Statistica Sinica

LINEAR DISCRIMINANT ANALYSIS WITH SPARSE AND DENSE SIGNALS

Ning Wang

Center for Statistics and Data Science, Beijing Normal University,
Zhuhai, Guangdong, China, 519087

Shaokang Ren

Microsoft Corporation, Redmond, WA, USA, 98052

and Qing Mai

Department of Statistics, Florida State University, Tallahassee, FL, USA, 32306

Abstract: Many high-dimensional linear discriminant analysis (LDA) methods have been proposed in recent years. A common theme of these methods is the sparsity assumption. However, in practice the sparsity assumption may be violated and, if so, sparse methods may be inaccurate. Motivated by this challenge, we propose a novel high-dimensional LDA method that relaxes the sparsity assumption. We assume that there exist a few sparse signals with large effects, and a large number of dense signals with small effects. In parameter estimation, we combine the group Lasso penalty and the ℓ_2 penalty to automatically identify these signals. Our estimation involves a convex optimization problem that can be solved straightforwardly. Theoretical and numerical results are obtained to support the application of our proposal.

Key words and phrases: Linear discriminant analysis, group Lasso, ℓ_2 penalty, regularization

1. Introduction

Linear discriminant analysis (LDA) has received considerable attention in its generalization to high-dimensional data. Some examples include Cai and Liu (2011); Clemmensen et al. (2011); Witten and Tibshirani (2011); Fan et al. (2012); Mai et al. (2012); Xu et al. (2015); Mai et al. (2019); Yang et al. (2022). These methods preserve the elegance and simplicity of classical LDA. They have explicit probabilistic models that yield highly interpretable final classifiers and enable researchers to understand the results. In the meantime, they recruit innovations in formulation, computation and theory to tackle the high dimensionality. These methods are shown to have impressive performance in a wide range of applications as well.

However, as with many other high-dimensional methods, sparsity is at the core of most, if not all, high-dimensional LDA methods. It is often assumed that some parameters in the high-dimensional LDA model are sparse, such as the covariance matrix, precision matrix, mean differences, or discriminant coefficients. It is known that, without additional parsimony assumptions, accurate model estimation is virtually impossible in high dimensions (Bickel and Levina, 2004; Fan and Fan, 2008), and there is no doubt that sparsity is a powerful assumption for this sake. The sparsity also facilitates interpretation as only a small subset of predictors are relevant for the prediction. But it remains an open question if we can relax the sparsity assumption. For example, Witten and Tibshirani (2011) explicitly enforced sparsity in their ℓ_1 Fisher's discriminant analysis (ℓ_1-FDA) method, but in the three real datasets they considered, ℓ_1-FDA produced non-sparse classifiers with thousands of nonzero coefficients and high classification accuracy. The authors argued that this was because sparsity is often only an approximation in practice. It is interesting to know if we could accommodate such situations with a new high-dimensional LDA model and method.

Motivated by this challenge, we propose a novel high-dimensional LDA method that gives a "sparse+dense" classifier. We assume that there exists a small subset of predictors with large coefficients, while the rest have small but possibly nonzero coefficients. Thus, we relax the sparsity assumption by allowing all the coefficients to be nonzero, but to some extent preserves the interpretability that only a few variables have large impacts in the final prediction. Under this assumption, we devise an estimator that automatically identifies and estimates the sparse and dense signals through convex optimization. Numerical and theoretical evidence is further obtained to support our proposal.

Our proposal is inspired by the so-called “lava” estimator in the regression problem Chernozhukov et al. (2017). The lava estimator estimates the coefficient in a linear regression problem with the sparse+dense structure. We borrow some of their techniques in our study, but we investigate the different problem of classification, where sparse+dense estimators have not been developed to the best of our knowledge. We also tackle many distinctive challenges in LDA problems. First, in regression, we can treat the predictors as fixed or at least condition on the predictors to make inference about the response, but in LDA we directly model the distribution of the predictors and have to deal with the randomness in them. Second, in regression, it is more obvious to pick the parameter of interest and then estimate it with a variant of the least squares formula. Yet in LDA we need to carefully determine the parametrization and formula for efficiency. Third, in a linear regression model we only need to estimate one parameter of the regression coefficient, while in multiclass problems we need to estimate several different directions to separate the classes.

The rest of the article is organized as follows. We explain the proposed model and method in Section 2. The theoretical properties of our proposal are given in Section 3. In Section 4, we present the numerical studies. We further examine our method on several real datasets in Section 5. We provide proofs of lemmas and theorems in the supplementary materials.

2. Methodology

2.1 Background

Consider a pair of random variables $(Y, \mathbf{X})$, where the predictor $\mathbf{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ and the class label $Y \in \{1, \ldots, K\}$, with K being a positive integer. Linear discriminant analysis (LDA) assumes that (Hastie et al., 2009, e.g)

$$\boldsymbol{X} \mid Y = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \quad \Pr(Y = k) = \pi_k, \tag{2.1}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^p$ is the within-class mean, $\boldsymbol{\Sigma}$ is a $p \times p$ covariance matrix and π_k is the prior probability of class k.

Our goal is to predict the label of any new sample $\boldsymbol{X}^*$. It is known that, under the LDA model, we can minimize the classification error by the so-called Bayes rule that (Friedman et al., 2001; Mai et al., 2019)

$$\hat{Y} = \arg\max_k \{(\boldsymbol{X} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_k))^T \boldsymbol{\theta}_k^* + \log(\pi_k/\pi_1)\}, \tag{2.2}$$

where the linear discriminant directions are given by

$$\boldsymbol{\theta}_k^* = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1), \quad k = 2, \ldots, K. \tag{2.3}$$

Hence, the linear discriminant directions $\boldsymbol{\theta}_k^*$ are critical to the classification. They project the p-dimensional predictor $\boldsymbol{X}$ to a $K-1$ dimensional subspace that retains all the information for optimal classification. Consequently, many existing sparse LDA methods assume that $\boldsymbol{\theta}_k^*$ is sparse in that the majority of its elements are zero (Cai and Liu, 2011; Fan et al., 2012; Mai et al., 2012, 2019). In particular, in a multiclass problem where $K > 2$, we have multiple directions $\boldsymbol{\theta}_k^*$ to estimate, and a variable X_j is unimportant for classification if and only if

$$\theta_{kj}^* = 0, \quad \text{for } k = 2, \ldots, K. \tag{2.4}$$

Therefore, the sparsity assumption indicates that (2.4) holds for most j.

2.2 The “Sparse+Dense” Assumption

Our interest is to relax the sparsity assumption in the LDA model. To this end, we decompose the discriminant direction as

$$\boldsymbol{\theta}_k^* = \boldsymbol{\beta}_k^* + \boldsymbol{\delta}_k^* \tag{2.5}$$

for any k, where $\boldsymbol{\delta}_k^* = (\delta_{k1}^*, \ldots, \delta_{kp}^*)^T \in \mathbb{R}^p$ is sparse with only a few nonzero and relatively large elements, while $\boldsymbol{\beta}_k^* = (\beta_{k1}^*, \ldots, \beta_{kp}^*)^T \in \mathbb{R}^p$ have small elements. Specifically, motivated by the sparsity assumption in (2.4), for most $j \in \{1, \ldots, p\}$, we have

$$\delta_{kj}^* = 0, \quad \text{for } k = 2, \ldots, K.$$

For ease of presentation, we also refer to $\boldsymbol{\delta}_k^*$ as “sparse signals”, and $\boldsymbol{\beta}_k^*$ as the “dense signals”.

Note that the entries in $\boldsymbol{\theta}_k^*$ are coefficients in the final classifier and quantify the impact of each predictor. Hence, our “sparse+dense” (SD) assumption implies that a few predictors stand out in their effects on classification, while most predictors have small effects. Our SD assumption includes the sparsity assumption as a special case, as when $\boldsymbol{\beta}_k^* = 0$, the discriminant direction is exactly sparse. However, in general our SD assumption is weaker than the sparsity assumption. By incorporating the dense signals, we are essentially assuming that the directions are approximately sparse, in which the sparse signals are most relevant for the classification. But the dense signals contribute to the classification as well, although with less noticeable impacts. As a result, our SD assumption allows us to perform variable selection similar to popular sparse methods. Even though $\boldsymbol{\theta}_k^*$ does not have to be sparse, we can still exploit the sparsity pattern in $\boldsymbol{\delta}_k^*$

to identify the most important variables.

In addition, the dense signals are assumed to have small magnitudes. Similar to sparsity, this is also a type of parsimony assumption that limits the parameter space. Such an assumption is important, as in high dimensions it is challenging to estimate the classifier accurately without appropriate parsimony assumptions. As will be seen later, this assumption brings in helpful regularization techniques in the estimation. We also note that, although the dense signals have small entries individually, jointly they can significantly improve the classification results.

Our SD assumption in (2.5) is imposed on the discriminant direction $\boldsymbol{\theta}_k^*$, because $\boldsymbol{\theta}_k^*$ is often viewed as the most "direct" parameter for classification, as aforementioned. In the literature of sparse LDA methods, sometimes researchers instead assume that the covariance matrix, precision matrix and the mean differences are sparse (Shao et al., 2011; Xu et al., 2015, e.g). In our context, we choose not to make the SD assumption on these parameters for two reasons. For one thing, as discussed above, assumption on the discriminant direction is easy to interpret. For the other, the discriminant direction has $O(p)$ parameters, but the covariance or the precision matrix have $O(p^2)$ parameters and are much more difficult to estimate than the discriminant directions.

As pointed out by a referee, our definition of the sparse and the dense signals could have identifiability issues. However, we will see in Section 2.3 that we are estimating unique target parameters when we employ regularization.

2.3 Estimation

To estimate our model, we first rewrite $\boldsymbol{\theta}_k^*$ as the solution to an optimization problem as suggested by Mai et al. (2019):

$$(\boldsymbol{\theta}_2^*, \ldots, \boldsymbol{\theta}_K^*) = \arg\min_{\boldsymbol{\theta}_k \in \mathbb{R}^p} \sum_{k=2}^{K} \{\frac{1}{2}\boldsymbol{\theta}_k^T \boldsymbol{\Sigma} \boldsymbol{\theta}_k - (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^T \boldsymbol{\theta}_k\}. \tag{2.6}$$

Equation (2.6) apparently cannot be used in estimation, as it involves the unknown parameters $\boldsymbol{\Sigma}, \boldsymbol{\mu}_k$. More importantly, it does not enforce our SD assumption. But we solve these two issues as follows.

To start with, suppose that we observe the dataset $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$, and let $\mathcal{C}_k$ be the set of indices of the n_k samples in class k. We find

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} \boldsymbol{X}_i \tag{2.7}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_k)^T \tag{2.8}$$

as estimates for $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}$. These are also the standard estimates for low-dimensional LDA (Hastie et al., 2009), and popular estimates in many sparse LDA methods (Cai and Liu, 2011; Fan et al., 2012, e.g).

Now we turn to the more interesting problem of imposing the SD assumption. We use the parametrization in (2.5) and regularize $\boldsymbol{\beta}_k^*$ and $\boldsymbol{\delta}_k^*$ respectively. For the sparse signals $\boldsymbol{\delta}_k^*$, we use the group lasso penalty (Yuan and Lin, 2006) to honor the sparsity assumption in (2.4). For the dense signals, we use the ridge penalty (Hoerl and Kennard, 1970; Hastie et al., 2009; Weisberg, 2005). In other words, we consider the penalized problem:

$$
\begin{aligned}
(\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\delta}}_k, k = 2, \ldots, K) = & \underset{\boldsymbol{\beta}_k \in \mathbb{R}^p, \boldsymbol{\delta}_k \in \mathbb{R}^p}{\arg\min} \\
& \sum_{k=2}^{K} \{ \frac{1}{2} (\boldsymbol{\beta}_k + \boldsymbol{\delta}_k)^T \hat{\boldsymbol{\Sigma}} (\boldsymbol{\beta}_k + \boldsymbol{\delta}_k) - (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1)^T (\boldsymbol{\beta}_k + \boldsymbol{\delta}_k) \} \qquad (2.9) \\
+ \lambda_1 \sum_{j=1}^{p} \sqrt{\sum_{k=2}^{K} \delta_{kj}^2} + \lambda_2 \sum_{j=1}^{p} \sum_{k=2}^{K} \beta_{kj}^2, &
\end{aligned}
$$

where $\lambda_1, \lambda_2 > 0$ are tuning parameters. After we obtain $\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\delta}}_k$, we estimate the discriminant direction $\boldsymbol{\theta}_k^*$ by $\hat{\boldsymbol{\theta}}_k = \hat{\boldsymbol{\beta}}_k + \hat{\boldsymbol{\delta}}_k$.

We name this method SD-LDA, where SD refers to the "sparse" and "dense" signals that we target. SD-LDA provides a non-sparse but interpretable classifier, as only a few variables have large effects. When the sparsity assumption does hold, SD-LDA is as powerful as existing sparse methods. However, in the SD problems of our primary interest, SD-LDA continues to be suitable.

It is easy to see that $(\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\delta}}_k, k = 2, \ldots, K)$ produced by (2.9) attempt to approximate

$$(\boldsymbol{\beta}_k^\dagger, \boldsymbol{\delta}_k^\dagger, k = 2, \ldots, K) = \underset{\boldsymbol{\beta}_k \in \mathbb{R}^p, \boldsymbol{\delta}_k \in \mathbb{R}^p}{\arg\min} \sum_{k=2}^{K} \{\frac{1}{2}(\boldsymbol{\beta}_k + \boldsymbol{\delta}_k)^T \boldsymbol{\Sigma}(\boldsymbol{\beta}_k + \boldsymbol{\delta}_k) - (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^T(\boldsymbol{\beta}_k + \boldsymbol{\delta}_k)\} + \lambda_1 \sum_{j=1}^{p} \sqrt{\sum_{k=2}^{K} \delta_{kj}^2} + \lambda_2 \sum_{j=1}^{p} \sum_{k=2}^{K} \beta_{kj}^2, \tag{2.10}$$

Note that, compared to (2.9), in (2.10) we make use of the true parameters $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_k$. Hence, for any fixed pair of tuning parameters (λ_1, λ_2), $(\boldsymbol{\beta}_k^\dagger, \boldsymbol{\delta}_k^\dagger, k = 2, \ldots, K)$ are parameters. With the corresponding penalties, $\boldsymbol{\beta}_k^\dagger$ is dense, while $\boldsymbol{\delta}_k^\dagger$ is sparse. Moreover, unlike $\boldsymbol{\delta}_k^*$ and $\boldsymbol{\beta}_k^*$ intuitively defined in Section 2.2, $(\boldsymbol{\beta}_k^\dagger, \boldsymbol{\delta}_k^\dagger, k = 2, \ldots, K)$ do not have identifiability issues, because (2.10) is strictly convex, and has a unique minimizer. Admittedly, similar to many penalized problems, $(\boldsymbol{\beta}_k^\dagger, \boldsymbol{\delta}_k^\dagger, k = 2, \ldots, K)$ are generally biased in the sense that in general $\boldsymbol{\theta}_k^* \neq \boldsymbol{\beta}_k^\dagger + \boldsymbol{\delta}_k^\dagger$. But if the tuning parameters (λ_1, λ_2) are properly chosen, the discrepancy will be small, and (2.9) will consistently estimate the discriminant directions. See Section 3 for rigorous theoretical justifications.

To gain some intuitions for SD-LDA, we further present the following toy example.

Example 1. Consider a binary classification problem where $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_p$ is known. Then we have the following solution to SD-LDA:

$$\hat{\delta}_{2j} = \text{sign}(\hat{\mu}_{2j} - \hat{\mu}_{1j}) \left(\left| \frac{\hat{\mu}_{2j} - \hat{\mu}_{1j}}{\sigma^2} \right| - \lambda_1 (\frac{1}{2\lambda_2} + \frac{1}{\sigma^2}) \right)_+ \tag{2.11}$$

and

$$\hat{\beta}_{2j} = \frac{\hat{\mu}_{2j} - \hat{\mu}_{1j} - \sigma^2 \hat{\delta}_{2j}}{2\lambda_2 + \sigma^2}, \tag{2.12}$$

for any j, where the soft-threshoding operator $(x)_+ = \max\{x, 0\}$ for any $x \in \mathbb{R}$.

Example 1 illustrates the mechanism of how (2.9) obtains the "sparse signals" and "dense signals" in $\hat{\boldsymbol{\delta}}_2$ and $\hat{\boldsymbol{\beta}}_2$. According to (2.11), each $\hat{\delta}_{2j}$ produces a shrunken standardized mean difference. Therefore, $\hat{\delta}_{2j}$ will only be non-zero when $\frac{\hat{\mu}_{2j} - \hat{\mu}_{1j}}{\sigma^2}$ is large enough with respect to the choices of λ_1 and λ_2. Benefiting from this feature, $\hat{\boldsymbol{\delta}}_2$ is able to identify the signals with large magnitude exclusively. On the other hand, $\hat{\beta}_{2j}$ is essentially a rescaled standardized mean difference reduced by λ_2. When λ_2 is large, (2.12) gives a small $\hat{\beta}_{2j}$., ensuring that $\hat{\boldsymbol{\beta}}_2$ contains the "dense signals". Therefore, the two types of penalties in (2.9) help identify the "sparse signals" and "dense signals" effectively. The ability to capture two types of signals is empowered by the ℓ_1 and ℓ_2 regularization.

Example 1 assumes that $\boldsymbol{\Sigma}$ is diagonal and known to obtain explicit formulas for the estimates, but in practice SD-LDA does not need any knowledge of $\boldsymbol{\Sigma}$. It simply plugs in our sample estimate in (2.8). In what follows, we further discuss another special case where $\boldsymbol{\Sigma}$ does not have any special structure and is unknown. Suppose that $\lambda_1 = \infty$ and thus the sparse signal is estimated as zero. Then the dense signals are estimated by

$$\hat{\boldsymbol{\beta}}_k = (\hat{\boldsymbol{\Sigma}} + 2\lambda_2 \boldsymbol{I})^{-1}(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1). \tag{2.13}$$

It is easy to see that the sample covariance is stabilized by adding $2\lambda_2\boldsymbol{I}$, and resembles the Ledoit-Wolf estimator (Ledoit and Wolf, 2004). The estimator in (2.13) also has a similar form to the regularized discriminant analysis (RDA; Friedman (1989)). However, in the Ledoit-Wolf estimator and RDA, the added identity matrix only intends to make the sample covariance well-conditioned, and usually λ_2 is chosen to be small. In our work, λ_2 is usually reasonably large to encourage the signals to have small magnitudes. When λ_2 is large, $\hat{\boldsymbol{\beta}}_k$ in (2.13) is close to shrunken mean difference. The nearest centroids classifier (Tibshirani et al., 2002, 2003) also use the shrunken mean difference to construct a classifier, but they use the soft-thresholding operator to obtain a sparse classifier, while (2.13) aims to obtain a dense coefficient with the ℓ_2 penalty.

Our proposal is inspired by the lava estimator in Chernozhukov et al. (2017) for regression. Similar to their estimator, we separate the coefficients into the sparse signals and the dense signals, and use a sparsity-inducing penalty and a ridge penalty accordingly. But our estimator is for the classification problem and has many unique challenges. For example, in regression problem there is one coefficient vector to be estimated, while in classification problems we need to estimate several discriminant directions when $K > 2$ so that we can separate all the classes.

Moreover, compared to the lava estimator in regression, the formulation for classification requires additional considerations. In regression problems, the least squares formula is the foundation of most methods, as is the case for the lava estimator. However, in discriminant analysis there are various approaches to find the directions in high dimensions, while no formula dominates the others like the least squares problem in regression. Consequently, we examined many different high-dimensional LDA formulas to find the most suitable one to be generalized to our context. Our SD-LDA is related to the MSDA method (Mai et al., 2019) in that when we are confident in the sparsity assumption, we can set the dense signals to zero, and (2.9) reduces to MSDA. In this sense, (2.9) is a generalization of MSDA to SD problems. We choose to generalize MSDA rather than other candidates for computational concerns. Note that SD-LDA is convex. This is partially because its predecessor MSDA is convex. However, many other high-dimensional LDA methods are non-convex, and their generalizations to the SD problem will continue to be non-convex and potentially challenging in computation. For example, Clemmensen et al. (2011); Fan et al. (2012) both consider optimization problems with equality constraints and are non-convex . In principle, we could also modify these methods by reparametrizing the parameter of interests into sparse and dense signals and adding appropriate penalty functions. But the resulting methods will be non-convex.

We also want to remark that our method is drastically different from the elastic net (Zou and Hastie (2005)), even though we also combine a non-smooth penalty function (group lasso) with the ridge penalty. Elastic net imposes both penalties on the same parameter to stablize the estimator when the predictors are highly correlated. In our method the penalties are enforced on the sparse and dense signals separately to exploit their own structure. On the other hand, in principle we could use other group selection penalty functions to pursue sparsity, such as group SCAD and group MCP (Fan and Li, 2001; Zhang, 2010; Huang et al., 2012). However, these penalty functions are nonconvex, which is likely to lead to instability in computation. The corresponding theoretical study is also expected to be more challenging, as there could be local minimums.

2.4 Algorithm

In this section, we derive an algorithm to solve (2.9). SD-LDA is jointly convex over $(\boldsymbol{\beta}_k, \boldsymbol{\delta}_k)$, but it is most straightforward to derive updates for one of $\boldsymbol{\beta}_k$ and $\boldsymbol{\delta}_k$ while fixing the other, and iterate between them. To this end, we derive the following lemma.

Lemma 1. *Denote* $\hat{\boldsymbol{Q}} = 2\lambda_2 \boldsymbol{I}_p + \hat{\boldsymbol{\Sigma}}$, $\bar{\boldsymbol{\Sigma}} = 2\lambda_2 \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{Q}}^{-1}$, $\hat{\boldsymbol{\mu}}_{dk} = \hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1$, *and* $\bar{\boldsymbol{\mu}}_{dk} = \hat{\boldsymbol{\mu}}_{dk}^T (2\lambda_2 \hat{\boldsymbol{Q}}^{-1})$. *Then we have*

1. *for a fixed* $\boldsymbol{\delta}_k$, *the optimizer of* $\boldsymbol{\beta}_k$ *to* (2.9) *is*

$$\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k) = (2\lambda_2 \boldsymbol{I}_p + \hat{\boldsymbol{\Sigma}})^{-1}(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\Sigma}}\boldsymbol{\delta}_k); \tag{2.14}$$

2. *The optimizer of* $\boldsymbol{\delta}_k$ *to* (2.9) *is*

$$(\hat{\boldsymbol{\delta}}_2, \ldots, \hat{\boldsymbol{\delta}}_K) = \arg\min_{\boldsymbol{\delta}_k \in \mathbb{R}^p} \sum_{k=2}^{K} \{\frac{1}{2}[\boldsymbol{\delta}_k^T \bar{\boldsymbol{\Sigma}} \boldsymbol{\delta}_k] - \bar{\boldsymbol{\mu}}_{dk}^T \boldsymbol{\delta}_k\} + \lambda_1 \sum_{j=1}^{p} \sqrt{\sum_{k=2}^{K} \delta_{kj}^2}. \tag{2.15}$$

According to Lemma 1, we first solve (2.15) and then plug its results into (2.14) to find the SD-LDA estimate. Note that the ℓ_2 regularization enables us to invert the covariance matrix in (2.14), so that (2.14) is feasible even in high dimensions. For the solution of (2.15), although it does not have an explicit form, we can find it by modifying the groupwise coordinate descent algorithm in Mai et al. (2019). We replace their $\widehat{\boldsymbol{\Sigma}}$ with $\bar{\boldsymbol{\Sigma}}$, and $\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1$ with $\bar{\boldsymbol{\mu}}_{dk}$, respectively, to solve (2.15).

3. Theory

In this section, we present the theoretical properties of SD-LDA. The entire theoretical study is based on the LDA model setup given by (2.1). For ease of presentation, for two quantities A and ξ, we write $A \lesssim \xi$ if $A \leq C\xi$ for some $C > 0$.

We also make the following assumptions.

(A1) $||\boldsymbol{\Sigma}||_2 \leq u$ and $||\boldsymbol{\Sigma}^{-1}||_2 \leq U$ for some constants U and u,

(A2) $\max_k ||\boldsymbol{\mu}_k - \boldsymbol{\mu}_1||_2 \leq w_1$ for some constant w_1,

(A3) $0 < c_2 < \pi_k < c_1 < 1$ for some constants c_1 and c_2.

Assumptions (A1) & (A2) are technical conditions that facilitate our proof. Assumption (A1) implies that the eigenvalues of $\boldsymbol{\Sigma}$ are finite and bounded away from 0 while Assumption (A2) requires a bound on the ℓ_2-norm of the mean difference. Assumption (A3) is made to guarantee that our dataset is not extremely unbalanced. We have the following theorem.

Theorem 1. *Let $\hat{\boldsymbol{\theta}}_k = \hat{\boldsymbol{\delta}}_k + \hat{\boldsymbol{\beta}}_k$, with $\hat{\boldsymbol{\beta}}_k$ and $\hat{\boldsymbol{\delta}}_k$ defined as in (2.14) and (2.15), and Assumptions (A1), (A2), and (A3) hold. Then with probability* *at least* $1 - O(p^{-1})$, *we have*

$$||\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*||_2 \lesssim \sqrt{\frac{p \log p}{n}} \tag{3.16}$$

for $\lambda_2 = O(\sqrt{\frac{\log p}{n}})$ *and* $\lambda_1 = O(\sqrt{\frac{\log p}{n}})$.

Theorem 1 shows that the estimators for the linear discriminant directions will consistently converge to the truth when $\frac{p \log p}{n} \to 0$. This implies that SD-LDA approaches the Bayes rule under the same dimensionality. Hence, in theory SD-LDA works similarly as the true Bayes rule as the sample size increases.

Although SD-LDA allows p to diverge, we acknowledge that it has a stronger requirement on the dimensionality than sparse methods. Sparse methods often allow p to diverge at an exponential rate (Cai and Liu, 2011; Fan et al., 2012; Mai et al., 2012). Theorem 1 has a stronger requirement because we no longer make the sparsity assumption and the problem is much harder. But the difficulty could be technical, because SD-LDA turns out to work well on high-dimensional problems with $p > n$ in our numerical studies. Also, in the special case of exact sparsity where we know $\boldsymbol{\beta}_k^* = 0$, SD-LDA reduces to the sparse classifier MSDA (Mai et al., 2019), which has an optimal convergence rate of $\sqrt{\frac{s \log p}{n}}$ (Min and Mai, 2022), with s being the number of important variables. Moreover, we had to conquer many technical challenges to obtain the current theoretical result. On one hand, even though SD-LDA is inspired by the lava estimator in regression, our proof is drastically different from theirs, as most of their proof conditions on $\boldsymbol{X}$, but we have to directly handle the randomness in $\boldsymbol{X}$ as a consequence of the LDA model assumption. Also, in multiclass problems we have multiple directions to estimate, which adds to the technical difficulty of the proof. On the other hand, compared to the sparse classifier MSDA, SD-LDA involves much more complicated functionals, such as $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}} + 2\lambda_2 \boldsymbol{I})^{-1}$. We need to establish bounds for these terms that are not available in the literature. Since the focus of this paper is method development, we leave more careful theoretical investigation for SD-LDA as a future topic, which we expect to be very challenging.

4. Simulation

We present the simulation study to examine the performance of our proposal in this section. We consider two scenarios separately: the settings where the sparsity assumption holds and the settings where the sparsity assumption does not but the SD assumption holds. The sparse models are given by S1 and S2 and the SD models are given by Models D1–D6. Throughout the simulation, we set the sample size $n_k = 50$ for each class and $p =$

$250K$. Simulations of imbalanced classes can be found in Section S1 in the Supplementary Materials. For each class, we set $\boldsymbol{X} \mid Y = k \sim N(\boldsymbol{\mu}_k, \sigma^2\boldsymbol{\Sigma})$, where σ^2 is a constant that varies from model to model. For each model, we have different $\boldsymbol{\delta}_k^*$ and $\boldsymbol{\beta}_k^*$, with $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_k = \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\beta}_k^* + \boldsymbol{\delta}_k^*)$ for $k = 2, \ldots, K$. In sparse models, $\boldsymbol{\beta}_k^* = \mathbf{0}$ for all k. For the five SD models, We choose q from $\{0.1, 0.15, 0.2\}$ for each model, where q presents the signal strength of "dense signals". The models are given as follows:

S1 : $K = 2$, $\sigma = 0.5$. $\boldsymbol{\Sigma}$ is block-diagonal, with each block $\boldsymbol{\Sigma}_s$ being a 4×4 auto-regressive matrix with parameter 0.5, $\boldsymbol{\delta}_2^* = (\mathbf{2}_5, \mathbf{0}_{495})$.

S2 : $K = 3$, $\sigma = 0.5$. $\boldsymbol{\Sigma}$ is block-diagonal, with each block $\boldsymbol{\Sigma}_s$ being a 4×4 auto-regressive matrix with parameter 0.5, $\boldsymbol{\delta}_2^* = (\mathbf{2}_5, \mathbf{0}_{745})$, $\boldsymbol{\delta}_3^* = (\mathbf{0}_5, -\mathbf{2}_5, \mathbf{0}_{740})$.

D1 : $K = 2$, $\sigma = 0.5$. $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $\boldsymbol{\delta}_2^* = (2.5, \mathbf{0}_{499})$ and $\boldsymbol{\beta}_2^* = (0, \boldsymbol{q}_{499})$, $q \in \{0.1, 0.2\}$.

D2 : $K = 2$, $\sigma = 0.5$. $\boldsymbol{\Sigma}$ is block-diagonal, with each block $\boldsymbol{\Sigma}_s$ being a 4×4 compound symmetry matrix with parameter 0.5, $\boldsymbol{\delta}_2^* = (2.5, \mathbf{0}_{499})$ and $\boldsymbol{\beta}_2^* = (0, \boldsymbol{q}_{499})$, $q \in \{0.1, 0.2\}$.

D3 : $K = 2$, $\sigma = 0.5$. $\boldsymbol{\Sigma}$ is block-diagonal, with each block $\boldsymbol{\Sigma}_s$ being a 4×4 auto-regressive matrix with parameter 0.5, $\boldsymbol{\delta}_2^* = (2.5, \mathbf{0}_{499})$ and

$\boldsymbol{\beta}_2^* = (0, \boldsymbol{q}_{499})$, $q \in \{0.1, 0.2\}$.

D4 : $K = 3$, $\sigma = 0.6$. $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $\boldsymbol{\delta}_2^* = (2.5, \mathbf{0}_{749})$, $\boldsymbol{\delta}_3^* = (-2.5, \mathbf{0}_{749})$, $\boldsymbol{\beta}_2^* = (0, \boldsymbol{q}_{749})$ and $\boldsymbol{\beta}_3^* = (0, -\boldsymbol{q}_{749})$, $q \in \{0.15, 0.2\}$.

D5 : $K = 3$, $\sigma = 0.5$. $\boldsymbol{\Sigma}$ is block-diagonal, with each block $\boldsymbol{\Sigma}_s$ being a 4×4 auto-regressive matrix with parameter 0.5, $\boldsymbol{\delta}_2^* = (2.5, \mathbf{0}_{749})$, $\boldsymbol{\delta}_3^* = (-2.5, \mathbf{0}_{749})$, $\boldsymbol{\beta}_2^* = (0, \boldsymbol{q}_{749})$ and $\boldsymbol{\beta}_3^* = (0, -\boldsymbol{q}_{749})$, $q \in \{0.1, 0.15\}$.

D6 : $K = 5$, $\sigma = 0.5$. The covariance matrix $\boldsymbol{\Sigma}$ has a auto-regressive structure, namely $\sigma_{ij} = 0.8^{|i-j|}$. The number of non-zero elements of parameter $\boldsymbol{\beta}'s$ is 5 instead of 1. Specifically, We let $(\boldsymbol{\delta}_2^*, \cdots, \boldsymbol{\delta}_5^*) = (\mathbf{1.5}_5, -\mathbf{1.5}_5, \mathbf{1}_5, -\mathbf{1}_5)$ and $(\boldsymbol{\beta}_2^*, \cdots, \boldsymbol{\beta}_5^*) = (\mathbf{2}\boldsymbol{q}_{495}, -\mathbf{2}\boldsymbol{q}_{495}, \boldsymbol{q}_{495}, -\boldsymbol{q}_{495})$, $q \in \{0.075, 0.1\}$.

In addition to SD-LDA, we include the following competitors in our simulation: MSDA (Mai et al. (2019)), logistic regression with Lasso penalty (Hastie et al. (2009)) or elastic-net penalty (Zou and Hastie (2005)) (denoted as Lasso and elastic-net), SVM (Joachims (1998)), and sparse optimal scoring (Clemmensen et al. (2011), denoted as SOS). MSDA is implemented by `R` package `msda`, Lasso and elastic-net are implemented by `R` package `glmnet`, and SVM is implemented by `R` package `e1071`. SOS is implemented by `R` package `sparseLDA` in the multiclass models and implemented by `R` package `TULIP` in binary models (Pan et al., 2020).

The tuning parameters in all methods are chosen by the 5-fold cross-validation, and a grid search is implemented if there are multiple tuning parameters. We run the simulation 100 times for each model and the means and standard errors of the prediction error (PE in %) are reported in Table 1. The mean of the number of correctly and incorrectly selected variables are given in Table 2. Recall that the sparse signals stand out in their effects, and thus we focus on the selection of them. For all the competitors, a variable is selected if and only if has a nonzero coefficent, while for SD-LDA, a variable is selected if it has a nonzero coefficient in $\hat{\boldsymbol{\delta}}_k$.

In Table 1, it can be seen that, even when the true models are sparse, SD-LDA still gives a comparable, or even significantly better result than the sparse competitors. The better results could be explained by the fact that SD-LDA can approximate a sparse classifier by using a large λ_2, but it explores more classifiers through cross validation over λ_2 that gives better empirical results. On the SD models, SD-LDA has a clear advantage over the sparse methods. Such an observation shows that the sparse methods are vulnerable when dense signals exist, and SD-LDA is preferrable under such circumstances.

In Table 2, we see that SD-LDA continues to give excellent variable

Table 1: The prediction accuracy result. The means and standard errors (in the parentheses) of the prediction error of 100 replicates are reported in percentage.

Models	q	BE	SD-LDA	MSDA	Lasso	elastic-net	SVM	SOS
S1	0	6.4	6.51(0.23)	7.83(0.28)	8.58(0.3)	7.2(0.27)	22.93(0.44)	8.69(0.3)
S2	0	8.3	8.54(0.24)	9.29(0.26)	11.45(0.3)	10.51(0.26)	29.15(0.36)	17.23(0.31)
D1	0.1	20.1	25.27(0.45)	27.26(0.47)	28.4(0.5)	26.28(0.46)	40.79(0.55)	28.2(0.5)
	0.2	10.0	21.08(0.45)	27.18(0.52)	28.35(0.53)	25.4(0.47)	28.18(0.48)	27.57(0.45)
D2	0.1	13.5	18.9(0.38)	26.59(0.47)	26.99(0.48)	24.18(0.42)	25.13(0.45)	25.91(0.49)
	0.2	2.8	5.63(0.22)	21.65(0.48)	19.46(0.44)	10.94(0.33)	5.92(0.23)	17.08(0.49)
D3	0.1	15.2	20.42(0.42)	26.24(0.46)	26.76(0.45)	24.34(0.41)	29.3(0.44)	27.14(0.45)
	0.2	4.1	7.99(0.28)	22.67(0.41)	21.18(0.42)	13.99(0.35)	9.33(0.28)	21.79(0.47)
D4	0.15	9.9	28.57(0.37)	31.23(0.41)	33.9(0.4)	32.47(0.32)	34.43(0.4)	35.61(0.41)
	0.2	8.3	20.76(0.38)	31.59(0.41)	33.75(0.37)	31.15(0.31)	23.38(0.32)	35.35(0.39)
D5	0.1	20.4	28.63(0.39)	35.36(0.42)	38(0.38)	35.11(0.38)	31.09(0.38)	39.06(0.41)
	0.15	5.9	17.54(0.4)	35.04(0.42)	35.41(0.39)	29.26(0.33)	14.66(0.3)	37.45(0.37)
D6	0.075	6.3	9.11 (0.18)	28.12 (0.26)	33.41 (0.36)	28.10 (0.29)	11.51 (0.20)	27.34 (0.31)
	0.1	2.4	3.1 (0.10)	26.14 (0.31)	28.52 (0.33)	18.47 (0.28)	4.0 (0.14)	24.15 (0.32)

selection results across all models. In the SD models, the variable selection becomes worse as q increases. This is expected, because as q increases, the boundary between sparse and dense signals is blurred. However, the sparse competitors struggle much more than SD-LDA. As they are incapable of modeling the dense signals, they drastically overselect the variables in hope to achieve higher accuracy.

Table 2: The means and standard errors of correctly selected variables (denoted as C) and incorrectly selected variables (denoted as IC) are reported. The true numbers of important/dense signals are 5 for Model S1, 10 for Model S2, and 1 for all SD models.

		SD-LDA		MSDA		Lasso		elastic net		SOS	
Models	q	C	IC	C	IC	C	IC	C	IC	C	IC
S1	0	4.5(0.7)	4.6(5)	4.5(0.8)	5.6(6.8)	4.7(0.5)	19.5(8.6)	4.9(0.2)	43.6(30.3)	4.8(0)	19.4(2.3)
S2	0	9.2(1)	5.1(6.7)	9.1(1.1)	4.6(6.9)	1.8(1.1)	33.2(7.4)	5.5(2.6)	67(36.2)	10(0)	4.4(0.1)
D1	0.1	1(0)	4.2(4.4)	1(0)	4.5(6.3)	1(0)	9.2(9)	1(0)	15(20.6)	1(0)	13.6(2.2)
	0.2	1(0.1)	4.7(5.2)	1(0)	6.7(7.5)	1(0)	18.2(13.4)	1(0)	56(52.1)	1(0)	30.8(3.5)
D2	0.1	1(0)	1.8(3.1)	1(0)	10.5(10.9)	1(0)	22.9(13.1)	1(0)	65.8(51)	1(0)	23.5(3)
	0.2	0.9(0.3)	1.5(2.8)	1(0)	24(12.1)	1(0)	46.2(7.5)	1(0)	189.1(29.7)	1(0)	68.8(3.1)
D3	0.1	1(0)	2.7(3.7)	1(0)	7.7(9)	1(0)	18.7(12.9)	1(0)	51.1(43.5)	1(0)	17.3(2.6)
	0.2	0.9(0.2)	1.1(2)	1(0)	21.9(11.1)	1(0)	42.1(10.1)	1(0)	175.1(44.2)	1(0)	58.2(3.3)
D4	0.15	1(0.1)	0.6(3.4)	1(0)	1.2(3.4)	0(0)	19.9(11.6)	0(0.1)	57(69.5)	1(0)	22.8(0)
	0.2	0.9(0.3)	0(0)	1(0)	5.4(13)	0(0)	23.1(12.8)	0(0.1)	109.7(87.4)	1(0)	26.7(0.1)
D5	0.1	1(0.1)	1.2(4.4)	1(0)	5.3(14.3)	0(0)	19.6(10.8)	0(0.1)	102.2(79.4)	1(0)	28.7(0.1)
	0.15	0.9(0.3)	1.6(6.2)	1(0)	30.4(26.5)	0(0)	22.9(10.1)	0.1(0.3)	168.2(66.5)	1(0)	18.9(0)
D6	0.075	2.6(0.2)	17.5(3.1)	3.3(0.1)	0.11(0)	0(0)	26.89(0.7)	0.31(0.1)	137.7(3.6)	2.3(0.1)	139.9(3.9)
	0.1	2.4(0.1)	11.3(2.7)	3.8(0.1)	0.81(0.2)	0(0)	26.7(0.6)	0.36(0.1)	153.0(1.3)	2.3(0.1)	139.0(3.7)

Finally, we report the computation cost for SD-LDA and its competitors in Table 3. For the sake of space, we only report the results for Models D1 & D6 with $q = 0.1$ at the optimal tuning parameters. The computation time is averaged over 100 replicates. Most methods take much less than 1 second to finish one replicate. We admit that SD-LDA is slower than most of the competitors, which is the price we pay to model the dense signals

Time ($s \times 10^{-2}$)	SD-LDA	MSDA	Lasso	elastic-net	SVM	SOS
D1	11.3	2.3	0.18	0.18	2.5	1.2
D6	17.9	5.2	8.4	9.0	10.4	104.7

Table 3: Average computation time based on 100 replicates.

and achieve better prediction accuracy. Among the discriminant analysis methods (SD-LDA, MSDA and SOS), SD-LDA is slower than MSDA because it needs to calculate $(\hat{\boldsymbol{\Sigma}} + 2\lambda_2 \boldsymbol{I})^{-1}$ in estimating the dense signals. For the comparison with SOS, SD-LDA is slower than SOS in the relatively simple Model (D1). However, in the more difficult Model (D6), SD-LDA is much faster than SOS, even though SOS only estimates the sparse signals.

5. Real Dataset Analysis

We further demonstrate the performance of SD-LDA on five real-world datasets: The IBD dataset from Burczynski et al. (2006), the small-blue-round-cell tumour dataset (SBRCT) from Khan et al. (2001), the prostate cancer dataset from Singh et al. (2002), the gene time data set, and the cancer genome atlas data set. The screened IBD dataset is imported from the `R` package `msda` directly. It contains 127 samples in three classes and 127 gene expressions. The SBRCT dataset has 84 samples in four classes and the prostate cancer dataset has 102 samples in two classes. As the dimensions of the SBRCT and the prostate cancer datasets are extremely high, we first apply t-test screening Fan and Fan (2008) on them before performing our proposal. The reduced datasets are generated by the t-test screening with p-values of screening set to 0.001 and 0.05, respectively. The numbers of their gene expressions are reduced to 594 and 477, respectively.

The gene time course data (GTC) describes clinical response to treatment for Multiple Sclerosis (MS) patients based on gene expression time course data. This dataset is originally described in Baranzini et al. (2004). Fifty-three patients were given recombinant human interferon beta (rIFNβ), which is often used to control the symptoms of MS. Gene expression was measured for 76 genes of interest before treatment (baseline) and at 6 follow-up time points over the next two years (3 months, 6 months, 9 months, 12 months, 18 months, 24 months). Afterward, patients were classified into good responders or poor responders to rIFNβ based on clinical characteristics. There are 20 good responders and 33 bad responders in all the 53 patients. The dimension for this data set is $76 \times 7 = 532$.

The Cancer Genome Atlas (TCGA) Research Network has profiled and analyzed large numbers of human tumors to discover molecular aberrations at the DNA, RNA, protein and epigenetic levels. This data is part of the Pan-Cancer data set, it is a random extraction of gene expressions of patients having different types of tumor: BRCA, KIRC, COAD, LUAD and PRAD. We download the data from *https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq*, which is kindly shared by Samuele Fiorini in 2016. The original data set is maintained by the cancer genome atlas pan-cancer analysis project (*https://www.synapse.org/*). The data set has 801 samples in five classes and 20531 gene expressions. As the previous data sets, we first use t-test screening method to select 801 genes.

We perform SD-LDA and all the other competitors included in Section 4 on the first 4 datasets. We run 100 replicates, and in each replicate the datasets are split with the 2:1 ratio in a balanced manner to form training and testing sets. The tuning parameters are chosen by 5-fold cross-validation and check their prediction errors. For the TCGA data, almost all the methods perform perfectly. To make the classification problem more challenging, we run 100 replicates with data set split with the 1:9 ratio in a balanced manner to form training and testing sets, namely we have 80 training samples and 721 test samples.

The average prediction errors are reported in Table 4. SD-LDA is either the best classifier, or statistically no different from the best classifier. These result support the application of SD-LDA in practice. We note that the

Table 4: The means and standard errors (in the parentheses) of the prediction error(%) of SD-LDA and other competitors of 100 replicates are reported for each dataset.

DataSet	SD-LDA	MSDA	Lasso	elastic-net	SVM	SOS
IBD	3.71(0.28)	8(0.37)	6.76(0.36)	5.1(0.34)	8.27(0.38)	6.93(0.33)
SBRCT	0.08(0.08)	14.08(0.9)	1.58(0.2)	0(0)	0.92(0.17)	2.15(0.26)
Prostate	0(0)	29.06(0.74)	19.06(0.58)	2.79(0.28)	0(0)	15.33(0.63)
GTC	17.4 (0.74)	33.6 (0.97)	31.7 (0.85)	21.2 (0.87)	31.6 (0.51)	45.2 (1.31)
TCGA	0.20(0.02)	1.57 (0.14)	0.58 (0.02)	0.42 (0.01)	0.37 (0.01)	4.90 (0.12)

sparse methods could have better performance if we preprocess the datasets a little differently. For example, Mai et al. (2012) reported the prediction error of 5.9% for MSDA (its binary equivalence to be exact) if the dataset is not screened. However, the error of 5.9% therein is still significantly larger than that of SD-LDA on the screened data in Table 4.

We further check the variable selection results in Table 5. It can be seen that by involving the "dense signals" into consideration, SD-LDA actually selects the fewest sparse signals. Therefore, SD-LDA could help researchers focus on a smaller set of key features for more in-depth study. Because the gene names for IBD, Prostate and TCGA data sets are missing in the resources where we obtained them. We only report the selected genes for

Table 5: The means and standard errors of the number of selected variables by SD-LDA and other competitors of 100 replicates are reported.

Methods	SD-LDA	MSDA	Lasso	elastic-net	SOS
IBD	9.4(0.82)	29.46(0.7)	10.67(0.18)	69.83(0.5)	52.36(1.89)
SBRCT	7.62(0.47)	17.32(0.54)	10.97(0.23)	140.45(0.67)	59.07(0.99)
Prostate	11.37(0.52)	12.66(0.4)	43.59(0.59)	233.7(0.93)	64.67(1.15)
GTC	2.04(0.18)	6.09(0.34)	12.52(0.66)	110.07(2.84)	25.65(1.15)
TCGA	19.13(0.76)	38.71(0.84)	9.01(0.19)	125.51(0.74)	108.34(2.71)

SBRCT and GTC data set. For SBRCT data, SD-LDA selects 10 genes, including WASp, CAV1, CDH2, HBE1, anti-CD99, Psmb10, HLA-DMA, SYNGR1, EHD1, and PSMB8. Some of those genes have been shown to have close relationship with the development of cancer. For example, CAV1 appears to act as a tumor suppressor protein at early stages of cancer progression (Sáinz-Jaspeado et al., 2011); CD99 appears to be a robust marker of cancer stem cells and a promising therapeutic target in these malignancies (Pasello et al., 2018); HLA-DMA antigen expression by tumor cells influences the tumor antigen (TA)-specific immune responses and, depending on the cancer type, the clinical course of the disease (Seliger et al., 2017). For GTC dataset, SD-LDA select 2 genes, including Caspase 6 and FLIP. This agrees with the existing literature e.g. Julien et al. (2016) showed that Caspase 6 is related to MS, (Hauser and Oksenberg, 2006) showed that FLIP is related to MS.

6. Discussion

SD-LDA is proposed as a convex high-dimensional classification method that is robust to the changing signal pattern in linear discriminant directions. It is obtained by separating the linear discriminant directions into “sparse signals” and “dense signals” and applying suitable penalties to estimate them. To the best of knowledge, this is the first SD classifier, which turns to have favorable performance over a wide range of datasets. Similar techniques could be combined with various linear classifiers, such as the logistic regression or SVM, to empower them with the ability to capture the “sparse signals” and “dense signals”. One can further consider a similar modification for nonlinear models, such as the quadratic discriminant analysis (QDA; Fan et al. (2015); Li and Shao (2015); Jiang et al. (2018)).

On the other hand, our work is developed under the LDA model in which the within-class distribution is normal. But we can easily extend it to a semiparametric framework that has been reasonably well studied (Lin and Jeon, 2003; Liu et al., 2009; Mai and Zou, 2015; Jiang and Leng, 2016). In this semiparametric framework, $\boldsymbol{X}$ does not have to be normal within each class, but there exists a set of unknown transformations $g = (g_1, \ldots, g_p)$ such that $(g(\boldsymbol{X}), Y)$ follows the LDA model. When considering the high-dimensional data, existing works frequently adopt the sparsity assumption. However, following our work, we can consider a semiparametric model with sparse+dense signals to achieve more flexibility. We leave this topic as an interesting future research direction, for which some recent theoretical works on estimating the transformation may be helpful as well (Mai et al., 2022). However, such studies are out of the scope of this paper.

Supplementary Materials

The derivation for algorithms and the proof for theorems are available in the supplementary materials. Section S2 contains the derivation for Lemma 1 and Section S3 contains the proof for Theorem 1.

Acknowledgements

The authors are grateful to the Editor, Associate Editor and referees for their helpful suggestions that greatly improved this paper. Mai's research is partially supported by CCF-1908969, National Science Foundation.

References

Baranzini, S. E., P. Mousavi, J. Rio, S. J. Caillier, A. Stillman, P. Villoslada, et al. (2004). Transcription-based prediction of response to ifnβ using supervised computational methods. *PLoS biology 3*(1), e2.

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

Burczynski, M. E., R. L. Peterson, N. C. Twine, K. A. Zuberek, B. J. Brodeur, L. Casciotti, V. Maganti, P. S. Reddy, A. Strahs, F. Immermann, et al. (2006). Molecular classification of crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *The journal of molecular diagnostics 8*(1), 51–61.

Cai, T. and W. Liu (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American statistical association 106*(496), 1566–1577.

Chernozhukov, V., C. Hansen, Y. Liao, et al. (2017). A lava attack on the recovery of sums of dense and sparse signals. *Annals of Statistics 45*(1), 39–76.

Clemmensen, L., T. Hastie, D. Witten, and B. Ersbøll (2011). Sparse discriminant analysis. *Technometrics 53*(4), 406–413.

Fan, J. and Y. Fan (2008). High dimensional classification using features annealed independence rules. *Annals of statistics 36*(6), 2605.

Fan, J., Y. Feng, and X. Tong (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(4), 745–771.

Fan, J., Z. T. Ke, H. Liu, and L. Xia (2015). Quadro: A supervised dimension reduction method via rayleigh quotient optimization. *Annals of statistics 43*(4), 1498.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics New York.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American statistical association 84*(405), 165–175.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Hauser, S. L. and J. R. Oksenberg (2006). The neurobiology of multiple sclerosis: genes, inflammation, and neurodegeneration. *Neuron 52*(1), 61–76.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Huang, J., P. Breheny, and S. Ma (2012). A selective review of group selection in high-dimensional models. *Statistical Science 27*.

Jiang, B. and C. Leng (2016). High dimensional discrimination analysis via a semiparametric model. *Statistics & Probability Letters 110*, 103–110.

Jiang, B., X. Wang, and C. Leng (2018). A direct approach for sparse quadratic discriminant analysis. *The Journal of Machine Learning Research 19*(1), 1098–1134.

Joachims, T. (1998). Making large-scale svm learning practical. Technical report, Technical report.

Julien, O., M. Zhuang, A. P. Wiita, A. J. O'Donoghue, G. M. Knudsen, C. S. Craik, and J. A. Wells (2016). Quantitative ms-based enzymology of caspases reveals distinct protein substrate specificities, hierarchies, and cellular roles. *Proceedings of the National Academy of Sciences 113*(14), E2001–E2010.

Khan, J., J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine 7*(6), 673–679.

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

Li, Q. and J. Shao (2015). Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica*, 457–473.

Lin, Y. and Y. Jeon (2003). Discriminant analysis through a semiparametric model. *Biometrika 90*(2), 379–392.

Liu, H., J. Lafferty, and L. Wasserman (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research 10*(10).

Mai, Q., D. He, and H. Zou (2022). Coordinatewise gaussianization: Theories and applications. *Journal of the American Statistical Association*, 1–15.

Mai, Q., Y. Yang, and H. Zou (2019). Multiclass sparse discriminant analysis. *Statistica Sinica 29*(1), 97–111.

Mai, Q. and H. Zou (2015). Sparse semiparametric discriminant analysis. *Journal of Multivariate Analysis 135*, 175–188.

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

Min, K. and Q. Mai (2022). Optimality in high-dimensional tensor discriminant analysis. *Manuscript*.

Pan, Y., Q. Mai, and X. Zhang (2020). Tulip: A toolbox for linear discriminant analysis with penalties. *The R Journal 12*(2).

Pasello, M., M. C. Manara, and K. Scotlandi (2018). Cd99 at the crossroads of physiology and pathology. *Journal of cell communication and signaling 12*(1), 55–68.

Sáinz-Jaspeado, M., J. Martin-Liberal, L. Lagares-Tena, S. Mateo-Lozano, X. G. Del Muro, and O. M. Tirado (2011). Caveolin-1 in sarcomas: friend or foe? *Oncotarget 2*(4), 305.

Seliger, B., M. Kloor, and S. Ferrone (2017). Hla class ii antigen-processing pathway in tumors:

molecular defects and clinical relevance. *Oncoimmunology 6*(2), e1171447.

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

Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell 1*(2), 203–209.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences 99*(10), 6567–6572.

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

Weisberg, S. (2005). *Applied linear regression*, Volume 528. John Wiley & Sons.

Witten, D. M. and R. Tibshirani (2011). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(5), 753–772.

Xu, P., J. Zhu, L. Zhu, and Y. Li (2015). Covariance-enhanced discriminant analysis. *Biometrika 102*(1), 33–45.

Yang, H., D. Lin, and Q. Li (2022). An efficient greedy search algorithm for high-dimensional linear discriminant analysis. *Statistica Sinica in press.*

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics 38*, 894–942.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology) 67*(2), 301–320.

Ning Wang

Center for Statistics and Data Science, Beijing Normal University, Zhuhai, Guangdong, China, 519087 E-mail: ningwangbnu@bnu.edu.cn

Shaokang Ren

Microsoft Corporation, Redmond, WA, USA, 98052 E-mail: sr17k@fsu.edu

Qing Mai

Department of Statistics, Florida State University, 600 W College Ave, Tallahassee, FL 32306, USA

E-mail: qmai@fsu.edu