Simultaneous Sparse Estimation of Canonical Vectors in the $p \gg N$ Setting

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

ABSTRACT
This article considers the problem of sparse estimation of canonical vectors in linear discriminant analysis when $p \gg N$. Several methods have been proposed in the literature that estimate one canonical vector in the two-group case. However, $G - 1$ canonical vectors can be considered if the number of groups is $G$. In the multi-group context, it is common to estimate canonical vectors in a sequential fashion. Moreover, separate prior estimation of the covariance structure is often required. We propose a novel methodology for direct estimation of canonical vectors. In contrast to existing techniques, the proposed method estimates all canonical vectors at once, performs variable selection across all the vectors and comes with theoretical guarantees on the variable selection and classification consistency. First, we highlight the fact that in the $N \gg p$ setting the canonical vectors can be expressed in a closed form up to an orthogonal transformation. Secondly, we propose an extension of this form to the $p \gg N$ setting and achieve feature selection by using a group penalty. The resulting optimization problem is convex and can be solved using a block-coordinate descent algorithm. The practical performance of the method is evaluated through simulation studies as well as real data applications. Supplementary materials for this article are available online.

1. Introduction
Recent technological advances have generated high-dimensional datasets across a wide variety of application areas such as finance, atmospheric science, astronomy, biology, and medicine. Not only do these datasets provide computational challenges, but they also motivate new statistical challenges as the traditional methods are no longer sufficient. Linear discriminant analysis (LDA) is a classical classification and data visualization tool that is used in the $N \gg p$ setting. LDA seeks the linear combinations of features that maximize between group variability with respect to within group variability (Mar-dia, Kent, and Bibby 1979, chap. 11). These linear combinations are called canonical vectors and they provide a low-dimensional representation of the data by reducing the original feature space dimension $p$ to $G - 1$, where $G$ is the total number of groups.

The classical use of LDA when $p \gg N$ fails to provide useful results because of the singularity of covariance matrix and over-selection of relevant features (Dudoit, Fridlyand, and Speed 2002; Bickel and Levina 2004). As a result, the extension of LDA to high-dimensional settings has recently received a lot of attention. A number of these proposals result in non-sparse classifiers. Friedman (1989), Krzanowski et al. (1995) and Xu, Brock, and Parrish (2009) regularize the within-class covariance matrix to obtain a positive definite estimate. Other approaches that lead to sparse discriminant vectors have also been considered. Tibshirani et al. (2002) propose the shrunk centroids methodology by adapting the naive Bayes classifier and soft-thresholding the mean vectors. Guo, Hastie, and Tibshirani (2007) combine the shrunk centroids approach with a ridge-type penalty on the within-class covariance matrix. Witten and Tibshirani (2011) apply an $\ell_1$ penalty to the Fisher’s discriminant problem to obtain sparse discriminant vectors. Clemmensen et al. (2011) use an optimal scoring approach which essentially reduces the sparse discriminant vector construction to a penalized regression problem.

In the two-group setting, Cai and Liu (2011) and Mai, Zou, and Yuan (2012) propose direct estimation of the canonical vector, thus, avoiding separate estimation of the covariance matrix. Simulations and real data applications show that the direct estimation approach results in reduced misclassification rates in comparison to alternative methods. The corresponding optimization problems can be solved efficiently for large datasets and have desirable theoretical properties. Unfortunately, the extension of two-group methods to the multi-group case is nebulous (Hastie, Tibshirani, and Friedman 2009, p. 658). Popular approaches include “one-versus-all” and “one-versus-one” methods, where the final classification assignment is usually based on the “majority vote.” As such, computation of more vectors is required ($G$ and $G(G - 1)/2$ versus $G - 1$).

Witten and Tibshirani (2011) and Clemmensen et al. (2011) propose estimating canonical vectors in a sequential fashion in the multi-group setting: starting with the first canonical vector $v_1$, with subsequent $v_i$ found subject to orthogonality constraints. This approach is undesirable from a computational viewpoint, as well as from an estimation perspective. Each subsequent canonical vector $v_i$ relies on all the previous estimates $v_k$ for $k < i$, hence propagation of the estimation error is possible. In addition, the corresponding optimization problems are nonconvex, hence, the convergence of the optimization algorithms to the global solution is not assured. This computational
2. Methodology

2.1 Notation

For a vector $b \in \mathbb{R}^p$ we define $\|b\|_\infty = \max_{i=1,\ldots,p} |b_i|$, $\|b\|_1 = \sum_{i=1}^p |b_i|$, and $\|b\|_2 = \sqrt{\sum_{i=1}^p b_i^2}$. For a matrix $M \in \mathbb{R}^{n \times p}$ we define $m_i$ to be its $i$th row and $M_j$ to be its $j$th column. We also define $\|M\|_\infty = \max_{i=1,\ldots,n} \|m_i\|_1$, $\|M\|_2 = \max_{i=1,\ldots,n} |m_i|_2$, $\|M\|_\infty,2 = \max_{i=1,\ldots,n} |m_i|_2$, $\|M\|_F = \sqrt{\sum_{i,j=1}^{p} m_{ij}^2}$, and $\|M\|_\ast = \sum_{i=1}^{n(p)} \Sigma_i(M)$, where $\Sigma_i(M)$ is the $i$th singular value of $M$. We define $\mathcal{O}^p$ to be the space of $p \times p$ matrices $R$ such that $RR^\top = R^\top R = I$.

2.2 Estimation Problem

We assume that $X_i \in \mathbb{R}^p$, $i = 1, \ldots, N$, are independent and come from $G$ groups with different means and the same covariance matrix, that is $E(X_i|Y_i=g) = \mu_g$ and Cov$(X_i|Y_i=g) = \Sigma_W$, where $Y_i \in \{1, \ldots, G\}$. The between-group population covariance matrix $\Sigma_B$ is defined as

$$\Sigma_B = \sum_{g=1}^{G} \pi_g (\mu_g - \mu)(\mu_g - \mu)^\top,$$

where $\pi_g = P(Y_i = g)$ are group-specific probabilities and $\mu = \sum_{g=1}^{G} \pi_g \mu_g$ is the overall population mean. The population canonical vectors $\Psi$ are defined as eigenvectors corresponding to nonzero eigenvalues of $\Sigma_B^{-1}$. Although the eigenvectors are unique only up to normalization (Golub and Van Loan 2012), we take advantage of the uniqueness of the eigenspace in defining a scale-invariant classification rule. For a new observation value of $X$, $x \in \mathbb{R}^p$, the population classification rule $h_\Psi(x)$ is defined as

$$h_\Psi(x) = \arg \min_{1 \leq g \leq G} \left\{ (x - \mu_g)^\top \Psi (\Psi^\top \Sigma_B \Psi)^{-1} \Psi^\top (x - \mu_g) - 2 \log \pi_g \right\}. \quad (1)$$

The classification rule is based on the closest Mahalanobis distance in the projected space defined by $\Psi$, after adjustment for potential discrepancy in the prior group probabilities $\pi_g$. Through the addition of $2 \log \pi_g$ term, the resulting classification rule mimics the optimal classification rule under the assumption that the data comes from the multivariate normal group-conditional distribution (McLachlan 1992, chap. 3.9.3). Our goal is to identify the eigenspace spanned by $\Psi$ based on the sample observations $X_i \in \mathbb{R}^p$ and sample labels $Y_i \in \{1, \ldots, G\}$.

Consider the within-group sample covariance matrix $W = \frac{1}{N-g} \sum_{g=1}^{G} (n_g - 1) S_g$ and the between-group sample covariance matrix $B = \frac{1}{G} \sum_{g=1}^{G} n_g (\bar{X}_g - \bar{X}) (\bar{X}_g - \bar{X})^\top$, where $n_g$ is the number of observations in group $g$, $S_g$ is the sample covariance matrix for group $g$, $\bar{X}_g$ is the sample mean for group $g$ and $\bar{X}$ is the overall sample mean. Recall that $W$ is nonsingular when $N \gg p$ and therefore, we can define the sample canonical vectors $V$ as $G - 1$ eigenvectors corresponding to nonzero eigenvalues of $W^{-1}B$ (Mardia, Kent, and Bibby 1979, chap. 11.5). Similarly to (1), the sample classification rule $\hat{h}_V(x)$ is defined as

$$\hat{h}_V(x) = \arg \min_{1 \leq g \leq G} \left\{ (x - \bar{X}_g)^\top V (V^\top W V)^{-1} V^\top (x - \bar{X}_g) - 2 \log \frac{n_g}{N} \right\}. \quad (2)$$

Further, we show that canonical vectors can be expressed in a closed form up to an orthogonal transformation. For this purpose, we establish the connection between the eigenspaces.
of matrices \( \Sigma_W^{-1} \Sigma_B \) and \( (\Sigma_W + \Sigma_B)^{-1} \Sigma_B \), as well as derive a closed form low-rank decomposition of matrix \( \Sigma_B \) and \( B \).

**Proposition 1.** Let \( Y \) be the matrix of eigenvectors corresponding to nonzero eigenvalues of \( (\Sigma_W + \rho \Sigma_B)^{-1} \rho \Sigma_B \) for some positive \( \rho \). Then \( Y \) is also the matrix of eigenvectors corresponding to nonzero eigenvalues of \( \Sigma_W^{-1} \Sigma_B \).

**Proposition 2.** The following decompositions hold: \( \Sigma_B = \Delta \Delta^\top \) and \( B = DD^\top \), where for \( r = 1, \ldots, G - 1 \) the \( r \)th column of \( \Delta \) has the form

\[
\Delta_r = \sqrt{\pi_{r+1}} \left( \sum_{i=1}^{G-1} \pi_i (\mu_i - \mu_{r+1}) \right) \sqrt{\sum_{i=1}^{r} \pi_i \sum_{i=r+1}^{G-1} \pi_i}
\]

and the \( r \)th column of \( \Delta \) has the form

\[
D_r = \sqrt{n_{r+1}} \left( \sum_{i=1}^{G-1} n_i (\tilde{x}_i - \tilde{x}_{r+1}) \right) \sqrt{\sum_{i=1}^{r} n_i \sum_{i=r+1}^{G-1} n_i}
\]

The low-rank decomposition of matrices \( \Sigma_B \) and \( B \) is not unique. Our choice of \( \Delta \) and \( D \) in Proposition 2 is motivated by the fact that these matrices can be expressed in a closed form (unlike the eigenvectors of \( \Sigma_B \) and \( B \)) and have intuitive interpretation in terms of the differences between the group means. Specifically, the columns of \( \Delta \) and \( D \) define orthogonal contrasts between the means of \( G \) groups. In the case \( G = 2 \), \( \Delta = \sqrt{\pi_1 \pi_2} (\mu_1 - \mu_2) \) and \( D = \sqrt{n_1 n_2} (\tilde{x}_1 - \tilde{x}_2) \).

We use Propositions 1 and 2 to derive the explicit form of the matrix of eigenvectors of \( \Sigma_W^{-1} \Sigma_B \).

**Proposition 3.** Define \( \Delta \) as in (3). There exist matrices \( P_1, P_2 \in \mathbb{O}^{G-1} \) such that \( \Sigma_W^{-1} \Delta P_1 \) and \( (\Sigma_W + \Sigma_B)^{-1} \Delta P_2 \) are matrices of eigenvectors of \( \Sigma_W^{-1} \Sigma_B \) corresponding to nonzero eigenvalues.

Both \( \Sigma_W^{-1} \Delta P_1 \) and \( (\Sigma_W + \Sigma_B)^{-1} \Delta P_2 \) satisfy the definition of the matrix of eigenvectors of \( \Sigma_W^{-1} \Sigma_B \) since the eigenvectors are only unique up to normalization. We discuss the advantages of using \( (\Sigma_W + \Sigma_B)^{-1} \Delta P_2 \) over \( \Sigma_W^{-1} \Delta P_1 \) in Section 2.3.

Finally, we show that the orthogonal transformation has no effect on the classification rule.

**Proposition 4.** For any matrix \( R \in \mathbb{O}^{G-1} \), the classification rule based on \( V \) is the same as the classification rule based on \( VR: h_V(x) = h_{VR}(x) \) and \( \hat{h}_V(x) = \hat{h}_{VR}(x) \) for all \( x \in \mathbb{R}^p \).

### 2.3 Proposed Estimation Criterion

From Section 2.2 it follows that for classification and variable selection it is sufficient to estimate the matrix of eigenvectors of \( \Sigma_W^{-1} \Sigma_B \) up to orthogonal transformation. Proposition 3 gives two possible population objectives: \( \Psi = \Sigma_W^{-1} \Delta \) and \( \Psi’ = (\Sigma_W + \Sigma_B)^{-1} \Delta \).

First, we consider \( \Psi = \Sigma_W^{-1} \Delta \) with the goal of choosing a suitable loss function to capture the deviations of the estimator from the target.

By definition \( \Psi \)

\[
\Psi = \arg \min_{V \in \mathbb{R}^{p \times (G-1)}} \frac{1}{2} \left\| \Sigma_W^{1/2} V - \Sigma_W^{-1/2} \Delta \right\|_F^2
\]

\[
= \arg \min_{V \in \mathbb{R}^{p \times (G-1)}} \frac{1}{2} \text{Tr} \left( V^\top \Sigma_W V - 2 \Delta^\top V \right).
\]

In the two-group case, \( V \) is a vector and the objective function in (5) reduces to

\[
\frac{1}{2} \left( \Sigma_W^{1/2} V - \Sigma_W^{-1/2} \Delta \right)^\top \left( \Sigma_W^{1/2} V - \Sigma_W^{-1/2} \Delta \right)
\]

\[
= \frac{1}{2} \left( V - \Sigma_W^{-1} \Delta \right)^\top \Sigma_W \left( V - \Sigma_W^{-1} \Delta \right)
\]

\[
= \frac{1}{2} \left( V - \Psi \right)^\top \Sigma_W \left( V - \Psi \right).
\]

This objective function is the same as the quadratic loss function considered by Rukhin (1992), who observed that it is invariant with respect to linear transformation of the data. Hence, we can define an estimator \( \hat{V} \) by substituting \( \Sigma_W \) and \( \Delta \) with \( W \) and \( D \):

\[
\hat{V} = \arg \min_{V \in \mathbb{R}^{p \times (G-1)}} \frac{1}{2} \text{Tr} \left( V^\top W V - 2 D^\top V \right).
\]

Unfortunately, the objective function in (6) is unbounded when \( W \) is singular due to the existence of nonzero \( \hat{V} \) with

\[
\text{Tr} (V^\top W V) = 0 \quad \text{and} \quad \text{Tr} (D^\top V) > 0.
\]

A simple solution is to use \( W = W + \rho I \) instead of \( W \), which is a common regularization in the LDA context (Friedman 1989; Guo, Hastie, and Tibshirani 2007; Cai and Liu 2011), and leads to

\[
\hat{V}(\rho) = \arg \min_{V \in \mathbb{R}^{p \times (G-1)}} \frac{1}{2} \text{Tr} \left( V^\top W V + \rho \|V - D\|_F^2 \right).
\]

The second component of the objective function encourages \( \hat{V}(\rho) \) to be close to \( D \), especially when \( \rho \) is large. In contrast, \( \hat{V}(\rho) \) should be close to \( W^{-1} D \) according to Proposition 3. This discrepancy suggests that strong regularization of \( W \) may have a negative affect on classification performance.

Consider now the second population objective from Proposition 3, \( \Psi’ = (\Sigma_W + \Sigma_B)^{-1} \Delta \). Following the same arguments as with \( \Psi = \Sigma_W^{-1} \Delta \) leads to

\[
\hat{V} = \arg \min_{V \in \mathbb{R}^{p \times (G-1)}} \frac{1}{2} \text{Tr} \left( V^\top W V + \frac{1}{2} \|D^\top V - I\|_F^2 \right).
\]

The objective function in (7) is bounded from below even when \( W \) is singular, hence, no additional regularization of \( W \) is needed. For this reason, we choose \((\Sigma_W + \Sigma_B)^{-1} \Delta \) as the population quantity for our estimation procedure.

Our next goal is to perform a variable selection, which in discriminant analysis corresponds to having zeros in the matrix of canonical vectors. An \( \ell_1 \) penalty \( \|V\|_1 = \sum_{i=1}^p \sum_{j=1}^{G-1} |v_{ij}| \) is commonly used for variable selection as it induces the element-wise sparsity. We are not using this penalty for two reasons. First, the element-wise sparsity leads to variable selection within each canonical vector, however, the number of variables used by all vectors may be very large. Secondly, the element-wise sparsity is not preserved under orthogonal transformation, leading to different sparsity patterns in \((\Sigma_W + \Sigma_B)^{-1} \Delta \) and \((\Sigma_W + \Sigma_B)^{-1} \Delta P_2 \).

Instead of an \( \ell_1 \) penalty, we consider the row-wise \( \ell_2 \) penalty \( \sum_{i=1}^p \|v_i\|_2 \) which induces row-sparsity. Unlike the element-wise sparsity, the row-sparsity eliminates the variables from all canonical vectors and is preserved under the orthogonal transformation. Alternative penalties include group SCAD and group MCP, we refer the reader to Huang, Brebeny, and Ma (2012) for an overview. Our choice of \( \sum_{i=1}^p \|v_i\|_2 \) is motivated by the fact that it preserves convexity of the underlying optimization.
problem. Combining (7) with this penalty suggests an estimator
\( \hat{V}(\lambda) \), defined as

\[
\hat{V}(\lambda) = \arg \min_{V \in \mathbb{R}^{p \times (G-1)}} \left\{ \frac{1}{2} \operatorname{Tr}(V^\top W V) + \frac{1}{2} \|D^\top V - I\|_F^2 + \lambda \sum_{i=1}^p \|v_i\|_2 \right\},
\]

(8)

where the objective function is convex and bounded below by zero. When \( W \) is nonsingular and \( \lambda = 0 \), \( \hat{V} = (W + B)^{-1} D \), which according to Proposition 3 is the matrix of sample canonical vectors up to an orthogonal transformation. The three components of the objective function in (8) attempt to minimize the within-group variability, control the level of the between-group variability, and provide regularization by inducing sparsity, respectively.

### 2.4 Connection with Other Sparse Discriminant Analysis Methods When \( G = 2 \)

The motivation for our method is based on the eigenstructure of the discriminant analysis problem in the multi-group setting, however, it has a direct connection with the two-group methods previously proposed in the literature. When \( G = 2 \), \( V \) is a vector in \( \mathbb{R}^p \), and (8) takes the form

\[
\hat{V}(\lambda) = \arg \min_{V \in \mathbb{R}^p} \left\{ \frac{1}{2} \operatorname{Tr}(V^\top W V) + \frac{1}{2} (D^\top V - 1)^\top (D^\top V - 1) + \lambda \|v\|_2 \right\}.
\]

**Proposition 5.** Consider \( \hat{V}_{\text{DSDA}}(\lambda) \) (Mai, Zou, and Yuan 2012), defined as

\[
\hat{V}_{\text{DSDA}}(\lambda) = \arg \min_{\beta \in \mathbb{R}^p, V \in \mathbb{R}^p} \frac{1}{2N} \sum_{i=1}^N (y_i - \beta_0 - X_i^\top V)^2 + \lambda \|V\|_2,
\]

where \( y_i = -\frac{\Sigma}{n_1} \) if the \( i \)th subject is in group 1 and \( y_i = \frac{\Sigma}{n_2} \) otherwise. Then

\[
\hat{V}(\lambda) = \frac{N}{\sqrt{n_1n_2}} \hat{V}_{\text{DSDA}} \left( \frac{N}{\sqrt{n_1n_2}} \lambda \right).
\]

Furthermore, Mai and Zou (2013) show an equivalence between the three methods for sparse discriminant analysis in the two-group setting: Wu et al. (2009), Clemmensen et al. (2011), and Mai, Zou, and Yuan (2012). It follows that our method belongs to the same class, however, it can be applied to any number of groups. Thus, it can be viewed as a multi-group generalization of this class of methods.

The optimization problem in (8) corresponds to the choice of \( \rho = 1 \) in Proposition 1. In general, any \( \rho > 0 \) leads to

\[
\hat{V}(\lambda, \rho) = \arg \min_{V \in \mathbb{R}^{p \times (G-1)}} \left\{ \frac{1}{2} \operatorname{Tr}(V^\top W V) + \frac{\rho}{2} \|D^\top V - I\|_F^2 + \lambda \sum_{i=1}^p \|v_i\|_2 \right\}.
\]

(9)

When \( \rho \to \infty \), (9) is equivalent to

\[
\hat{V}(\lambda, \rho = \infty) = \arg \min_{D^\top V = 1} \left\{ \frac{1}{2} \operatorname{Tr}(V^\top W V) + \lambda \sum_{i=1}^p \|v_i\|_2 \right\},
\]

(10)

hence, the optimization problem (9) can be considered a convex relaxation to (10) for large values of \( \rho \). When \( G = 2 \), the optimization problem (9) is equivalent to the proposal of Fan et al. (2012), who also observe the connection between (9) and (10). They perform a simulation study to assess the effect of the tuning parameter \( \rho \) and note that its value doesn’t significantly affect the classification results as long as the best \( \lambda \) is chosen for each \( \rho \). They keep the value of \( \rho \) at a fixed level \( \rho = 10 \).

### 2.5 Optimization Algorithm

The optimization problem in (8) can be solved efficiently using a block-coordinate descent algorithm. The convexity of the problem guarantees the convergence to the global optimum from any initial starting value. We refer the reader to Bach et al. (2011) for the overview of convex optimization with sparsity-inducing norms, including alternative algorithms like proximal gradient and interior-point methods. We chose to use the block-coordinate descent algorithm as it takes advantage of warm starts when solving for a range of tuning parameters and is one of the fastest algorithms for smooth losses with separable regularizers (Bach et al. 2011; Qin, Scheinberg, and Goldfarb 2013). Define the usual sample covariance matrix as

\[
T = W + B = W + DD^\top.
\]

(11)

By convexity, the solution to (8) satisfies the KKT conditions (Boyd and Vandenberghe 2004, chap. 5.5). Differentiating (8) with respect to the \((G-1) \times 1\) vector \( v_j \) formed by the \( j \)th row of \( V \) leads to

\[
V^\top T_j - d_j + \lambda u_j = 0,
\]

(12)

where \( T_j \) is the \( j \)th column of matrix \( T \) in (11), \( d_j \) is a \((G-1) \times 1\) vector formed by the \( j \)th row of matrix \( D \) in (4), and \( u_j \) is the subgradient of \( \|v_j\|_2 \):

\[
u_j = \begin{cases} u_j / \|u_j\|_2, & \text{if } u_j \neq 0; \\ \in \{u : \|u\|_2 \leq 1\}, & \text{if } u_j = 0. \end{cases}
\]

Solving (12) with respect to \( v_j \) leads to \( v_j = (d_j - \sum_{i \neq j} t_{ij} v_i - \lambda u_i) / t_{jj} \), where \( t_{ij} \) are the elements of matrix \( T \). This leads to the block-coordinate descent algorithm.

**Algorithm 1** Block-coordinate descent algorithm.

Given: \( k = 1, V^{(0)}, \epsilon \)

repeat

\[
\begin{align*}
\bar{V} & \leftarrow V^{(k-1)} \\
\text{for } j = 1 \text{ to } p & \text{ do} \\
\quad v_j^{(k)} & \leftarrow \left(1 - \frac{\lambda}{\|d_j - \sum_{i \neq j} t_{ij} v_i\|_2} \right) + (d_j - \sum_{i \neq j} t_{ij} \bar{v}_i) / t_{jj} \\
\text{end for} \\
k & \leftarrow k + 1
\end{align*}
\]

until \( k = k_{\text{max}} \) or \( V^{(k)} \) satisfies \( \max_{j} |v_{ij}^{(k)} - v_{ij}^{(k-1)}| < \epsilon \)

More implementation details, including the choice of \( V^{(0)} \) and tuning parameter \( \lambda \), are described in Section 4.3.

If \( T \) is nonsingular, by applying the vectorization operator (8) can be rewritten as

\[
\hat{V}(\lambda) = \arg \min_{V \in \mathbb{R}^{p \times (G-1)}} \frac{1}{2} \|\text{vec}(D^\top T^{-1/2})\|
\]
This formulation corresponds to a group lasso optimization problem (Yuan and Lin 2006) with the response vector \( (Y) \) and the design matrix \( X \). Due to the form of the design matrix, each block subproblem can be solved in a closed form, making the implementation of block-coordinate descent algorithm straightforward.

3. Theoretical Guarantees

In this section, we analyze the variable selection and classification performance of the estimator \( \hat{\lambda} \) defined in (8). In Section 2.4 we established an equivalence between our proposal and the proposal of Mai, Zou, and Yuan (2012) for the two-group case. We use this connection to extend the variable selection consistency results of Mai, Zou, and Yuan (2012) to the multi-group case. In particular, to prove Theorem 1, we derive Lemmas A.2 and A.3 that serve as a multi-group version of Lemma A1 in Mai, Zou, and Yuan (2012). We also show that the variable selection consistency implies classification consistency.

Let \( \Sigma = \Sigma_W + \Sigma_R \) and \( \Sigma = \Sigma_W \). Then, \( \lambda \) is the support of \( \Sigma \) with \( \lambda \geq 1 \), where \( \lambda \) is the \( j \)th row of \( \Sigma \).

Denote the support of \( \hat{\lambda} \) by \( \hat{\lambda} = \{ j : \parallel \hat{\lambda}_j \parallel_2 \neq 0 \} \). Furthermore, let \( s = \text{card}(\lambda) \), \( \phi = \sum_{\lambda} \parallel \hat{\lambda} \parallel_2 \), \( \delta = \parallel \hat{\lambda} \parallel_2 \), and \( \phi = \sum_{\lambda} \parallel \hat{\lambda} \parallel_2 \), where \( \lambda \) is the submatrix of \( \Sigma \) formed by the intersection of the rows and columns in \( \lambda \). In Theorem 1 we establish lower bounds on \( P(\lambda = \hat{\lambda}) \) and \( P \left( \parallel \hat{\lambda} \parallel_2 \leq 2 \phi \right) \).

\[ \text{Theorem 1. Assume } \kappa < 1 \text{ and } (X_i|Y_i = g) \sim N(\mu_g, \Sigma_W), i = 1, \ldots, N. \]

1. For any \( \lambda > 0 \) and positive \( \epsilon = \frac{\lambda(1-\epsilon)}{\epsilon(1-\epsilon)}, \)
\[ \hat{\lambda} = 0 \text{ with a probability of at least } 1 - t_1, \]
where \( t_1 = c_1 \parallel \psi^0 \parallel_2 \). For any \( \lambda < \phi \) and \( \epsilon < \frac{\phi - \phi_0}{\phi - \phi_0} \) none of the elements of \( \hat{\lambda} \) are zero with a probability of at least \( 1 - t_2 \), where \( t_2 = c_1 \parallel \psi^0 \parallel_2 \).

2. For any \( \lambda > 0 \) and positive \( \epsilon = \frac{\lambda(1-\epsilon)}{\epsilon(1-\epsilon)}, \)
\[ P \left( \parallel \hat{\lambda} \parallel_2 \leq 2 \phi \right) \geq 1 - c_1 \parallel \psi^0 \parallel_2 \lambda \]

Although the motivation for the proposed optimization problem doesn't rely on the normality assumption, the normality assumption does simplify the proof. We discuss possible extensions to the nonnormal case in Supplement. We further use Theorem 1 to establish variable selection consistency of the estimator \( \hat{\lambda} \) defined in (8). Specifically, Theorem 1 implies the asymptotic conditions under which \( P(\lambda = \hat{\lambda}) \rightarrow 1 \), which coincide with asymptotic conditions for the two-group case (Mai, Zou, and Yuan 2012):

\[ (C1) N \rightarrow \infty, p \rightarrow \infty, G = O(1) \text{ and } \frac{\log(p\lambda)}{N} \rightarrow 0. \]

\[ (C2) \sqrt{\frac{\log(p\lambda)}{N}} < \lambda < \Psi_{\min}. \]

Corollary 1. If (C1) and (C2) hold, then \( P(\lambda = \hat{\lambda}) \rightarrow 1. \)

We also show that under the same asymptotic conditions the same classification rule based on \( \hat{\lambda} \) coincides with the population classification rule \( h_{\psi} \) defined in (1). Let \( X \in \mathbb{R}^p \) be a new observation with a value \( x \in \mathbb{R}^p \).

Corollary 2. If (C1) and (C2) hold, then \( P(\parallel \hat{\lambda} \parallel_2 \leq 2 \phi) \rightarrow \lambda \rightarrow 0 \), and \( P \left( \hat{\lambda}(x) = h_{\psi}(x) \right) \rightarrow 1. \)

4. Simulation Results

In this section, we evaluate the performance of the estimator \( \hat{\lambda} \) defined in (8) against the alternative methods proposed in the literature. We refer to our proposal as MGSDA for multigroup sparse discriminant analysis. The test datasets are the same size as the training datasets and are generated independently. The following structures for \( \Sigma_W \) are considered in all the simulations:

1. Identity: \( \Sigma_W = I \).
2. Autoregressive: \( \Sigma_W = (\Sigma_{ij})_{p \times p} \text{ with } \Sigma_{ij} = 0.8^{\mid i-j \mid} \) for \( 1 \leq i, j \leq p \).
3. Data-based: \( \Sigma_W = (1 - \alpha)S + \alpha I \), \( \alpha = 0.01 \) and \( S \) is a sample correlation matrix estimated from the most variable \( p = 800 \) features of Ramaswamy dataset (Ramaswamy et al. 2001). The dataset is available from http://www-stat.stanford.edu/tibs/ElemStatLearn/.

Structures 1–2 have been used in simulation studies in LDA literature (Cai and Liu 2011; Witten and Tibshirani 2011; Mai, Zou, and Yuan 2012). We view the data-based structure as an approximation to a covariance structure that is more realistic in practical settings.

Simulations for other covariance structures, as well as the cases when \( p < n = 3 \), are in Supplement.

4.1 The Two-Group Case

This simulation scenario considers the classification between the two groups with \( \lambda_1 = 0 \) and \( \lambda_2 = (110, 0, p-10) \) for covariance structures 1–2. For the data-based structure, we take \( \mu_2 = (d_{10}, 0, p-10) \) with \( d \) ranging from 0.1 to 0.5 since in this case the Bayes error is almost zero for \( \mu_2 = (110, 0, p-10) \). The sample size for each group is \( n = 100 \). Conditional on group \( g \), the samples are drawn independently from the multivariate normal distribution \( N(\mu_g, \Sigma_W) \).

Mai, Zou, and Yuan (2012) performed extensive simulations to compare their proposal with the methods of Wu et al. (2009), Witten and Tibshirani (2011), Tibshirani et al. (2003), and Fan and Fan (2008). In all the settings, the method of Mai, Zou, and Yuan (2012) performs the best in terms of misclassification error. Given Proposition 5, we do not compare MGSDA with any of these methods. On the other hand, Cai and Liu (2011) also show that their proposal performs the best when
compared to Shao et al. (2011), Fan and Fan (2008), and Tibshirani et al. (2003). To our knowledge, no comparison was performed between the methods of Mai, Zou, and Yuan (2012) and Cai and Liu (2011), therefore, in this section we compare our results to the results of Cai and Liu (2011). We follow the terminology of Cai and Liu (2011) and refer to their method as linear programming discriminant (LPD). We also evaluate the performance of $\tilde{\Psi} = \sum_{i=1}^{p} \Delta$. We refer to $\tilde{\Psi}$ as the Oracle. We note that the LPD requires additional regularization of the within-group sample covariance matrix: $\tilde{W} = W + \rho I$. This regularization is needed to generate a feasible starting point for the optimization algorithm. Cai and Liu (2011) suggested taking $\rho \leq \sqrt{\log p/N}$. In our simulations $N = 200$ and therefore, $\rho = 0.15$ satisfies this requirement for both $p = 100$ and $p = 800$. We also try $\rho = 2$ to examine how the choice of $\rho$ affects the misclassification rate.

The misclassification error rates are reported in Table 1 and the corresponding number of selected features in Table 2. The methods have similar error rates, with MGSDA performing better on the Autoregressive covariance structure. They also select comparable numbers of features in all scenarios. Note that the matrix of population canonical vector $\Psi$ is truly sparse only in the Identity case, it is only approximately sparse in other scenarios.

The mean misclassification rates for the Data Based covariance structure are reported in Figure 1. In this case MGSDA performs significantly better than LPD regardless of the choice of $\rho$. The difference in misclassification rates is especially noticeable when the difference in means $d$ is small. The error rates of LPD with $\rho = 0.15$ and $\rho = 2$ are comparable except for the data-based structure. This suggests that the choice of $\rho$ can significantly affect the performance of the LPD, with smaller values of $\rho$ likely to result in smaller misclassification error. Unfortunately, it remains unclear how to choose the optimal $\rho$ in practical settings.

### 4.2 The Multi-Group Case

This simulation scenario considers the classification between the five groups with $\mu_1 = 0_p$, $\mu_2 = (d_5, -d_5, 0_{p-10})$, $\mu_3 = (-d_5, d_5, 0_{p-10})$, $\mu_4 = (d_{10}, 0_{p-10})$, and $\mu_5 = (d, -d, d, -d, \ldots, 0_{p-10})$ with $d = 1.5$ for Identity and Autoregressive covariance structures, and $d = 0.3$ for Data Based covariance structure. The sample size for each group is $n = 50$. We consider both multivariate normal and multivariate $t$ distribution with 5 degrees of freedom.

The LPD method of Cai and Liu (2011) is developed for the two-group setting. Though it can be generalized to the multi-group case, this generalization is not unique. Among the popular methods are “one-versus-one” and “one-versus-all” approaches (Hastie, Tibshirani, and Friedman 2009, p. 658). In addition to requiring the computation of a larger number of discriminant vectors ($G(G-1)/2$ and $G$ correspondingly), these approaches can disagree in their classification rules, as well as, in selected features. Given this ambiguity, we do not compare our method to the LPD in the multi-group case.

We were able to find only two methods in the literature that specifically consider sparse discriminant analysis in the multi-group case: penalizedLDA by Witten and Tibshirani (2011) and sparseLDA by Clemmensen et al. (2011). Therefore, we compare the performance of MGSDA with these two methods.

For all three methods, the misclassification errors are higher when the data comes from multivariate $t_5$ distribution. This is not surprising since $t_5$ distribution has heavier tails than normal distribution leading to higher oracle misclassification error rate. The comparative performance of MGSDA, sparseLDA, and penalizedLDA is the same for each distribution. The misclassification error rates are reported in Table 3 and the number of selected features in Table 4.

All three methods have similar misclassification rates for Identity case. For the Autoregressive and data-based structures, penalizedLDA performs the worst. MGSDA and sparseLDA have comparable error rates for the Identity and Autoregressive structure, with MGSDA selecting a smaller number of features. Hence, MGSDA achieves the best tradeoff between the misclassification error and sparsity of the solution.

For the data-based structure, MGSDA has significantly smaller error rate than sparseLDA, however, it selects a larger number of features. Note that sparseLDA is restricted to have at most 160 features as in our simulations sparseLDA package produced errors otherwise. To make the comparison between MGSDA and sparseLDA for data-based covariance structure clearer, we restricted the range of the tuning parameter for MGSDA to select the comparable number of features. The results are reported in Table 5. Although the misclassification rate of MGSDA is worse with smaller number of features, it is still better than the misclassification rate for sparseLDA.

### 4.3 Implementation Details

The method of Cai and Liu (2011) is implemented using linprogPD function from the package CLIME from CRAN. Note that linprogPD almost never returns a sparse solution. However, all the values below the precision level should be treated as zeroes (Cai, Liu, and Luo 2011). We used the default value of $10^{-3}$ for precision. The grid for the tuning parameter is chosen from 0.01 to 0.5 by 0.01. The method of Witten and Tibshirani (2011) is implemented using the package penalizedLDA from CRAN. The grid for tuning parameter is chosen from 0 to 1 by 0.01. The method of Clemmensen et al. (2011) is implemented using the package sparseLDA from CRAN. Each canonical vector is constrained to have between 3 and $0.8n$ features. This is quite a restrictive range for tuning, however, the sparseLDA
package produced errors when we used a wider range of features. MGSDA is implemented using the R package MGSDA. The grid for the tuning parameter $\lambda_1 \leq \cdots \leq \lambda_{\text{max}}$ is chosen adaptively for each dataset with $\lambda_{\text{max}} = \max_j \|d_j\|_2$, which corresponds to zero selected features. For each $\lambda_j < \lambda_{\text{max}}$ we set $V^{(0)} = \hat{V}(\lambda_{j+1})$. For all the methods, the final tuning parameter is chosen from the respective grid through 5-fold cross-validation to minimize the error rate.

Witten and Tibshirani’s penalizedLDA has significantly faster running time than all other methods since penalizedLDA assumes that the covariance matrix has diagonal structure. This assumption results in a simplified optimization algorithm, for details we refer to Witten and Tibshirani (2011). The running time of penalizedLDA is followed by MGSDA and sparseLDA. Surprisingly, LPD has the slowest performance. We suspect that this is not due to the method itself, but due to the use of linprogPD function in its implementation. A different linear program solver is likely to result in much faster running time, however, the use of a general solver makes the method implementation less straightforward.

5. Real Data

5.1 Metabolomics Dataset

Metabolomics is the global study of all metabolites in a biological system under a given set of conditions. Metabolites are the final products of enzymes and enzyme networks whose substrates and products often cannot be deduced from genetic information and whose levels reflect the integrated product of the genome, proteome, and environment. Metabolomic readouts, thus, represent the most direct (or phenotypic) readout of a cell’s physiologic state. From a technical standpoint, analytical studies of metabolism have been historically limited to one or a limited set of metabolites. However, advances in liquid chromatography and mass spectrometry have recently made it possible to measure hundreds of metabolites and with enough biomass well over 1000, in parallel. Such technologies have, thus, opened the door to obtaining global biochemical readouts of a cell’s physiologic state and response to perturbation. Cornell researchers have developed and applied a state-of-the-art metabolomic platform to track the intrabacterial pharmacokinetic fates and pharmacodynamic actions of a given compound within Mycobacterium tuberculosis (Pethe et al. 2010; de Carvalho et al. 2010, 2011; Chakraborty et al. 2013). These studies demonstrate the highly unpredictable nature and identities of these properties even for well-studied antibiotics.

We investigate a (currently unpublished) metabolomics data obtained from Dr. Kyu Rhee, which seeks to systematically elucidate the intrabacterial pharmacokinetic and pharmacodynamic fates and actions of antimycobacterial hit or lead compound series identified in high throughput screens against replicating and non- or slowly replicating forms of Mycobacterium tuberculosis.

The data contains measurements of 171 metabolic responses of 68 patients to 25 antibiotics that are administered at different dosage levels. Each measurement is an average of three replicates, normalized to the vehicle control and log2 transformed. And 14 out of 25 antibiotics can be divided into the following 5 groups: STREP_AMI (streptomycin, amikacin), FLQ (levofloxacin, moxifloxacin), DHFR (nitrofurantoin, sri8210, sri710, sri8857), DHPS (sulfamethoxazole, sulfisoxazole, aps), and InhA (ethambutol, isoxyl, gsk93). These antibiotics are administered to 35 patients out of 68. In the subsequent analysis, we only focus on these 5 groups of antibiotics and do not consider the dosage levels.

We compare the performance of MGSDA, penalizedLDA (Witten and Tibshirani 2011), and sparseLDA (Clemmensen et al. 2011) on this dataset using the following measures: the mean number of mismatched samples and the mean number of selected features over 100 replications of 5-fold cross-validation. We do not perform random splits into the training and test set due to the small sample size. The results are reported in Table 6. The results show that all three methods perform very well in terms of misclassification error; the mean number of mismatched samples is significantly less than one indicating that all three methods lead to almost perfect classification performance. Such a good performance suggests that there is a significant difference in the metabolic responses between the 5 groups of antibiotics. However, penalizedLDA achieves this performance by selecting almost all of the metabolites, whereas MGSDA and sparseLDA use less than 20% of the original features. Note that there is a substantial variation between the replications due to the small sample size of the data.

We further estimate four canonical vectors using MGSDA with $\lambda = 0.57$ and illustrate the projected data in Figure 2. Note that 8 selected metabolites provide perfect linear separation between the groups. $\lambda = 0.57$ is chosen as one of the hundred tuning parameters from above replications of cross-validation splits. We have tried the other values of $\lambda$ as well, however, they all provided perfect linear separation between the groups with projected data being very similar to Figure 2. Though there is a variation between the cross-validation replications due to the small sample size of the data, this variation has negligible effect on the final projection.
Table 3. Mean misclassification error rates as percentages over 100 replications, $G = 5$, standard deviation is given in parentheses.

| Covariance  | $p$ | $t_5$ distribution | Normal distribution |
|-------------|-----|--------------------|---------------------|
|             |     | penLDA | sparseLDA | MGSDA | penLDA | sparseLDA | MGSDA |
| Identity    | 100 | 6.39 (1.35) | 7.2 (1.6) | 6.47 (1.62) | 1.93 (0.91) | 2.45 (1.18) | 1.89 (0.84) |
|             | 800 | 6.34 (1.8)  | 7.84 (1.83) | 6.78 (1.58) | 1.7 (0.85)  | 2.98 (1.21) | 2.06 (0.94) |
| Autoregressive | 100 | 16.02 (2.49) | 8.76 (2.01) | 9.04 (2.02) | 11.14 (2.21) | 5.11 (1.56) | 5.63 (1.55) |
|             | 800 | 15.51 (2.25) | 10.41 (2.08) | 10.84 (1.89) | 10.72 (1.98) | 6.42 (1.82) | 7.13 (1.78) |
| Data Based  | 100 | 72.89 (4.09) | 20.34 (3.04) | 15.64 (2.64) | 70.47 (5.39) | 14.24 (2.73) | 9.33 (2.21) |
|             | 800 | 78.79 (2.2)  | 37.44 (5.39) | 7.54 (1.76)  | 78.7 (2.36)  | 26.47 (4.06) | 3.4 (1.45)  |

Table 4. Mean number of selected features over 100 replications, $G = 5$, standard deviation is given in parentheses.

| Covariance  | $p$ | $t_5$ distribution | Normal distribution |
|-------------|-----|--------------------|---------------------|
|             |     | penLDA | sparseLDA | MGSDA | penLDA | sparseLDA | MGSDA |
| Identity    | 100 | 24 (24) | 36 (20) | 16 (11) | 34 (32) | 31 (19) | 13 (10) |
|             | 800 | 41 (39) | 47 (32) | 16 (14) | 26 (29) | 33 (23) | 11 (3)  |
| Autoregressive | 100 | 12 (9)  | 41 (15) | 29 (18) | 15 (15) | 43 (17) | 31 (20) |
|             | 800 | 25 (26) | 70 (43) | 29 (33) | 19 (26) | 57 (34) | 15 (11) |
| Data Based  | 100 | 59 (33) | 80 (4)  | 93 (8)  | 48 (34) | 80 (5)  | 92 (9)  |
|             | 800 | 458 (340) | 139 (6)* | 330 (29)* | 411 (354) | 138 (5)* | 302 (29)* |

Figure 2. Metabolomics dataset projected onto 4 column vectors of $V$, $k = 8$ metabolite features are used.

Table 5. Comparison of MGSDA and sparseLDA on Data Based covariance structure, $G = 5$ and $p = 800$, the tuning parameter for MGSDA is restricted to allow for comparable number of features with sparseLDA, standard deviation is given in parentheses.

| $t_5$ distribution | Normal distribution |
|--------------------|---------------------|
| sparseLDA | MGSDA | sparseLDA | MGSDA |
| Error | 37.44 (5.39) | 23.75 (4.15) | 26.47 (4.06) | 16.46 (3.72) |
| Features | 139 (6) | 152 (14) | 138 (5) | 143 (13) |

5.2 14 Cancer Dataset

In this section, we compare the performance of MGSDA, penalizedLDA (Witten and Tibshirani 2011), and sparseLDA (Clemmensen et al. 2011) on the 14 cancer dataset by Ramaswamy et al. (2001). This dataset contains 16063 gene expression measurements collected on 198 samples. Each sample belongs to one of the 14 cancer classes. The dataset can be obtained from http://statweb.stanford.edu/tibs/ElemStatLearn/.

We selected this dataset as it is publicly available and has been previously analyzed by a number of authors including Witten and Tibshirani (2011).

Following the recommendation of Hastie, Tibshirani, and Friedman (2009, p. 654), we first standardize the data to have mean zero and standard deviation one for each patient. To reduce the overall computational cost, we restrict the analysis to 3000 genes. We select these genes following the novel model-free feature screening procedure for discriminant analysis of Cui,
Li, and Zhong (2014). Following the approach taken by Witten and Tibshirani (2011), we perform 100 independent splits of the dataset into the training set containing 75% of the samples and the test set containing 25% of the samples. The tuning parameter for all methods is selected using 5-fold cross-validation on the training set. The mean number of misclassified samples on the test set and the mean number of selected features over 100 splits are reported in Table 7. MGSDA and sparseLDA perform better than penalizedLDA in terms of the misclassification error and select much smaller number of features. MGSDA and sparseLDA select comparable number of features, with the misclassification error of MGSDA being the smallest.

6. Discussion

This article introduces a novel procedure that estimates population canonical vectors in the multi-group setting, and a corresponding R package MGSDA is available on CRAN. The proposed method is a natural generalization of the two-group methods that were previously studied in the literature. In addition to being computationally tractable, the method performs feature selection which results in sparse canonical vectors. The group penalty eliminates features from all canonical vectors at once with the remaining nonzero features being the same for all the vectors.

One possible extension of the proposed method is to allow canonical vectors to have different sparsity patterns. This goal is achievable through the addition of the within-group penalty term to the objective function (8). Such an estimation procedure has already been considered in the regression context; for example, Simon et al. (2013) propose the following optimization problem:

$$
\hat{\beta}(\lambda, \alpha) = \arg \min_{\beta \in \mathbb{R}^{p \times (g-1)}} \left\{ \frac{1}{2n} \|Y - X\beta\|^2_2 + (1 - \alpha)\lambda \sum_{g=1}^G \sqrt{p_g}\|\beta^{(g)}\|_2^2 + \alpha\lambda\|\beta\|_1 \right\},
$$

where $$p_g$$ is the size of group $$g$$. In our case this approach results in

$$
\hat{V}(\lambda, \alpha) = \arg \min_{V \in \mathbb{R}^{p \times (G-1)}} \left\{ \frac{1}{2} \text{Tr}(V^TWV) + \frac{1}{2}\|D^TV - I\|^2_F + (1 - \alpha)\lambda \sum_{i=1}^p \|v_i\|_2 + \alpha\lambda\|V\|_F \right\}.
$$

Another possible extension is to perform canonical vectors selection in addition to feature selection, which will enhance the interpretability, especially when the number of groups $$G$$ is large. This goal can be achieved through the addition of the nuclear norm penalty term to the objective function (8):

$$
\hat{V}(\lambda, \alpha) = \arg \min_{V \in \mathbb{R}^{p \times (G-1)}} \left\{ \frac{1}{2} \text{Tr}(V^TWV) + \frac{1}{2}\|D^TV - I\|^2_F + \lambda \sum_{i=1}^p \|v_i\|_2 + \alpha\|V\|_F \right\}.
$$

Depending on the value of $$\alpha > 0$$, the resulting matrix $$\hat{V}$$ has rank that is less than $$G - 1$$, effectively resulting in a lower-dimensional eigenspace.

Both extensions result in convex optimization problems, but require additional modifications to the optimization algorithm. An interesting direction for future research is to examine how these extensions compare to the original method in different scenarios.

An underlying assumption of MGSDA is the equality of covariance structures between groups. A small simulation study of the effect of covariance misspecification is given in Supplement. It indicates that MGSDA is robust with respect to this assumption, outperforming the traditional quadratic discriminant analysis (QDA) in the $$p < n$$ setting. This result is supported by previous research, which demonstrated the superiority of LDA over QDA for small samples and moderate values of $$p$$ (see Chapter 6.3.2 in Seber (1984) and the references therein).

It is of interest to investigate whether a direct estimation procedure can be applied to QDA to improve its performance in high-dimensional settings.

We established the variable selection and classification consistency of proposed estimator in the regime where $$\frac{\log(p)s^2}{N} \to 0$$. Although preparing this manuscript, we became aware of the work of Kolar and Liu (2013), who show variable selection consistency of the sparse discriminant analysis under the conditions that $$G = 2$$ and $$N \geq C\log((p - s)\log(N))$$ for some constant $$C > 0$$. These improved rates directly apply to our proposal in the case $$G = 2$$, however, the extension of these results to the case $$G > 2$$ is not clear. This is another direction for future research.

### Appendix A

#### A.1 Proof of Proposition 1

**Proof.** From the definition of $$\Upsilon_{\rho}$$, $$(\Sigma_W + \rho\Sigma_B)^{-1}\rho\Sigma_B\Upsilon_{\rho} = \Upsilon_{\rho}A$$. It follows that

$$
\rho\Sigma_B\Upsilon_{\rho} = \Sigma_W\Upsilon_{\rho}A + \rho\Sigma_B\Upsilon_{\rho}A;
$$

$$
\rho\Sigma_B\Upsilon_{\rho}(I - A) = \Sigma_W\Upsilon_{\rho}A;
$$

$$
\Sigma_W^{-1}\Sigma_B\Upsilon_{\rho} = \Upsilon_{\rho}\frac{1}{\rho} (I - A)^{-1}.
$$

From the last equation it follows that $$\Upsilon_{\rho}$$ is the matrix of eigenvectors of $$\Sigma_W^{-1}\Sigma_B$$.

#### A.2 Proof of Proposition 2

**Proof.** The proof is only given for matrix $$B$$, the proof for matrix $$\Sigma_B$$ is similar.

1. Consider the equal group case: $$n_1 = \ldots = n_G = n$$ and $$N = Gn$$. It follows that $$\bar{X} = \frac{1}{N}\sum_{g=1}^G X_g/G$$ and therefore $$B = \frac{1}{N}\sum_{g=1}^G n(\bar{X}_g - \bar{X})(\bar{X}_g - \bar{X})^\top = \frac{1}{N}\bar{X}^\top\left(\frac{1}{G}\mathbb{1}\right)C\left(\frac{1}{G}\mathbb{1}\right)^\top\bar{X}$$, where $$C$$ is the centering matrix and
$[\frac{1}{\sqrt{n_i}}1_i]$ is a $N \times G$ matrix formed by $G$ columns $\frac{1}{\sqrt{n_i}}1_i$ such that $(1_{ij}) = 1$ if $i$th observation belongs to the $j$th group and $(1_{ij}) = 0$ otherwise. Note that $C = H^T H$ where $H$ is the Helmert matrix of size $G$ with its first row removed (Searle 2006). Therefore $B = DD^T$, where $D = \frac{1}{\sqrt{n}}X^T [\frac{1}{\sqrt{n_i}}1_i]H^T$.

2. Consider the general case where each group has size $n_g$. Similar to the equal group case, $B = \frac{1}{\sqrt{n}}X^T [\frac{1}{\sqrt{n_i}}1_i]C(\frac{1}{\sqrt{n_i}}1_i)^T X$, where $C = I_G - \frac{1}{\sqrt{n}}KK^T$ and $K = (\frac{1}{\sqrt{n}} \ldots \frac{1}{\sqrt{n}})^T$. Next we show that similar to $C$, $C$ can be decomposed as $\bar{C} = H^T \bar{H}$ and $\bar{H}$ is a $G \times 1$ adjusted Helmert matrix. Since $\bar{H}$ satisfies $I_G - \frac{1}{\sqrt{n}}KK^T = H^T \bar{H}$, $(K, \bar{H}^T)$ is an orthogonal matrix. The $G \times 1$ orthogonal contrasts for unbalanced data have the following form (Searle 2006, p 51):

$$\delta_r = \sqrt{n_{r+1}} \left( \sum_{i=1}^{r} n_i (\tilde{X}_i - \bar{X}_{r+1}) \right).$$

Denote by $h_r$ the rows of $\bar{H}$. Then it follows that for some constant $C_r$, $h_r (\frac{1}{\sqrt{n_r}}1_r)^T X = C_r \delta_r$. This means that $h_{r+1} = C_r \sqrt{n_{r+1}} n_{r+1} \delta_{r+1}$ and $h_j = 0$ for $j > (r + 1)$. To find $C_r$, we use the fact that $h_{r+1} h_r^\top = 1$. Let $s_r = \sum_{i=1}^{r} n_i$. Then $C_r$ satisfies $C_r (\delta_r)^T = 1$, or equivalently $C_r s_r \delta_r = 1$. From the last equation $C_r = \frac{1}{\delta_r}$. Combining the results it follows that $B = DD^T$, where $D = \frac{1}{\sqrt{n}}X^T [\frac{1}{\sqrt{n_i}}1_i]H^T$ and $D_r = \frac{1}{\delta_r} C_r \delta_r = \frac{1}{\sqrt{n_{r+1}}} \sum_{i=1}^{r} n_i (\tilde{X}_i - \bar{X}_{r+1})$.

\[ \delta_r = \sqrt{n_{r+1}} \left( \sum_{i=1}^{r} n_i (\tilde{X}_i - \bar{X}_{r+1}) \right). \]

A.3 Proof of Proposition 3

**Proof.** Denote $\Psi = \Sigma_w A \Delta P$, where $P$ is an orthogonal matrix such that $\Delta^\top \Sigma^2 \Delta = P A P^\top$. It follows that $\Sigma_w = \Sigma_w A \Psi = \Sigma_w A \Sigma^\top \Sigma \Delta^\top \Sigma \Delta P = \Sigma_w A \Sigma \Delta P = \Psi A$. Hence, $\Psi$ is the matrix of eigenvectors of $\Sigma_w A$. The proof for $(\Sigma_w + \Sigma_b) - \Delta P = \Psi A$. The proof for Proposition 1.

A.4 Proof of Proposition 4

**Proof.** The proof is only given for the sample classification rule $\hat{h}_V(x)$, the proof for the population classification rule $\hat{h}_P(x)$ is analogous. Define $Z = XV$. Using $V$, a new observation $x \in \mathbb{R}^p$ is classified to group $h_v(x)$, where

$$\hat{h}_V(x) = \arg \min_{1 \leq j \leq G} \left( (V \top x - \bar{Z}_j)^\top (V \top WWV)^{-1} (V \top x - \bar{Z}_j) - 2 \log \frac{n_j}{N} \right).$$

Consider a new classification rule $\hat{h}_V(x)$ based on $V' = VR$ with $R \in \mathbb{R}^{G-1}$. Then $Z' = XV' = XV = ZR$ and

$$\hat{h}_V(x) = \arg \min_{1 \leq j \leq G} \left( (V' \top x - \bar{Z}_j)^\top (V' \top WWV')^{-1} (V' \top x - \bar{Z}_j) - 2 \log \frac{n_j}{N} \right)$$

$$= \arg \min_{1 \leq j \leq G} \left\{ (R \top V \top x - R \top \bar{Z}_j)^\top (R \top WWR)^{-1} (R \top V \top x - R \top \bar{Z}_j) - 2 \log \frac{n_j}{N} \right\}$$

$$= \arg \min_{1 \leq j \leq G} \left( (V \top x - \bar{Z}_j)^\top RR^{-1} (V \top WWV)^{-1} (V \top x - \bar{Z}_j) - 2 \log \frac{n_j}{N} \right)$$

$$= \hat{h}_V(x).$$

A.5 Proof of Proposition 5

**Proof.** By definition $D = \frac{\sqrt{np}}{N} (\tilde{X}_i - \bar{X}_j)$. Therefore

$$\hat{V}_{DSDA}(\lambda) = \arg \min_{V \in \mathbb{R}^p} \frac{1}{2} V^\top (W + DD^T) V - \frac{N}{\sqrt{n_1 n_2}} D^\top V + \lambda \|V\|_1 \leq \frac{1}{2} V^\top (W + DD^T) V - D^\top V + \frac{\lambda \sqrt{n_1 n_2}}{N} \|V\|_1.$$ 

Similarly, $\hat{V}(\lambda) = \arg \min_{V \in \mathbb{R}^p} f(V, \lambda)$, where

$$f(V, \lambda) = \frac{1}{2} V^\top (W + DD^T) V - D^\top V + \lambda \|V\|_1.$$ 

It follows that $\hat{V}(\lambda) = \frac{N}{\sqrt{n_1 n_2}} \hat{V}_{DSDA} \left( -\frac{\sqrt{n_1 n_2}}{N} \lambda \right).$

A.6 Auxiliary Lemmas for Theorem 1

**Lemma A.1.** \(\|AB\|_{\infty,2} \leq \|A\|_{\infty,2} \|B\|_{\infty,2} \).

**Lemma A.2.** Let $F = D - \Delta$. There exists constant $c_1 > 0$ such that

$$P(\|F\|_{\infty,2} \geq \epsilon) \leq 2p(G - 1) \exp(-c_1 N \epsilon^2).$$

**Lemma A.3.** Let $T = W + B$ and $\Sigma = \Sigma_w + \Sigma_b$. There exist constants $c_1 > 0$ and $c_2 > 0$ such that

$$P(\|T_{AA} - \Sigma_{AA}\|_{\infty,2} \geq \epsilon) \leq c_2 s \epsilon^2 \exp(-c_2 N \epsilon^2);$$

$$P(\|T_{AA} - \Sigma_{AA}\|_{\infty,2} \geq \epsilon) \leq c_2 s (p - s) \epsilon^2 \exp(-c_2 N \epsilon^2).$$

**Lemma A.4.** Let $F_1 = T_{AA}(T_{AA})^{-1} - \Sigma_{AA}(T_{AA})^{-1}$. There exists constant $c_1 > 0$ such that

$$P(\|F_1\|_{\infty,2} \geq \epsilon (\kappa + 1)(1 - \phi) \epsilon^{-1}) \leq 2s(G + 1) \phi s \epsilon^2 \exp(-c_2 N \epsilon^2)^2.$$

A.7 Proof of Theorem 1.

**Proof.** The proof follows the proof of Theorem 1 in Mai, Zou, and Yuan (2012) using the results of auxiliary lemmas. For simplicity of illustration, we consider the case $\pi_i = \frac{1}{p}$ and $n_i = \frac{N}{p}$. The full proof as well as extension to the general $\pi_i$ and $n_i$ are in Supplemental Materials.

Supplementary Materials

Supplementary materials include additional simulation studies and technical proofs.
Acknowledgments

We are grateful to Jacob Bien for a valuable discussion of block-coordinate descent algorithms. We thank the Editor, the Associate Editor and two anonymous referees for valuable comments and suggestions that helped to significantly improve this manuscript.

Funding

This research was partially supported by NSF-DMS 1208488, NSF-DMS 0808864, NIH grant U19 AI111143 and ASAP grant FA9550-13-1-0137.

References

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2011), Optimization for Machine Learning. Chapter: Convex Optimization with Sparsity-Inducing Norms, Cambridge, MA: MIT Press. [699]

Bickel, P. J. J., and Levina, E. (2004), “Some Theory for Fisher’s Linear Discriminant Function, ‘Naive Bayes,’ and Some Alternatives When There are Many More Variables Than Observations,” Bernoulli, 10, 989–1010. [696]

Boyd, S. P., and Vandenberghe, L. (2004), Convex Optimization, Cambridge, UK: Cambridge University Press. [699]

Cai, T., and Liu, W. (2011), “A Direct Estimation Approach to Sparse Linear Discriminant Analysis,” Journal of the American Statistical Association, 106, 1566–1577. [696,697,700,701]

Cai, T., Liu, W., and Luo, X. (2011), “A Constrained $\ell_1$ Minimization Approach to Sparse Precision Matrix Estimation,” Journal of the American Statistical Association, 106, 594–607. [701]

Chakraborty, S., Gruber, T., Barry, C. E., Boshoff, H. L., and Rhee, K. Y. (2013), “Para-Aminosalicylic Acid Acts as an Alternative Substrate of Folate Metabolism in Mycobacterium Tuberculosis,” Science, 339, 88–91. [702]

Clemmensen, L., Hastie, T., Witten, D. M., and Ersbøll, B. (2011), “Sparse Discriminant Analysis,” Technometrics, 53, 406–413. [696,697,698,701,702]

Cui, H., Li, R., and Zhong, W. (2014), “Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis,” Journal of the American Statistical Association, 110, 630–641. [704]

de Carvalho, L. P. S., Darby, C. M., Rhee, K. Y., and Nathan, C. (2011), “Nita-oxazinide Disrupts Membrane Potential and Interalacital pH Homeostasis of Mycobacterium tuberculosi,” ACS Medicinal Chemistry Letters, 2, 849–854. [702]

de Carvalho, L. P. S., Fischer, S. M., Marrero, J., Nathan, C., Ehrtr, S., and Rhee, K. Y. (2010), “Metabolomics of Mycobacterium Tuberculosis Reveals Compartmentalized Co-Catabolism of Carbon Substrates,” Chemistry and Biology, 17, 1122–1131. [702]

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002), “Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data,” Journal of the American Statistical Association, 97, 77–87. [696]

Fan, J., and Fan, Y. (2008), “High Dimensional Classification Using Features Annealed Independence Rules,” Annals of Statistics, 36, 2605–2637. [700]

Fan, J., Feng, Y., and Tong, X. (2012), “A Road to Classification in High Dimensional Space: The Regularized Optimal Affine Discriminant,” Journal of the Royal Statistical Society: Series B, 74, 745–771. [699]

Friedman, J. H. (1989), “Regularized Discriminant Analysis,” Journal of the American Statistical Association, 84, 165–175. [696,698]

Golub, G. H., and Van Loan, C. F. (2012). Matrix Computations (4th ed.), Baltimore, MD: The Johns Hopkins University Press. [697]

Guo, Y., Hastie, T., and Tibshirani, R. (2007), Regularized Linear Discriminant Analysis and Its Application in Microarrays, Biostatistics, 8, 86–100. [696,698]

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009), The Elements of Statistical Learning (2nd ed.), New York: Springer. [696,701,703]

Huang, J., Breheny, P., and Ma, S. (2012), “A Selective Review of Group Selection in High-Dimensional Models,” Statistical Science, 27, 481–499. [698]

Kolar, M., and Liu, H. (2013), “Optimal Feature Selection in High-Dimensional Discriminant Analysis,” IEEE Transactions on Information Theory, 61, 1063–1083. [704]

Krzanowski, W. J., Jonathan, P., McCarthy, W. V., and Thomas, M. R. (1995), “Discriminant Analysis With Singular Covariance Matrices: Methods and applications to Spectroscopic Data,” Journal of the Royal Statistical Society Series C, 44, 101–115. [696]

Mai, Q., and Zou, H. (2013), “A Note on the Connection and Equivalence of Three Sparse Linear Discriminant Analysis Methods,” Technometrics, 55, 243–246. [699]

Mai, Q., Zou, H., and Yuan, M. (2012), “A Direct Approach to Sparse Discriminant Analysis in Ultra-High Dimensions,” Biometrika, 99, 29–42. [696,697,700,705]

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), Multivariate Analysis, New York: Academic Press. [696,697]

McLachlan, G. J. (1992), Discriminant Analysis and Statistical Pattern Recog- nition, New York: Wiley-Interscience. [697]

Pethe, K., Sequeira, P. C., Agarwalla, S., Rhee, K., Kuhen, K., Phong, W. Y., Patel, V., Beer, D., Walker, J. R., Duraiswamy, J., Jiricek, J., Keller, T. H., Chatterjee, A., Tan, M. P., Ujjini, M., Rao, S. P., Camacho, L., Bifani, P., Mak, P. A., Ma, I., Barnes, S. W., Chen, Z., Plouffe, D., Thayalan, P., Ng, S. H., Au, M., Lee, B. H., Tan, B. H., Ravindran, S., Nanjundappa, M., Lin, X., Goh, A., Lakshminarayana, S. B., Shoen, C., Cynamon, M., Kreiswirth, B., Dartois, V., Peters, E. C., Glynne, J., Brenner, S., and Dick, T. (2010), “A Chemical Genetic Screen in Mycobacterium Tuberculosis Identifies Carbon-Source-Dependent Growth Inhibitors Devoid of in Vivo Efficacy,” Nature Communications, 1, 57. Available at http://www.nature.com/ncomms/journal/v1/n5/full/ncomms1060.html [702]

Quin, Z., Scheinberg, K., and Goldfarb, D. (2013), “Efficient Block-Coordinate Descent Algorithms for the Group Lasso,” Mathematical Programming Computation, 5, 143–169. [699]

Ramaseswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Loda, M., Lander, E. S., and Golub, T. R. (2001), “Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures,” Proceedings of the National Academy of Sciences, 98, 15149–15154. [703]

Rukhin, A. L. (1992), “Generalized Bayes Estimators of a Normal Discriminant Function,” Journal of Multivariate Analysis, 41, 154–162. [698]

Searle, S. R. (2006), Linear Models for Unbalanced Data, New York: Wiley-Interscience. [704]

Seber, G. A. F. (1984), Multivariate Observations, New York: Wiley. [704]

Shao, J., Wang, Y., Deng, X., and Wang, S. (2011), “Sparse Linear Discriminant Analysis by Thresholding for High Dimensional Data,” Annals of Statistics, 39, 1244–1265. [701]

Simon, N., Friedman, J. H., Hastie, T., and Tibshirani, R. (2013). “A Sparse Group Lasso,” Journal of Computational and Graphical Statistics, 22, 231–245. [704]

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002), “Diagnos- is of Multiple Cancer Types by Shrunken Centroids of Gene Expression,” Proceedings of the National Academy of Sciences, 99, 6567–6572. [696]

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003), “Class Prediction by Nearest Shrunken Centroids, With Applications to DNA Microarrays,” Statistical Science, 18, 104–117. [700]

Witten, D. M., and Tibshirani, R. (2011), “Penalized Classification Using Fisher’s Linear Discriminant,” Journal of the Royal Statistical Society. Series B, 73, . 753–772. [696,697,701,702,703]

Wu, M. C., Zhang, L., Wang, Z., Christianti, D. C., and Lin, X. (2009), “Sparse Linear Discriminant Analysis for Simultaneous Testing for the Significance of a Gene Set/Pathway and Gene Selection,” Bioinformatics, 25, 1145–1151. [699,700]

Xu, P., Brock, G. N., and Parrish, R. S. (2009), “Modified Linear Dis- criminant Analysis Approaches for Classification of High-Dimensional Microarray Data,” Computational Statistics and Data Analysis, 53, 1674–1687. [696]

Yuan, M., and Lin, Y. (2006), “Model Selection and Estimation in Regression With Grouped Variables,” Journal of the Royal Statistical Society, Series B, 68, 49–67. [700]