}$ and $\mathbf{y}$, and $\delta_{B_p} : \mathbf{R}^p \to \{0, +\infty\}$ is the indicator function defined by $\delta_{B_p}(\mathbf{y}) = 0$ if $\mathbf{y}^T\mathbf{y} \le 1$ and is equal to $+\infty$ otherwise; here B_p denotes the unit ball in $\mathbf{R}^p$ centered at the origin. Suppose that after k iterations we have the iterates $(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)$. We update $(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^{k+1})$ sequentially by the following steps:

$$\mathbf{y}^{k+1} = \arg\min_{\mathbf{y}\in\mathbf{R}^p} L_\beta\left(\mathbf{x}^k, \mathbf{y}, \mathbf{z}^k\right) \tag{3.5}$$

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathbf{R}^p} L_\beta\left(\mathbf{x}, \mathbf{y}^{k+1}, \mathbf{z}^k\right) \tag{3.6}$$

$$\mathbf{z}^{k+1} = \mathbf{z}^k + \beta\left(DN\mathbf{x}^{k+1} - \mathbf{y}^{k+1}\right). \tag{3.7}$$

That is, $\mathbf{y}$ and $\mathbf{x}$ are updated by minimizing the augmented Lagrangian with all other variables fixed in (3.5) and (3.6), respectively, and $\mathbf{z}$ is updated by taking a dual ascent step

in (3.7).

We now describe the solution of (3.5) and (3.6). It is easy to see that (3.5) is equivalent to

$$\mathbf{y}^{k+1} = \arg\min_{\mathbf{y}\in\mathbf{R}^p} \left\{ \gamma\rho(\mathbf{y}) + \frac{\beta}{2}\mathbf{y}^T\mathbf{y} - \mathbf{y}^T(\beta DN\mathbf{x}^k + \mathbf{z}^k) : \mathbf{y}^T\mathbf{y} \leq 1 \right\}. \tag{3.8}$$

Applying the Karush-Kuhn-Tucker conditions [5, Section 5.5.3] to (3.8), we see that $\mathbf{y}^{k+1}$ must satisfy

$$0 \in \gamma\partial\rho(\mathbf{y}^{k+1}) + (\beta+\lambda)\mathbf{y}^{k+1} - (\beta DN\mathbf{x}^k + \mathbf{z}^k), \qquad \lambda((\mathbf{y}^{k+1})^T\mathbf{y}^{k+1} - 1) = 0 \tag{3.9}$$

for some $\lambda \geq 0$. By the form of the subdifferential of ρ (see [6, Section 3.4]), each component of $\mathbf{y}^{k+1}$ must satisfy

$$0 = (\beta+\lambda)y_i^{k+1} + \gamma\phi_i - b_i,$$

where $\mathbf{b} = \beta DN\mathbf{x}^k + \mathbf{z}^k$, for some $\boldsymbol{\phi} \in \mathbf{R}^p$ satisfying $\boldsymbol{\phi}^T\mathbf{y}^{k+1} = \rho(\mathbf{y}^{k+1})$ and $|\phi_i| \leq \sigma_i$ for all $i \in \{1,\ldots,p\}$. Rearranging and solving for y_i^{k+1} shows that

$$(\beta+\lambda)y_i^{k+1} = \operatorname{sign}(b_i)\cdot\max\{|b_i| - \gamma\sigma_i, 0\}$$

for all $i \in \{1,\ldots,p\}$. Letting $\mathbf{s}^{k+1} \in \mathbf{R}^p$ be the vector such that $s_i^{k+1} = \operatorname{sign}(b_i)\cdot\max\{|b_i| - \gamma\sigma_i, 0\}$ for all $i \in \{1,\ldots,p\}$ and applying the complementary slackness condition $\lambda((\mathbf{y}^{k+1})^T\mathbf{y}^{k+1} - 1) = 0$ shows that

$$\mathbf{y}^{k+1} = \frac{\mathbf{s}^{k+1}}{\beta + \max\{0, \|\mathbf{s}^{k+1}\| - \beta\}}.$$

That is, we update $\mathbf{y}^{k+1}$ by applying soft thresholding to $\mathbf{b}$ with respect to $\gamma\boldsymbol{\sigma}$, and then normalizing the obtained solution if it has norm greater than 1.

On the other hand, (3.6) is equivalent to

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathbf{R}^d} \frac{1}{2}\mathbf{x}^T(\beta I - A)\mathbf{x} + \mathbf{x}^T(DN)^T(\mathbf{z}^k - \beta\mathbf{y}^{k+1}). \tag{3.10}$$

For sufficiently large choice of β, (3.10) is an unconstrained convex program. Taking the derivative of the objective of (3.10) shows that $\mathbf{x}^{k+1}$ is the solution of the linear system

$$(\beta I - A)\mathbf{x}^{k+1} = (DN)^T(\beta\mathbf{y}^{k+1} - \mathbf{z}^k), \tag{3.11}$$

by the fact that $\mathbf{x}^{k+1}$ must be a critical point of the objective of (3.10). Putting everything together we have the following algorithm for identifying the set of $K-1$ discriminant vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{K-1}$ corresponding to the given data set:

0. Given $K-1$ regularization parameters $\{\gamma_1, \gamma_2, \ldots, \gamma_k\}$, and sets of initial solutions $\{(\mathbf{x}^0, \mathbf{y}^0, \mathbf{z}^0)_i\}_{i=1}^{K-1}$. Set $i = 1$.

1. Compute basis N for the null space of W.

2. Approximately solve (3.4) with regularization parameter $\gamma = \gamma_i$ using the ADMM algorithm described by (3.5), (3.6), and (3.7) and the initial solution $(\mathbf{x}^0, \mathbf{y}^0, \mathbf{z}^0)_i$. to obtain $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*)$. Take $\mathbf{w}_i = DN\mathbf{x}^*$ as the ith zero variance discriminant vector.

3. Append $\mathbf{w}_i$ to W: $[W; \mathbf{w}_i^T] \mapsto W$. Update $i = i + 1$.

4. Repeat steps 1 through 3 until all nontrivial discriminant vectors $\{\mathbf{w}_i\}_{i=1}^{K-1}$ are computed.

The ADMM algorithm in Step 2 is stopped when the primal and dual residuals of the current iterate $(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)$, $DN\mathbf{x}^k - \mathbf{y}^k$ and $\beta(\mathbf{y}^k - \mathbf{y}^{k-1})$, respectively, are sufficiently small. That is, we declare the algorithm to have converged when

$$\begin{aligned} \|DN\mathbf{x}^k - \mathbf{y}^k\| &\leq tol_{abs} \cdot \sqrt{p} + tol_{rel} \cdot \max\{\|\mathbf{x}^k\|, \|\mathbf{y}^k\|\} \\ \beta\|\mathbf{y}^k - \mathbf{y}^{k-1}\| &\leq tol_{abs} \cdot \sqrt{p} + tol_{rel} \cdot \|\mathbf{y}^k\|, \end{aligned}$$

for desired absolute and relative error tolerances tol_{abs} and tol_{rel}; motivation for this choice of stopping tolerance is provided in [4, Sect. 3.3.1].

We conclude this section with a brief discussion of the per-iteration complexity of performing (3.5), (3.6), and (3.7). The soft thresholding and dual ascent operations in the update of $\mathbf{y}$ and $\mathbf{z}$, respectively, each can be performed in $O(p)$ flops provided that $(DN)\mathbf{x}^k$ has been computed; computing DN requires $O(p^2 d)$ flops, while performing the matrix-vector multiplication $(DN)\mathbf{x}^{k+1}$ requires $O(pd)$ flops per iteration. On the other hand, (3.6) requires the solution of the linear system (3.11). The coefficient matrix $\beta I - A$ is fixed for all iterations. Its Cholesky decomposition can be precomputed during initialization of the algorithm at a cost of $O(d^3)$ flops; afterward each $\mathbf{x}$ update comes at a cost of two triangular system solves and a single matrix-vector multiplication, costing $O(d^2)$ and $O(pd)$ operations, respectively. It should be noted that the cost of this update can be significantly improved by exploiting the structure of A in some special cases. For example, if $K = 2$, we have the decomposition $A = \mathbf{v}\mathbf{v}^T$, where $\mathbf{v} = N^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Applying the Sherman-Morrison-Woodbury formula [20, Equation (2.1.4)] to (3.11) shows that

$$\mathbf{x}^{k+1} = \frac{1}{\beta}\left((DN)^T(\beta\mathbf{y}^{k+1} - \mathbf{z}^k) - \left(\frac{(\beta\mathbf{y}^{k+1} - \mathbf{z}^k)^T DN\mathbf{v}}{\beta - \mathbf{v}^T\mathbf{v}}\right)\mathbf{v}\right),$$

which can be computed in $O(pd)$ flops (the cost of the matrix-vector multiplication $(DN)^T(\beta \mathbf{y}^{k+1} - \mathbf{z}^k)$ plus the linear cost of the inner products and other vector operations). In addition to the per-iteration costs, the algorithm requires an eigenvalue decomposition or QR decomposition of the within-class covariance matrix W to compute N (at a cost of $O(p^3)$ flops). Therefore, the algorithm has a total time complexity of $O((K-1)p^3) + O(\#\text{its} \cdot pd)$.

3.2 Connection to Sparse PCA and the General Problem

Recall that the leading principal component of a given data set can be identified by solving the optimization problem

$$\mathbf{w}_1 = \arg\max_{\mathbf{w} \in \mathbf{R}^p} \left\{ \mathbf{w}^T \Sigma \mathbf{w} : \mathbf{w}^T \mathbf{w} \leq 1 \right\} \tag{3.12}$$

where $\Sigma \in \mathbf{S}_+^p$ is the sample covariance matrix of the centered data. A frequently used approach to simultaneously perform feature selection and principal component analysis is to require the obtained principal component to be k-sparse, with respect to the orthonormal basis $D \in \mathbf{O}^p$, for some integer k:

$$\mathbf{w}_1 = \arg\max_{\mathbf{w} \in \mathbf{R}^p} \left\{ \mathbf{w}^T \Sigma \mathbf{w} : \mathbf{w}^T \mathbf{w} \leq 1,\ \|D\mathbf{w}\|_0 \leq k \right\}.$$

Moving the cardinality constraint to the objective as a penalty and relaxing the ℓ_0-norm with the ℓ_1-norm yields the relaxation

$$\mathbf{w}_1 = \arg\max_{\mathbf{w} \in \mathbf{R}^p} \left\{ \mathbf{w}^T \Sigma \mathbf{w} + \gamma \|D\mathbf{w}\|_1 : \mathbf{w}^T \mathbf{w} \leq 1 \right\}. \tag{3.13}$$

Clearly, (3.13) is a special case of (3.2) with $B = \Sigma$, $W = 0$ (or equivalently, $N = I$), and $\boldsymbol{\sigma} = \mathbf{e}$. Therefore, the algorithm outlined in the previous section is also applicable to this relaxation of the sparse PCA problem (3.13). More generally, this algorithm is immediately applicable to all problems of the form given by (1.1).

3.3 Convergence Analysis

It is known that ADMM converges to the optimal solution of

$$\min_{\mathbf{x} \in \mathbf{R}^{n_1}, \mathbf{y} \in \mathbf{R}^{n_2}} \{ f(\mathbf{x}) + g(\mathbf{y}) : A\mathbf{x} + B\mathbf{y} = \mathbf{c} \}$$

if both f and g are convex functions (see [17, Theorem 8], [4, Section 3.2], and [24]). However, general convergence results for minimizing nonconvex separable functions, such as the

objective of (3.4), are unknown. In this section, we establish that, under certain assumptions on the within-class covariance matrix W and the dictionary matrix D, the ADMM algorithm described by (3.5), (3.6), and (3.7) converges to a stationary point of (3.4). Let us define a new matrix $M = DN$. Clearly the columns of M are also orthogonal, as we have $M^T M = N^T D^T DN = I$. We have the following theorem.

Theorem 3.1 *Suppose that the columns of the matrix* $[M, C] \in \mathbf{O}^p$ *form an orthonormal basis for* $\mathbf{R}^p$. *Suppose further that the sequence of iterates* $(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)$ *generated by* (3.5), (3.6), *and* (3.7) *satisfies*

$$C^T(\mathbf{z}^{k+1} - \mathbf{z}^k) = 0 \tag{3.14}$$

for all k *and*

$$\beta > \|A\| \left(\frac{\lambda_0 + 2}{\lambda_0} \right), \tag{3.15}$$

where $\|A\|$ *denotes the square root of the largest singular value of* $A = N^T BN$, *and* λ_0 *denotes the smallest nonzero eigenvalue of the matrix* MM^T. *Then* $\left\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\right\}_{k=0}^{\infty}$ *converges to a stationary point of* (3.4).

Although the assumption (3.14) that the successive difference of the multipliers lies in the null space of M may seem unrealistically restrictive, it is satisfied for two special classes of problems. We have the following corollary.

Corollary 3.1 *Suppose that* $[M, C]$ *forms the standard Euclidean basis for* $\mathbf{R}^p$ *and* $\mathbf{z}^0 = M\mathbf{a}^0$ *for some vector* $\mathbf{a}^0 \in \mathbf{R}^p$ *with bounded norm. Then* (3.14) *is satisfied for all* k *and the sequence* $\left\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\right\}_{k=0}^{\infty}$ *generated by* (3.5), (3.6), *and* (3.7) *converges to a stationary point of* (3.4) *if* β *satisfies* (3.15).

Proof: If $[M, C]$ forms the standard basis, then we can write $\mathbf{y} \in R^p$ as $\mathbf{y} = M\mathbf{c} + C\mathbf{d}$ for some $\mathbf{c}$ and $\mathbf{d}$ with appropriate dimensions. It follows that we may decompose the subdifferential of the ℓ_1-norm at any $\mathbf{y} \in \mathbf{R}^p$ as

$$\partial\rho(\mathbf{y}) = \partial\rho\left(M\mathbf{c} + C\mathbf{d}\right) = \partial\rho\left(M\mathbf{c}\right) + \partial\rho\left(C\mathbf{d}\right) = M\partial\rho\left(\mathbf{c}\right) + C\partial\rho\left(\mathbf{d}\right) \tag{3.16}$$

by the fact that $\rho \circ M$ and $\rho \circ C$ are separable functions of $\mathbf{y}$. Substituting into the gradient condition of (3.9) we have

$$\begin{aligned} 0 &\in \gamma\partial\rho(M\mathbf{c}^1) + (\beta + \lambda)M\mathbf{c}^1 - (\beta M\mathbf{x}^0 + \mathbf{z}^0) \\ 0 &\in \gamma\partial\rho(C\mathbf{d}^1) + (\beta + \lambda)C\mathbf{d}^1. \end{aligned}$$

Therefore, we must have $(\beta + \lambda)C\mathbf{d}^1 \in -\gamma\partial\rho(C\mathbf{d}^1)$, which implies that $C\mathbf{d}^1 = 0$ by the structure of the subdifferential of the ℓ_1-norm. Extending this argument inductively shows that $C\mathbf{d}^k = 0$ for all k, or equivalently, $C^T\mathbf{y}^k = 0$ for all k. Substituting into (3.7) shows that

$$C^T(\mathbf{z}^{k+1} - \mathbf{z}^k) = \beta C^T(M\mathbf{x}^{k+1} - \mathbf{y}^{k+1}) = -\beta C^T(\mathbf{y}^{k+1}) = 0.$$

This completes the proof.

Similarly, if N has both full row and column rank, as it would if $W = 0$, then our ADMM algorithm converges. In particular, the algorithm converges when applied to (3.13) to identify the leading sparse principal component.

Corollary 3.2 *Suppose that N forms a basis for $\mathbf{R}^p$. Then* (3.14) *is satisfied for all k and the sequence $(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)$ generated by* (3.5), (3.6), *and* (3.7) *converges to a stationary point of* (3.4) *if β satisfies* (3.15).

Proof: If N forms a basis for $\mathbf{R}^p$, then M also forms a basis of $\mathbf{R}^p$, so its null space is spanned by $C = 0$. Clearly $C^T(\mathbf{z}^{k+1} - \mathbf{z}^k) = 0$ in this case.

The remainder of this section consists of a proof of Theorem 3.1. To establish Theorem 3.1, we will show that the value of the augmented Lagrangian of (3.4) decreases each iteration and the sequence of augmented Lagrangian values is bounded below. We will then exploit the fact that sequence of augmented Lagrangian values is convergent to show that the sequence of ADMM iterates is convergent. We conclude by establishing that a limit point of the sequence $\{(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}_{k=0}^{\infty}$ must be a stationary point of the original problem (3.4). We begin with the following lemma, which establishes that the augmented Lagrangian is decreasing provided the hypothesis of Theorem 3.1 is satisfied.

Lemma 3.1 *Suppose that $C^T(\mathbf{z}^{k+1} - \mathbf{z}^k) = 0$ for all k and $\beta > (\lambda_0 + 2)\|A\|/\lambda_0$. Then*

$$\begin{aligned} &L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^{k+1}) - L_\beta(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k) \\ &\quad \leq -\frac{\beta}{2}\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 - \frac{1}{2}\left(\beta - \|A\| - \frac{2\|A\|^2}{\beta\lambda_0}\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \end{aligned} \tag{3.17}$$

and the right-hand side of (3.17) *is strictly negative if $\mathbf{x}^{k+1} \neq \mathbf{x}^k$ or $\mathbf{y}^{k+1} \neq \mathbf{y}^k$.*

Proof: We will obtain the necessary bound on the improvement in Lagrangian value given by $L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^{k+1}) - L_\beta(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)$ by decomposing the difference as

$$L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^{k+1}) - L_\beta(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)$$

$$= \left(L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^{k+1}) - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^k)\right)$$
$$+ \left(L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^k) - L_\beta(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\right)$$

and bounding each summand in parentheses separately. We begin with the first summand $L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^{k+1}) - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^k)$. Recall that $\mathbf{x}^{k+1}$ satisfies (cf. (3.11))

$$(\beta I - A)\mathbf{x}^{k+1} = M^T\left(\beta \mathbf{y}^{k+1} - \mathbf{z}^k\right).$$

Multiplying (3.7) by M^T, using the fact that $M^T M = I$, and substituting the formula above yields

$$M^T\mathbf{z}^{k+1} = M^T\mathbf{z}^k + \beta\mathbf{x}^{k+1} - \beta M^T\mathbf{y}^{k+1} = \beta\mathbf{x}^{k+1} - M^T(\beta\mathbf{y}^{k+1} - \mathbf{z}^k) = A\mathbf{x}^{k+1}.$$

This implies that

$$\|M^T(\mathbf{z}^{k+1} - \mathbf{z}^k)\| = \|A(\mathbf{x}^{k+1} - \mathbf{x}^k)\| \leq \|A\|\|\mathbf{x}^{k+1} - \mathbf{x}^k\|.$$

Applying the assumption (3.14), we have

$$\lambda_0\|\mathbf{z}^{k+1} - \mathbf{z}^k\| \leq \|M^T(\mathbf{z}^{k+1} - \mathbf{z}^k)\| \leq \|A\|\|\mathbf{x}^{k+1} - \mathbf{x}^k\|.$$

It follows immediately that

$$\begin{aligned} &L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^{k+1}) - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^k) \\ &\quad = (\mathbf{z}^{k+1} - \mathbf{z}^k)^T(M\mathbf{x}^{k+1} - \mathbf{y}^{k+1}) \\ &\quad = \frac{1}{\beta}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 \leq \frac{\|A\|^2}{\beta\lambda_0^2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2. \end{aligned} \tag{3.18}$$

It remains to derive the necessary bound on $L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^k) - L_\beta(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)$.

To do so, note that subproblem (3.10) is strongly convex with modulus $(\beta - \|A\|)/2$. Let $f(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^T A\mathbf{x} + \frac{\beta}{2}\|M\mathbf{x} - \mathbf{y}^{k+1} + \mathbf{z}^k/\beta\|^2$. Then

$$\begin{aligned} L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^k) - L_\beta(\mathbf{x}^k, \mathbf{y}^{k+1}, \mathbf{z}^k) &= f(\mathbf{x}^{k+1}) - f(\mathbf{x}) \\ &\leq -\nabla f(\mathbf{x}^{k+1})^T(\mathbf{x}^k - \mathbf{x}^{k+1}) - \left(\frac{\beta - \|A\|}{2}\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \\ &\leq -\left(\frac{\beta - \|A\|}{2}\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \end{aligned} \tag{3.19}$$

by the fact that $\mathbf{x}^{k+1}$ is a minimizer of f and, consequently, $\nabla f(\mathbf{x}^{k+1})^T(\mathbf{x}^k - \mathbf{x}^{k+1}) \geq 0$. Note that (3.18) and (3.19) in tandem imply that

$$\begin{aligned} &L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^{k+1}) - L_\beta(\mathbf{x}^k, \mathbf{y}^{k+1}, \mathbf{z}^k) \\ &\quad \leq \frac{1}{2}\left(\|A\| - \beta - \frac{\|A\|^2}{\beta\lambda_0^2}\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2, \end{aligned}$$

which is strictly negative if $\mathbf{x}^{k+1} \neq \mathbf{x}^k$, by the assumption that $\beta > (\lambda_0 + 2)\|A\|/\lambda_0$. An identical argument yields the upper bound

$$L_\beta(\mathbf{x}^k, \mathbf{y}^{k+1}, \mathbf{z}^k) - L_\beta(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k) \leq -\frac{\beta}{2}\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2, \tag{3.20}$$

because (3.8) is strongly convex with modulus $\beta/2$. Combining (3.18), (3.19), and (3.20) gives the desired bound on the decrease of L.

Having established sufficient decrease of the augmented Lagrangian during each iteration, we next establish that the sequence of augmented Lagrangian values is bounded and, thus, is convergent. We have the following lemma.

Lemma 3.2 *Suppose that $C^T(\mathbf{z}^{k+1} - \mathbf{z}^k) = 0$ for all k and $\beta > (\lambda_0 + 2)\|A\|/\lambda_0$. Then the sequence $\{L_\beta(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}_{k=0}^\infty$ of augmented Lagrangian values is bounded. As a bounded monotonic sequence, $\{L_\beta(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}_{k=0}^\infty$ is convergent.*

Proof: Note that the fact that $C^T(\mathbf{z}^{k+1} - \mathbf{z}^k) = 0$ implies that $C^T\mathbf{y}^{k+1} = 0$ for all k. Indeed, in this case

$$0 = C^T(\mathbf{z}^{k+1} - \mathbf{z}^k) = \beta C^T(M\mathbf{x}^{k+1} - \mathbf{y}^{k+1}) = \beta C^T\mathbf{y}^{k+1}$$

because the columns of M are orthogonal to those of C. Thus, there exists $\{\mathbf{b}^k\}_{k=0}^\infty \in \mathbf{R}^d$, with $\|\mathbf{b}^k\| \leq 1$, such that $\mathbf{y}^k = M\mathbf{b}^k$. In this case,

$$\begin{aligned} &L_\beta(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k) \\ &\quad = -\frac{1}{2}(\mathbf{x}^k)^T A\mathbf{x}^k + \gamma\rho(\mathbf{y}^k) + (\mathbf{z}^k)^T(M\mathbf{x}^k - \mathbf{y}^k) + \frac{\beta}{2}\|M\mathbf{x}^k - \mathbf{y}^k\|^2 \\ &\quad = -\frac{1}{2}(\mathbf{x}^k)^T A\mathbf{x}^k + \gamma\rho(\mathbf{y}^k) + (\mathbf{z}^k)^T M(\mathbf{x}^k - \mathbf{b}^k) + \frac{\beta}{2}\|M\mathbf{x}^k - \mathbf{y}^k\|^2 \\ &\quad = -\frac{1}{2}(\mathbf{x}^k)^T A\mathbf{x}^k + \gamma\rho(\mathbf{y}^k) + (A\mathbf{x}^k)^T(\mathbf{x}^k - \mathbf{b}^k) + \frac{\beta}{2}\|M\mathbf{x}^k - \mathbf{y}^k\|^2 \qquad (3.21) \\ &\quad = -\frac{1}{2}(\mathbf{x}^k)^T A\mathbf{x}^k + \gamma\rho(\mathbf{y}^k) + (A\mathbf{x}^k)^T(\mathbf{x}^k - \mathbf{b}^k) + \frac{1}{2}(\mathbf{b}^k)^T A\mathbf{b}^k \\ &\qquad\quad - \frac{1}{2}(\mathbf{b}^k)^T A\mathbf{b}^k + \frac{\beta}{2}\|M\mathbf{x}^k - \mathbf{y}^k\|^2 \qquad (3.22) \end{aligned}$$

$$= \frac{1}{2}\Big(\|A^{1/2}(\mathbf{x}^k - \mathbf{b}^k)\|^2 - (\mathbf{b}^k)^T A \mathbf{b}^k + \beta \|M\mathbf{x}^k - \mathbf{y}^k\|^2 \Big) + \gamma\rho(\mathbf{y}^k)$$

where (3.21) follows from the identity $M^T\mathbf{z}^k = A\mathbf{x}^k$ and (3.22) follows from adding and subtracting $\frac{1}{2}(\mathbf{b}^k)^T A\mathbf{b}^k$. Note that $A^{1/2}$ is well-defined since A is a positive semidefinite matrix. Because both $\{\mathbf{b}^k\}_{k=0}^{\infty}$ and $\{\mathbf{y}^k\}_{k=0}^{\infty}$ are bounded, we may conclude that the sequence $\{L_\beta(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k)\}_{k=0}^{\infty}$ is lower bounded.

As an immediate consequence of Lemma 3.2, we see that each of the sequences $\{\mathbf{x}^k\}_{k=0}^{\infty}$, $\{\mathbf{y}^k\}_{k=0}^{\infty}$, and $\{\mathbf{z}^k\}_{k=0}^{\infty}$ is convergent. Indeed, we have the following corollary.

Corollary 3.3 *Suppose that $C^T(\mathbf{z}^{k+1} - \mathbf{z}^k) = 0$ for all k and $\beta > (\lambda_0 + 2)\|A\|/\lambda_0$. Then $\{\mathbf{x}^k\}_{k=0}^{\infty}$, $\{\mathbf{y}^k\}_{k=0}^{\infty}$, $\{\mathbf{z}^k\}_{k=0}^{\infty}$, and $\{M\mathbf{x}^k - \mathbf{y}^k\}_{k=0}^{\infty}$ are convergent, with*

$$\lim_{k\to\infty} M\mathbf{x}^k - \mathbf{y}^k = 0.$$

Proof: The fact that $L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \mathbf{z}^{k+1}) - L_\beta(\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k) \to 0$ and (3.17) imply that $\mathbf{x}^{k+1} - \mathbf{x}^k \to 0$ and $\mathbf{y}^{k+1} - \mathbf{y}^k \to 0$. The assumption that $C^T(\mathbf{z}^{k+1} - \mathbf{z}^k) = 0$ and the identity $M^T\mathbf{z}^k = A\mathbf{x}^k$ implies that

$$M^T(\mathbf{z}^{k+1} - \mathbf{z}^k) = A(\mathbf{x}^{k+1} - \mathbf{x}^k) \to 0.$$

Thus, $\mathbf{z}^{k+1} - \mathbf{z}^k \to 0$ because the columns of $[M, C]$ form an orthonormal basis for $\mathbf{R}^p$, which further implies that the constraint violation satisfies $M\mathbf{x}^k - \mathbf{y}^k = DN\mathbf{x}^k - \mathbf{y}^k \to 0$.

It remains to establish the following lemma, which states that any limit point of the sequence generated by (3.5), (3.6), and (3.7) is a stationary point of (3.4).

Lemma 3.3 *Let $\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}}$ be limit points of the sequences $\{\mathbf{x}^k\}_{k=0}^{\infty}$, $\{\mathbf{y}^k\}_{k=0}^{\infty}$, and $\{\mathbf{z}^k\}_{k=0}^{\infty}$, respectively. Then*

$$\bar{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbf{R}^d} \left\{ \mathbf{x}^T A\mathbf{x} + \bar{\mathbf{z}}^T(M\mathbf{x} - \bar{\mathbf{y}}) \right\} \tag{3.23}$$

$$\bar{\mathbf{y}} = \arg\min_{\mathbf{y}\in\mathbf{R}^p} \left\{ \gamma\rho(\mathbf{y}) + \bar{\mathbf{z}}^T(M\bar{\mathbf{x}} - \mathbf{y}) : \mathbf{y}^T\mathbf{y} \le 1 \right\} \tag{3.24}$$

$$\bar{\mathbf{y}} = M\bar{\mathbf{x}}. \tag{3.25}$$

Therefore, is a $(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}})$ stationary point of $L_0(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}})$, i.e., $0 \in \partial L_0(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}})$.

Proof: That (3.25) holds is a consequence of the fact that $M\mathbf{x}^k - \mathbf{y}^k \to 0$. The fact that $0 \in \partial L_0(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}})$ if (3.23) and (3.24) hold is an immediate consequence of the optimality

conditions for the subproblems for $\mathbf{x}$ and $\mathbf{y}$ applied at $(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}})$. It remains to prove that (3.23) and (3.24) hold.

We begin with (3.23); (3.24) will follow by a similar argument. Fix k. Recall that $\mathbf{x}^{k+1}$ is a minimizer of the function $f(x) = -\frac{1}{2}\mathbf{x}^T A\mathbf{x} + \frac{\beta}{2}\|M\mathbf{x} - \mathbf{y}^{k+1} + \mathbf{z}^k/\beta\|^2$. Therefore, $\mathbf{x}^{k+1}$ satisfies $\nabla f(\mathbf{x}^{k+1})^T(\mathbf{x} - \mathbf{x}^{k+1}) \geq 0$ for all $\mathbf{x} \in \mathbf{R}^d$. Evaluating the gradient of f at $\mathbf{x}^{k+1}$ shows that

$$\begin{aligned} 0 &\leq (\mathbf{x} - \mathbf{x}^{k+1})^T(-A\mathbf{x}^{k+1} + \beta(\mathbf{x}^{k+1} - M^T\mathbf{y}^{k+1}) + M^T\mathbf{z}^k) \\ &= (\mathbf{x} - \mathbf{x}^{k+1})^T(-A\mathbf{x}^{k+1} + M^T\mathbf{z}^{k+1}) \end{aligned}$$

by (3.7). This implies that $\mathbf{x}^{k+1}$ is also a minimizer of $-\frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{x}^T M^T\mathbf{z}^{k+1}$. It follows that

$$-\frac{1}{2}\mathbf{x}^T A\mathbf{x} + (\mathbf{z}^{k+1})^T(M\mathbf{x} - \mathbf{y}^{k+1}) \geq -\frac{1}{2}(\mathbf{x}^{k+1})^T A\mathbf{x}^{k+1} + (\mathbf{z}^{k+1})^T(M\mathbf{x}^{k+1} - \mathbf{y}^{k+1})$$

for all $\mathbf{x} \in \mathbf{R}^d$. Taking the limit as $k \to \infty$ shows that

$$-\frac{1}{2}\mathbf{x}^T A\mathbf{x} + \bar{\mathbf{z}}^T(M\mathbf{x} - \bar{\mathbf{y}}) \geq -\frac{1}{2}\bar{\mathbf{x}}^T A\bar{\mathbf{x}} \ + \bar{\mathbf{z}}^T(M\bar{\mathbf{x}} - \bar{\mathbf{y}})$$

for all $\mathbf{x} \in \mathbf{R}^d$, which establishes (3.23).

By an identical argument, every iterate $\mathbf{y}^{k+1}$ satisfies

$$\gamma(\rho(\mathbf{y}) - \rho(\mathbf{y}^{k+1}) + (\mathbf{y} - \mathbf{y}^{k+1})^T(\mathbf{z}^{k+1} + \beta M(\mathbf{x}^{k+1} - \mathbf{x}^k)) \geq 0$$

for all $\mathbf{y} \in \mathbf{R}^p$. Taking the limit as $k \to \infty$ shows that

$$\gamma\rho(\mathbf{y}) + \bar{\mathbf{z}}^T(M\bar{\mathbf{x}} - \mathbf{y}) \geq \gamma\rho(\bar{\mathbf{y}}) + \bar{\mathbf{z}}^T(M\bar{\mathbf{x}} - \bar{\mathbf{y}})$$

as required. This completes the proof.

4 Numerical Results

We performed a series of numerical experiments to compare our proposed algorithm (SZVD) with two recently proposed heuristics for penalized discriminant analysis, namely the PLDA [41] and SDA [11] methods discussed in Section 2.3 implemented as the R packages **penalizedLDA** [40] and **sparseLDA** [10] respectively.

In each experiment, we learn sets of $K - 1$ discriminant vectors from given training data using each heuristic, and then test classification performance on a given test set. For each

method, we apply validation to choose regularization parameters minimizing the validation criterion

$$\frac{\#\text{misclassified}}{\#\text{validation obs}} + \left(\frac{1}{2}\right)\frac{\#\text{ nonzero features}}{p(K-1)};$$

that is, for each method, we choose regularization parameters minimizing a weighted sum of the fraction of misclassified observations in the validation set and fraction of nonzero features of the obtained discriminant vectors. For each data set and $i = 1, \ldots, K-1$, this validation, applied to our ADMM heuristic (SZVD), selects the regularization parameter γ_i in (3.4) from a set of m evenly spaced values in the interval $[0, \tilde{\gamma}_i]$, where

$$\tilde{\gamma}_i := \frac{(\mathbf{w}_0)_i^T B(\mathbf{w}_0)_i}{\rho((\mathbf{w}_0)_i)}$$

and $(\mathbf{w}_0)_i$ is the ith unpenalized zero-variance discriminant vector; this choice of $\tilde{\gamma}_i$ is made to ensure that the problem (3.4) has a nontrivial optimal solution by guaranteeing that at least one nontrivial solution with nonpositive objective value exists for each potential choice of γ_i. Similarly, in PLDA we perform validation on the tuning parameter λ controlling sparsity of the discriminant vectors in (2.8). Finally, SDA employs two tuning parameters, λ_1, which controls the ridge regression penalty, and λ_2 ("loads" in the R package), which controls the number of nonzero features; in each experiment we fix λ_1 and perform validation to choose λ_2.

In all experiments, we use the dictionary matrix $D = I$, regularization parameter $\beta = 2$, and stopping tolerances $tol_{abs} = tol_{rel} = 10^{-4}$ in our ADMM heuristic SZVD. The initial primal iterates were set equal to the unpenalized zero-variance discriminant vectors given by (2.7) and the initial dual solution $\mathbf{z}^0$ was set equal to 0. All features of discriminant vectors $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{K-1}\}$ found using SZVD with magnitude less than 0.025 were rounded to 0. Any obtained trivial discriminant vectors were discarded and dimensionality-reduction and classification was performed using the nontrivial discriminant vectors. All experiments were performed in R; R and Matlab code for SZVD and R code for generating the synthetic data sets can be found on the authors' webpages.

4.1 Simulated Data

Two sets of simulations were considered. For each $K \in \{2, 4\}$ and $p = 500$, we generate $25K$, $25K$, and $250K$ training, validation, and test observations in $\mathbf{R}^p$, respectively. For each $i \in \{1, 2, \ldots, K\}$, we sample 25, 25, and 250 Gaussian training, validation, and test observations belonging to class C_i from the distribution $N(\boldsymbol{\mu}_i, \Sigma)$, where the mean vector $\boldsymbol{\mu}_i$

is defined by

$$[\boldsymbol{\mu}_i]_j = \begin{cases} 0.7, & \text{if } 100(i-1)+1 \leq j \leq 100i \\ 0, & \text{otherwise} \end{cases}$$

and the covariance matrix Σ is chosen in one of two ways:

- In the first set of experiments, all features are correlated with

$$[\Sigma_r]_{k\ell} = \begin{cases} 1, & \text{if } k = \ell \\ r, & \text{otherwise.} \end{cases}$$

  The experiment was repeated for each choice of $r \in \{0, 0.1, 0.5, 0.9\}$.

- In the second set of experiments, Σ is a block diagonal matrix with 100×100 diagonal blocks. For each (k, ℓ) pair with k and ℓ belonging to the same block, we have

$$[\Sigma_\alpha]_{k\ell} = \alpha^{|k-\ell|}.$$

  We let $[\Sigma_\alpha]_{k\ell} = 0$ for all remaining (k, ℓ) pairs. The experiment was repeated for all choices of $\alpha \in \{0.1, 0.5, 0.9\}$.

For each (K, r) and (K, α) pair, we applied unpenalized zero-variance discriminant analysis (ZVD), our ADMM heuristic for penalized zero-variance discriminant analysis (SZVD), Witten and Tibshirani's penalized linear discriminant analysis with ℓ_1-norm and fused lasso penalties (PL1 and PFL), and the SDA algorithm of Clemmensen et al. (SDA) to obtain sets of $K-1$ discriminant vectors from the sampled training set. These discriminant vectors were then used to perform dimensionality reduction of the test data, and each observation in the test set was assigned to the class of the nearest projected training class centroid; an identical process was applied to the validation data to train any regularization parameters. Both versions of PLDA chose the tuning parameter λ using 20 equally spaced values on the interval $[0, 0.15]$ by validation, while SDA used the ridge regression parameter $\lambda_1 = 10$ and chose the sparsity tuning parameter "loads" from the set

$$\{-500, -400, -300, -250, -200, -150, -120, -100, -80, -70, -60, -50\}$$

by validation; the inner optimization of the SDA algorithm was stopped after a maximum of 5 iterations. This process was repeated 20 times for each (K, r) and (K, α) pair. Tables 4.1 and 4.2 report the average and standard deviation over all 20 trials for each set of (K, r) and (K, α) pairs, respectively, of the number of misclassification errors, number of nonzero features of discriminant vectors, and time (in seconds) required to obtain each set of discriminant vectors per validation step for each of the five heuristics.

Simulation 1		ZVD	SZVD	PL1	PFL	SDA
$K = 2$	*Err*	0 (0)	2.8 (2.8)	2.1 (3.2)	0.1 (0.2)	8.7 (5.9)
$r = 0$	*Feat*	490.9 (3.1)	98.6 (13.5)	104.4 (26.5)	190.8 (13.9)	60.8 (10.1)
	Time	0.1 (0.003)	1.2 (0.08)	0.02 (0.003)	0.03 (0.005)	2.0 (0.02)
$K = 2$	*Err*	0 (0)	1.6 (1.9)	12.2 (17.6)	11 (26.7)	17.0 (14.7)
$r = 0.1$	*Feat*	490.9 (2.7)	104.2 (10.5)	113.2 (33.4)	178.1 (31.9)	57.8 (8.2)
	Time	0.10 (0.004)	1.1 (0.08)	0.02 (0.002)	0.03 (0.002)	2.04 (0.025)
$K = 2$	*Err*	0 (0)	0.1 (0.447)	86.55 (49.55)	58.6 (57.62)	56.9 (46.09)
$r = 0.5$	*Feat*	488.7 (3.8)	112.7 (10.3)	139.0 (44.0)	184.3 (44.5)	51.4 (4.7)
	Time	0.10 (0.003)	1.1 (0.06)	0.02 (0.001)	0.04 (0.003)	2.0 (0.01)
$K = 2$	*Err*	0 (0)	0 (0)	100.2 (68.5)	95.8 (71.7)	115.2 (58.0)
$r = 0.9$	*Feat*	485.5 (3.3)	143.8 (9.4)	171.4 (60.6)	160.3 (50.7)	49.2 (0.9)
	Time	0.10 (0.004)	1.1 (0.051)	0.03 (0.004)	0.04 (0.005)	2.01 (0.02)
$K = 4$	*Err*	0.9 (1.0)	23 (10.8)	6.7 (7.9)	0.2 (0.5)	18 (17.1)
$r = 0$	*Feat*	1473.8 (4.6)	312.4 (59.3)	364.8 (68.8)	391.2 (12.2)	280.8 (119.0)
	Time	0.19 (0.005)	5.8 (0.4)	0.12 (0.006)	0.20 (0.01)	14.0 (0.7)
$K = 4$	*Err*	0.6 (0.7)	23.3 (16.7)	115.2 (36.4)	116.6 (23.5)	53.1 (38.3)
$r = 0.1$	*Feat*	1474.0 (3.8)	305.0 (66.1)	402.6 (109.0)	428.4 (59.1)	313.7 (211.1)
	Time	0.19 (0.003)	5.8 (0.4)	0.13 (0.01)	0.21 (0.02)	14.3 (1.4)
$K = 4$	*Err*	0.1 (0.2)	30.9 (28.6)	318.4 (40.3)	307.6 (43.7)	39.5 (32.2)
$r = 0.5$	*Feat*	1475.1(3.8)	239.8 (73.7)	369.9 (61.1)	388.9 (81.5)	369.3 (194.9)
	Time	0.20 (0.01)	6.1 (0.3)	0.14 (0.01)	0.23 (0.02)	14.1 (1.4)
$K = 4$	*Err*	0 (0)	0 (0)	412.0 (55.9)	415.5 (60.5)	1.0 (4.5)
$r = 0.9$	*Feat*	1471.1(6.2)	336.0 (136.2)	386.2 (100.4)	369.7 (92.4)	371.3 (288.7)
	Time	0.20 (0.01)	6.0 (0.4)	0.15 (0.02)	0.23 (0.02)	14.0 (2.5)

Table 4.1: Comparison of performance for synthetic data in $\mathbf{R}^{500}$ drawn from classes $C_1, \ldots, C_K \sim N(\boldsymbol{\mu}_i, \Sigma_r)$ where Σ_r is matrix with diagonal equal to 1 and all other entries equal to r. All values reported in the format "mean (standard deviation)".

4.2 Time-Series Data

We performed similar experiments for three data sets drawn from the UCR time series data repository [28], namely the *Coffee, OliveOil,* and *ECGFiveDays* data sets. The *ECGFiveDays* data set consists of 136-dimensional electrocardiogram measurements of a 67-year old male. Each observation corresponds to a measurement of the electrical signal of a single heartbeat of the patient. The data consists of two classes, 884 observations in total, corresponding to measurements taken on two dates, five days apart. We randomly divided the data into training, validation, and testing sets containing 25, 1000, and 759 observations, respectively.

Simulation 2		ZVD	SZVD	PL1	PFL	SDA
$K = 2$	*Errs*	0.3 (0.7)	5.1 (4.8)	3.5 (2.9)	0.1 (0.3)	8.9 (4.5)
$\alpha = 0.1$	*Feat*	491.3 (2.7)	102.1 (17.8)	110.7 (28.5)	186.9 (11.9)	61.8 (11.1)
	Time	0.96 (0.02)	1.1 (0.09)	0.5 (0.06)	0.03 (0.003)	2.0 (0.02)
$K = 2$	*Errs*	10.6 (3.7)	28.4 (11.2)	15.9 (9.5)	6.1 (3.6)	25.0 (8.3)
$\alpha = 0.5$	*Feat*	490.8 (2.4)	100.4 (14.8)	112.6 (31.9)	173.1 (22.1)	65.7 (18.5)
	Time	0.96 (0.02)	1.2 (0.1)	0.45 (0.02)	0.03 (0.002)	2.0 (0.02)
$K = 2$	*Errs*	214.8 (37.6)	229.4 (38.0)	88.4 (19.7)	85.3 (26.1)	101.8 (18.7)
$\alpha = 0.9$	*Feat*	491.0 (3.1)	89.8 (22.7)	136.9 (45.0)	163.1 (53.6)	55.3 (8.3)
	Time	1.0 (0.04)	1.6 (0.1)	0.45 (0.02)	0.04 (0.004)	2.1 (0.04)
$K = 4$	*Errs*	3.1 (2.3)	32.8 (17.3)	7.3 (10.0)	0.8 (1.3)	27.9 (28.7)
$\alpha = 0.1$	*Feat*	1473.7 (3.6)	306.6 (47.7)	380.0 (54.7)	384.5 (29.7)	333.3 (198.7)
	Time	2.0 (0.03)	6.0 (0.4)	2.5 (0.1)	0.20 (0.01)	14.4 (1.3)
$K = 4$	*Errs*	54.1 (10.9)	122.8 (36.9)	47.0 (17.1)	36.2 (14.9)	96.0 (35.2)
$\alpha = 0.5$	*Feat*	1472.8 (5.7)	313.4 (66.4)	391.0 (57.7)	394.5 (46.6)	256.9 (161.9)
	Time	2.0 (0.2)	6.1 (0.4)	2.5 (0.1)	2.0 (0.01)	15.5 (1.4)
$K = 4$	*Errs*	473.5 (28.0)	505.9 (35.1)	363.2 (29.8)	365.6 (38.8)	382.2 (29.5)
$\alpha = 0.9$	*Feat*	1472.4 (4.7)	330.6 (77.0)	369.1 (78.6)	368.4 (86.1)	235.9 (113.1)
	Time	2.0 (0.05)	7.2 (0.4)	2.6 (0.3)	0.2 (0.02)	17.0 (0.9)

Table 4.2: Comparison of performance for synthetic data in $\mathbf{R}^{500}$ drawn from classes $C_1, \ldots, C_K \sim N(\boldsymbol{\mu}_i, \Sigma_\alpha)$ where Σ_α is a 500×500 diagonal block matrix with 100×100 diagonal blocks with (i, j) nonzero entry equal to $\alpha^{|i-j|}$.

We then applied each of our five heuristics to obtain discriminant vectors using each training and validation set pair and perform nearest centroid classification on each corresponding test set in the projected space. The tuning parameter λ in PLDA was selected from twenty equally spaced values in the interval $[0, 0.15]$, and we set $\lambda_1 = 0.001$ and chose the tuning parameter "loads" from the set

$$\{-500, -400, -300, -250, -200, -150, -120, -100, -80, -60, -50, -40, -30, -20, -10\}$$

by validation when using SDA. As before, we stop the SDA inner optimization after 5 iterations. We repeated this process for 20 (training, validation, testing)-splits of the data and recorded the results in Table 4.3.

The *OliveOil* and *Coffee* data sets comprise 60 and 56 food spectrogram observations of different kinds of olive oil and coffee, respectively. Here, mass spectroscopy is applied to generate signals (spectra) corresponding to the molecular composition of samples of each food. The goal is to distinguish between different varieties of olive oil and coffee from

Time-Series		ZVD	SZVD	PL1	PFL	SDA
OliveOil	*Err*	1.8 (1.1)	2.5 (1.9)	5.1 (2.3)	5.0 (2.3)	2.0 (1.1)
$p = 570$	*Feat*	1669.4 (6.2)	319.4 (63.5)	224.2 (145.7)	231.7 (138.9)	113.6 (112.7)
$K = 4$	*Time*	3.2 (0.06)	16.6 (0.9)	0.04 (0.005)	0.10 (0.01)	10.1 (0.4)
Coffee	*Err*	0.1 (0.2)	0.1 (0.3)	7.7 (3.1)	7.4 (3.3)	2.2 (2.3)
$p = 286$	*Feat*	281.7 (2.2)	53.8 (13.2)	115.3 (62.0)	131.9 (79.8)	18.0 (19.4)
$K = 2$	*Time*	0.23 (0.02)	1.1 (0.3)	0.02 (0.005)	0.03 (0.009)	0.91 (0.03)
ECG	*Err*	35.2 (21.5)	42.6 (26.0)	160.7 (46.7)	164.8 (45.6)	60.6 (31.0)
$p = 136$	*Feat*	127.1 (0.8)	32.8 (7.8)	34.6 (15.7)	35.4 (21.4)	15.0 (7.6)
$K = 2$	*Time*	0.05 (0.006)	0.17 (0.03)	0.02 (0.001)	0.03 (0.003)	0.36 (0.02)

Table 4.3: Comparison of performance for the *OliveOil, Coffee, and ECGFiveDays* data sets.

these spectral signals. The *OliveOil* data set [36] consists of 570-dimensional spectrograms corresponding to samples of extra virgin olive oil from one of four countries (Greece, Italy, Portugal, or Spain); 286-dimensional spectrograms of either Arabica or Robusta variants of instant coffee compose the *Coffee* data set [7]. As before, we divide the *OliveOil* data into training, validation, and testing sets containing $(30, 10, 20)$ observations, respectively. We then applied each of our five heuristics to learn a classification rule from the training and validation data and classify the given test data. For each PLDA heuristic, we used the same range of tuning parameter λ as in the *ECGFiveDays* trials; we stopped the inner optimization step after 5 iterations, set $\lambda_1 = 0.1$, and used the same set of potential values of the tuning parameter "loads" as we did in *ECGFiveDays* trials for SDA. This process was repeated for 20 different data splits, and we then repeated the experiment for the *Coffee* data set using training, validation, test splits of size $(25, 10, 21)$. The results of these trials are summarized in Table 4.3.

4.3 Commentary

As can be seen from the experiments of the previous section, our proposed algorithm, SZVD, compares favorably to the current state of the art. When compared to the zero-variance discriminant, adding penalization in the form of a ℓ_1-penalty results in a modest degradation in classification performance, as may be expected. However, this penalization significantly increased sparsity of the obtained discriminant vectors from that of the zero-variance discriminants. Moreover, the average misclassification error for SZVD was smaller than (or comparable to) that for PLDA and SDA in almost all trials, while obtaining discriminant vectors with a similar number of nonzero features. In terms of computational complexity, SZVD was slower than PLDA, while representing a significant increase in classification per-

formance for correlated data, but faster than SDA for most trials, which gave only a modest improvement in classification performance, if any. It should also be noted that, although we did not verify that the conditions guaranteeing convergence of our ADMM heuristic given by Theorem 3.1 are satisfied (and there is no reason to expect them to be), the ADMM heuristic converged in all trials after at most a few hundred iterations.

There were two notable exceptions. First, for uncorrelated data, i.e., the $r = 0$ case in Simulation 1, both variants of PLDA outperformed SZVD in terms of classification error. This is not surprising, as the implicit assumption that the data are uncorrelated made when using the diagonal approximation holds for this special case. However, the performance of PLDA degrades significantly as r is increased, while that of SZVD improves. Second, SZVD performs very poorly for highly correlated data in Simulation 2; roughly half of all test observations are misclassified in the $\alpha = 0.9$ trials. It should be noted that SZVD performs only marginally worse than unpenalized zero-variance discriminant analysis (ZVD), which suggests that the classes may not be linearly separable in the null space of the within-class covariance matrix in this case. It should also be noted that none of the heuristics perform well for these particular synthetic data sets, with PLDA and SDA misclassifying at least one third of test observations on average.

The use of penalization, aside from encouraging sparsity of the discriminant vectors, also seems to increase interpretability of the discriminant vectors. For example, the nonzero entries of the discriminant vector used to classify the *ECGFiveDays* data align with features where data in different classes appear to differ significantly (on average). Specifically, both the zero-variance and SZVD discriminant vectors closely follow the trajectory of the difference of the class-means vectors $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. However, most of the entries of the SZVD discriminant vector corresponding to small magnitude entries of the zero-variance discriminant are set equal to zero; the remaining nonzero entries of the SZVD discriminant vector correspond to features where the two class mean vectors differ the most significantly. This is most apparent when comparing the discriminant vectors to the class mean vectors of the data after centering and normalizing, although this phenomena is also weakly visible when comparing class means for the original data set. See Figure 4.1 for more details.

5 Acknowledgments

This research was supported in part by the Institute for Mathematics and its Applications with funds provided by the National Science Foundation. We are grateful to Fadil Santosa, Krystal Taylor, Zhi-Quan Luo, and Meisam Razaviyayn for their insight and helpful suggestions.

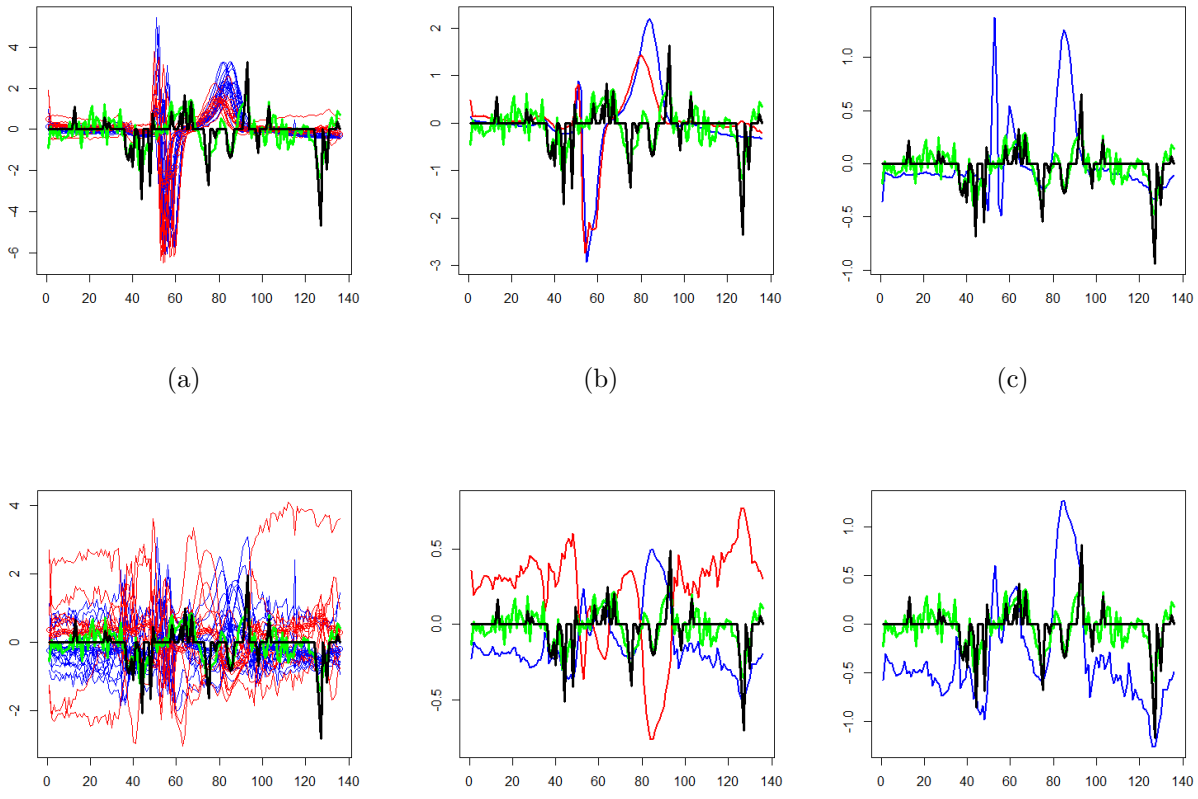

Figure 4.1: Plots of the ZVD (green) and SZVD (black) discriminant vectors for *ECGFiveDays* data set plotted with (a) all observations in Class 1 (blue) or Class 2 (red), class means (blue or red circles), (b) class means only, and (c) difference of class means (blue). Plots (d), (e), and (f) contain identical plots for the centered and normalized time-series. Nonzero components of the SZVD discriminant vectors occur at peaks of the difference of class-means vector after centering and normalization. The discriminant vectors were rescaled in each plot to to emphasize nonzero values of largest magnitude aligning with large differences between the two classes.

References

[1] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012. 7

[2] Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006. 7

[3] Peter Bickel and Elizaveta Levina. Some theory for Fisher's linear discriminant func-

tion, naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004. 5

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011. 9, 11, 12

[5] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004. 10

[6] Stephen Boyd and Lieven Vandenberghe. Subgradients. *Lecture Notes for EE364b, Stanford University, Winter 2006-07*, 2008. Available from `http://see.stanford.edu/materials/lsocoee364b/01-subgradients_notes.pdf`. 10

[7] Romain Briandet, E Katherine Kemsley, and Reginald H Wilson. Discrimination of Arabica and Robusta in instant coffee by Fourier transform infrared spectroscopy and chemometrics. *Journal of agricultural and food chemistry*, 44(1):170–174, 1996. 23

[8] Emmanuel J Candès. Compressive sampling. In *Proceedings oh the International Congress of Mathematicians: Madrid, August 22-30, 2006: invited lectures*, pages 1433–1452, 2006. 2

[9] Line Clemmensen. *On discriminant analysis techniques and correlation structures in high dimensions*. Technical Report-2013. Technical University of Denmark, 2013. 8

[10] Line Clemmensen and contributions by Max Kuhn. *sparseLDA: Sparse Discriminant Analysis*, 2012. R package version 0.1-6. 18

[11] Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4), 2011. 7, 8, 18

[12] Alexandre d'Aspremont, Francis Bach, and Laurent El Ghaoui. Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research*, 9:1269–1294, 2008. 7

[13] Alexandre d'Aspremont, Francis Bach, and Laurent El Ghaoui. Approximation bounds for sparse principal component analysis. *arXiv preprint arXiv:1205.0121*, 2012. 7

[14] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert RG Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM review*, 49(3):434–448, 2007. 7

[15] Sandrine Dudoit, Jane Fridlyand, and Terence P Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87, 2002. 5

[16] Jurjen Duintjer Tebbens and Pavel Schlesinger. Improving implementation of linear discriminant analysis for the high dimension/small sample size problem. *Computational Statistics & Data Analysis*, 52(1):423–437, 2007. 6

[17] Jonathan Eckstein and Dimitri P Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992. 12

[18] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936. 3

[19] Jerome H Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989. 5

[20] Gene H Golub and Charles F Van Loan. *Matrix computations*. Johns Hopkins University Press, 1996. 11

[21] Yaqian Guo, Trevor Hastie, and Robert Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100, 2007. 8

[22] Trevor Hastie, Andreas Buja, and Robert Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, pages 73–102, 1995. 7

[23] Trevor Hastie, Robert Tibshirani, and J Jerome H Friedman. *The elements of statistical learning*. Springer New York, 2013. 3, 6

[24] Mingyi Hong and Zhi-Quan Luo. On the linear convergence of the alternating direction method of multipliers. *arXiv preprint arXiv:1208.3922*, 2012. 12

[25] Iain M Johnstone and Arthur Yu Lu. Sparse principal components analysis. *Unpublished manuscript*, 2004. 7

[26] Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003. 7

[27] Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research*, 11:517–553, 2010. 7

[28] Eamonn Keogh, Xiaopeng Xi, Li Wei, and Chotirat Ann Ratanamahatana. The UCR time series classification/clustering homepage, 2006. 21

[29] Wojtek Krzanowski, Philip Jonathan, WV McCarthy, and MR Thomas. Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Applied statistics*, pages 101–115, 1995. 5, 8

[30] Gitta Kutyniok. Compressed sensing: Theory and applications. *CoRR, vol. abs/1203.3815*, 2012. 2

[31] Ronny Luss and Marc Teboulle. Convex approximations to sparse PCA via lagrangian duality. *Operations Research Letters*, 39(1):57–61, 2011. 7

[32] Dimitris S Papailiopoulos, Alexandros G Dimakis, and Stavros Korokythakis. Sparse PCA through low-rank approximations. *arXiv preprint arXiv:1303.0551*, 2013. 7

[33] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617, 2007. 7

[34] Peter Richtárik, Martin Takáč, and Selin Damla Ahipaşaoğlu. Alternating maximization: Unifying framework for 8 sparse PCA formulations and efficient parallel codes. *arXiv preprint arXiv:1212.4137*, 2012. 7

[35] Jun Shao, Yazhen Wang, Xinwei Deng, and Sijian Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, 39(2):1241–1265, 2011. 8

[36] Henri S Tapp, Marianne Defernez, and E Katherine Kemsley. Ftir spectroscopy and multivariate analysis can distinguish the geographic origin of extra virgin olive oils. *Journal of agricultural and food chemistry*, 51(21):6110–6115, 2003. 23

[37] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 2

[38] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, pages 104–117, 2003. 8

[39] Max Welling. Fisher linear discriminant analysis. *Department of Computer Science, University of Toronto*, 2005. 3

[40] Daniela Witten. *penalizedLDA: Penalized classification using Fisher's linear discriminant*, 2011. R package version 1.0. 18

[41] Daniela M Witten and Robert Tibshirani. Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772, 2011. 7, 8, 18

[42] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009. 7

[43] Ping Xu, Guy N Brock, and Rudolph S Parrish. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computational Statistics & Data Analysis*, 53(5):1674–1687, 2009. 5

[44] Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *arXiv preprint arXiv:1112.2679*, 2011. 7

[45] Youwei Zhang, Alexandre dAspremont, and Laurent El Ghaoui. Sparse PCA: Convex relaxations, algorithms and applications. In *Handbook on Semidefinite, Conic and Polynomial Optimization*, pages 915–940. Springer, 2012. 7

[46] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. 7

[47] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006. 7