Electronic Journal of Statistics
Vol. 16 (2022) 814–861
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1939

New hard-thresholding rules based on data splitting in high-dimensional imbalanced classification*

Arezou Mojiri

Department of Mathematical Sciences, Isfahan University of Technology, Isfahan, 19395-5746, Iran. e-mail: arezu.mojiri@math.iut.ac.ir

Abbas Khalili †

Department of Mathematics and Statistics, McGill University, Montreal, H3A 0B9, Canada e-mail: abbas.khalili@mcgill.ca

and

Ali Zeinal Hamadani

Department of Industrial and Systems Engineering, Isfahan University of Technology, Isfahan, 19395-5746, Iran. e-mail: hamadani@cc.iut.ac.ir

Abstract: In binary classification, imbalance refers to situations in which one class is heavily under-represented. This issue is due to either a data collection process or because one class is indeed rare in a population. Imbalanced classification frequently arises in applications such as biology, medicine, engineering, and social sciences. In this paper, for the first time, we theoretically study the impact of imbalance class sizes on the linear discriminant analysis (LDA) in high dimensions. We show that due to data scarcity in one class, referred to as the minority class, and high-dimensionality of the feature space, the LDA ignores the minority class yielding a maximum misclassification rate. We then propose a new construction of hard-thresholding rules based on a data splitting technique that reduces the large difference between the misclassification rates. We show that the proposed method is asymptotically optimal. We further study two well-known sparse versions of the LDA in imbalanced cases. We evaluate the finite-sample performance of different methods using simulations and by analyzing two real data sets. The results show that our method either outperforms its competitors or has comparable performance based on a much smaller subset of selected features, while being computationally more efficient.

MSC2020 subject classifications: 62H30.
Keywords and phrases: Classification, high-dimensionality, imbalanced, linear discriminant analysis, thresholding.

Received November 2020.

*Abbas Khalili was supported by the Natural Sciences and Engineering Research Council of Canada through Discovery Grants (NSERC RGPIN-2015-03805 and NSERC RGPIN-2020-05011).
†Corresponding Author.

Contents

1. Introduction

The rise of high-dimensional data has affected many areas of research in statistics and machine learning, including classification. Linear Discriminant Analysis (LDA) has been extensively studied in high-dimensional classification. [4], [12], and [38] showed that when the number of features is larger than the sample size, the LDA can perform as badly as a random guess. To deal with the curse of dimensionality, several developments have been made over the last decade or so. For example, among others, new developments include the nearest shrunken centroids [40], shrunken centroids regularized discriminant analysis [18], features annealed independence rule (FAIR) [12], sparse and penalized LDA [38, 42], regularized optimal affine discriminant (ROAD) [13], multi-group sparse discriminant analysis [16], pairwise sure independent screening [29], and the ultra high-dimensional multiclass LDA [23]. The general idea of these methods is to incorporate a feature selection strategy in a classifier in order to obtain certain optimality properties in the sense of misclassification rates.

To the best of our knowledge, most of the existing developments in high dimensions focus on problems with comparable class sizes in the training data. However, in applications such as clinical diagnosis [2], fraud detection [9], drug

discovery [44], or equipment malfunction detection [31], classification often suffers from imbalanced class sizes where, for example, in a binary problem one class (referred to as the minority class) is heavily under-represented. This is due to either a data collection process or because one class is indeed rare in a population. In such situations, the minority class is of primary interest as it carries substantial information, and often has higher misclassification costs compared to the larger class, referred to as the majority class. For example, in a study of a certain rare disease, the cost of misclassifying a positive case is often higher than the cost of misclassifying a negative one [36]. In banking or telecommunication studies, few customers are voluntarily willing to terminate their contracts and leave their provider. In these applications, misclassification of a potential churner is more expensive than that of a non-churner for a provider [41]. Due to data scarcity in the minority class, conventional discriminant methods are often biased toward the majority class resulting in much higher misclassification rate for the minority class. This error dramatically increases in high-dimensional cases, as empirically shown by [7]. In this paper, we study imbalanced binary classification with the class sizes $n_2 \ll n_1$, when the number of features, p_n, grows to infinity as the total sample size $n = (n_1 + n_2)$ grows to infinity. We refer to Class 1 with size n_1 as the majority class, and Class 2 with size n_2 as the minority class. A specific limiting relationship between n_1 and n_2 is given in Section 2.2.

Imbalanced classification under various settings have attracted attention in recent years. A common approach to deal with the imbalanced issue is to make virtual class sizes comparable by using resampling methods, for example, the synthetic minority over-sampling technique (SMOTE) of [10]. The recent work of [15] provides a review of the common re-sampling techniques for fixed dimensional imbalanced problems. In other methods, such as the weighted extreme learning machine [45] and the cost-sensitive support vector machine (SVM) [21], the idea is to strengthen the relative impact of the minority class by either assigning different weights to sample units or different costs to misclassification instances in each class. [3] studied distributional properties of the correct classification probabilities of the minority and majority classes of a hard-thresholding independence rule. Using a non-asymptotic approach, they adjusted the bias of correct classification probabilities which is rooted on the imbalanced class sizes. [20] and [30] proposed bias-corrected discriminant functions. [28] studied limiting form of the logistic regression under a so-called infinitely imbalanced case in which the size of one class is fixed and the other grows to infinity. [32] proposed new evaluation criteria and weighted learning procedures that increase the impact of a minority class. [33] developed a distance weighted discrimination method (DWD), originally proposed to overcome the well-known data-piling issue [1] in high-dimensions, by an adaptive weighting scheme to reduce sensitivity to unequal class sizes. [34] proposed a linear classifier that is a hybrid of DWD and SVM, thus haivng advantages of both techniques. [35] introduced a new family of classifiers including SVM and DWD that provides a trade-off between imbalanced and high-dimensionality. [19, 27] theoretically showed that under certain conditions, SVM suffers from data-piling in high-dimensions, meaning that all

data points become the support vectors which may result in ignorance of one of the classes. [27] proposed a biased-corrected SVM that improves its performance even when the class sizes are imbalanced. [26] proposed a robust SVM which is less sensitive to class sizes and choice of a regularization parameter. [43] used a repeated case-control sampling technique coupled with a fused feature screening procedure to deal with imbalanced and high-dimensionality.

The behaviour of LDA in high-dimensional imbalanced classification has often been studied empirically. In this paper, we first theoretically show that in such cases this classifier ignores the minority class, yielding a maximum misclassification rate for this class. On the other hand, a common approach to deal with high-dimensionality is to use a hard-thresholding operator for feature selection. However, our simulations show large differences between the misclassification rates of the hard-thresholding rule (HR) in imbalanced settings. Thus, we face both high-dimensionality and an inflated bias in the difference between the two misclassification rates. To address the issues, we propose a new construction of the HR based a multiple data splitting (Msplit) technique as described below, and thus called Msplit-HR. We randomly split the training data in each class into two parts of sizes $\lfloor n_k/2 \rfloor, k = 1, 2$, and use one part only for feature selection and the other part is then used to construct a bias-corrected classifier based on the selected features. As shown in Section 3, the splitting facilitates the correction of the inflated bias in the difference between the two misclassification rates. To reduce the effect of randomness in single-split, we repeat the process several ($\mathcal{L}$) times which maximizes the usage of training data in finite-sample situations. In general, as pointed out by [25], multiple splitting also helps reproducibility of finite sample results. As shown numerically in Figures 1 and 2, respectively discussed in Sections 3.1 and 3.2, the classification results of Msplit-HR corresponding to $\mathcal{L} \approx 30$ are unsurprisingly more powerful than a single-split ($\mathcal{L} = 1$). We show that our method is asymptotically optimal. We also study asymptotic properties of two well-known linear classifiers, namely the sparse LDA [38], and the regularized optimal affine discriminant analysis [13], under the imbalanced setting. Our simulations show that Msplit-HR either outperforms its competitors or has comparable performance based on a much smaller subset of selected features, while being computationally more efficient as discussed in Section 5.

The rest of the paper is organized as follows. Section 2 gives the problem setup and investigates the behaviour of the LDA in high-dimensional imbalanced binary classification. Section 3 introduces our proposed method, Msplit-HR. Large-sample properties of the method are also discussed in this section. Two well-known high-dimensional variants of the LDA, under the imbalanced setting, are studied in Section 4. The finite-sample performance of several binary classifiers is examined using simulations in Section 5. Analysis of two real data sets are given in Section 6. A summary and discussion are given in Section 7. Technical Lemmas and proofs of our main results are given in Appendices A-C.

Notation: All vectors and matrices are shown in bold letters. For any vector $\mathbf{a} \in \mathbb{R}^p$, $\|\mathbf{a}\|_0 = \#\{j : a_j \neq 0\}$, $\|\mathbf{a}\|_1 = \sum_{j=1}^p |a_j|$, $\|\mathbf{a}\|_2 = (\sum_{j=1}^p a_j^2)^{1/2}$, $\|\mathbf{a}\|_\infty = \max_{j=1,\ldots,p} |a_j|$. For any symmetric matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\|\mathbf{A}\|_1 =$

$\max_{i=1,\ldots,p} \sum_{j=1}^{p} |a_{ij}|$, $\|\mathbf{A}\| = \max_{j=1,\ldots,p} |\lambda_j(\mathbf{A})|$, where $\lambda_j(\mathbf{A})$ are the eigenvalues of the matrix $\mathbf{A}$, and $\|\mathbf{A}\|_\infty = \max_{i,j=1,\ldots,p} |a_{ij}|$. A diagonal matrix is denoted by $\mathbf{D}$. For any two sequences a_n and b_n, we write $a_n \lesssim b_n$ or $a_n = O(b_n)$, if for sufficiently large n there exists a constant C such that $a_n \leq C\, b_n$. We write $a_n \sim b_n$, if $a_n/b_n \to 1$, as $n \to \infty$. And $a_n \asymp b_n$, if $a_n = O(b_n)$ and $b_n = O(a_n)$. Also, $a_n = o(b_n)$, when $a_n/b_n \to 0$ as $n \to \infty$. The notations o_p and O_p are respectively used to indicate convergence and boundedness in probability. An indicator function is denoted by $\mathbf{1}\{\cdot\}$.

2. The LDA

In this section, we first describe the setting of the binary classification problem under our consideration. We then study the effect of dimension and imbalanced class sizes on the LDA, which motives the topics of the remaining sections.

2.1. Overview

We consider the class labels $Y \in \{1, 2\}$, class prior probabilities $\pi_k = \Pr(Y = k)$, and a p-dimensional feature vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)^\top$ such that $\mathbf{X}|Y = k \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), k = 1, 2$. The LDA is a well-known classification technique for this setting. More specifically, given the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ and assuming $\pi_1 = \pi_2$, the optimal rule classifies a subject with an observed feature vector $\mathbf{x}^* = (x_1^*, \ldots, x_p^*)^\top$ to Class 1 if and only if

$$\delta^{\text{opt}}(\mathbf{x}^*; \boldsymbol{\theta}) = \boldsymbol{\mu}_d^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}^* - \boldsymbol{\mu}_a) < 0, \tag{2.1}$$

where $\boldsymbol{\mu}_d = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \neq \mathbf{0}$, $\boldsymbol{\mu}_a = (\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)/2$.

The misclassification rate (MCR) of a classifier is typically used to quantity its performance. The classifier in (2.1) which is the Bayes' rule, is referred to as the optimal rule since it has the smallest average MCR, Π^{opt} in (2.2) below, among all classifiers. For δ^{opt}, the class-specific MCRs are equal and given by

$$\Pi_k = \Pr\left((-1)^k \delta^{\text{opt}}(\mathbf{X}^*; \boldsymbol{\theta}) < 0 \middle| Y = k\right) = \Phi\left(-\Delta_p/2\right) \equiv \Pi^{\text{opt}}, \quad k = 1, 2, \tag{2.2}$$

where Φ is the cumulative distribution function of the standard normal, and $\Delta_p^2 = \boldsymbol{\mu}_d^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_d$ is referred to as the discriminative power or signal value. It is seen that as $\Delta_p \to \infty$, high discriminative power, then $\Pi^{\text{opt}} \to 0$; and as $\Delta_p \to 0$, low discriminative power, then $\Pi^{\text{opt}} \to \frac{1}{2}$ implying that the classifier performs as a random guess. From now on, we assess the performance of other classifiers under consideration by comparing them with the optimal rule.

In practice, the parameter vector $\boldsymbol{\theta}$ is unknown and needs to be estimated using a training data $\mathcal{D}_n = \{\mathbf{x}_{ik},\ i = 1, ..., n_k,\ k = 1, 2\}$, where $\mathbf{x}_{ik}$ is the i-th observed value of $\mathbf{X}$ in Class k, and the n_k are the class sample sizes with the

total sample size $n = n_1 + n_2$. For a new feature vector $\mathbf{x}^*$, a so-called plug-in discriminant function based on the parameter estimates is given by

$$\delta^{\text{LDA}}(\mathbf{x}^*; \hat{\boldsymbol{\theta}}_n) = \hat{\boldsymbol{\mu}}_d^\top \, \widehat{\boldsymbol{\Sigma}}_n^{-1}(\mathbf{x}^* - \hat{\boldsymbol{\mu}}_a), \tag{2.3}$$

where $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\mu}}_{n,1}, \hat{\boldsymbol{\mu}}_{n,2}, \widehat{\boldsymbol{\Sigma}}_n)$ and

$$\hat{\boldsymbol{\mu}}_{n,k} \equiv \hat{\boldsymbol{\mu}}_k \;=\; \frac{1}{n_k}\sum_{i=1}^{n_k} \mathbf{x}_{ik}, \quad k = 1, 2, \tag{2.4}$$

$$\widehat{\boldsymbol{\Sigma}}_n \;=\; \frac{1}{n-2}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(\mathbf{x}_{ik} - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_{ik} - \hat{\boldsymbol{\mu}}_k)^\top. \tag{2.5}$$

The matrix $\widehat{\boldsymbol{\Sigma}}_n^{-1}$ in (2.3) is a generalized inverse when $\widehat{\boldsymbol{\Sigma}}_n$ is not invertible. Given $\mathcal{D}_n$, the conditional MCR of the plug-in linear discriminant rule based on (2.3) corresponding to Class $k \in \{1, 2\}$, is given by

$$\Pi_k^{\text{LDA}}(\mathcal{D}_n) = \Pr\left((-1)^k \delta^{\text{LDA}}(\mathbf{X}^*; \hat{\boldsymbol{\theta}}_n) < 0 \,\middle|\, Y = k,\ \mathcal{D}_n\right) = \Phi\left(\frac{\Psi_k^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)}{\sqrt{\Upsilon^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)}}\right), \tag{2.6}$$

where $\Psi_k^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n) = (-1)^k \hat{\boldsymbol{\mu}}_d^\top \widehat{\boldsymbol{\Sigma}}_n^{-1}(\hat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_k)$, and $\Upsilon^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n) = \hat{\boldsymbol{\mu}}_d^\top \widehat{\boldsymbol{\Sigma}}_n^{-1} \boldsymbol{\Sigma} \widehat{\boldsymbol{\Sigma}}_n^{-1} \hat{\boldsymbol{\mu}}_d$. As is common in the literature, we study large-sample properties of a classifier through its conditional MCR.

2.2. Impact of the dimension and imbalanced class sizes

The effect of the dimension p on the LDA's performance is well studied in the literature. [38] showed that when p is fixed or diverges to infinity at a slower rate than $\sqrt{n}$, the classifier is asymptotically optimal [38, Definition 1]. When $p \to \infty$ such that $p/n \to \infty$, [4], [12], and [38] showed that this classifier performs no better than a random guess. Hence, feature selection is essential when p is large compared to the sample size n.

In the aforementioned works, the impact of dimensionality is studied under particular limiting settings on the class sizes n_1 and n_2. [4] and [38] respectively considered equal class sizes ($n_1 = n_2$) and unequal sizes where $n_1,\ n_2 \to \infty$ such that $\frac{n_2}{n} \to \pi$, $0 < \pi < 1$. [12] developed their results by considering compatible class sizes, such that $c_1 \le \frac{n_1}{n_2} \le c_2$, with $0 < c_1 \le c_2 < \infty$. [3] investigated the case where $n_1,\ n_2 \to \infty$, such that $\frac{n_1 - n_2}{n_1 + n_2} = \rho > 0$ is fixed. All in all, it is seen that the sizes of the two classes grow similarly and proportional to the total sample size n, that is $n_k = O(n), k = 1, 2$. We refer to these settings as a *balanced* classification problem. [28] analyzed the binary logistic regression models with fixed dimension p in a so-called infinitely imbalanced case in which $n_1 \to \infty$ but the class size n_2 is fixed. In this paper, we study imbalanced classification in which n_1 and n_2 grow to infinity such that $n_2 = o(n_1)$, implying a different growth rate of the class sizes.

In the balanced classification, typically average (over the two classes) MCR [38, 13] or the MCR of one arbitrary class [4, 12] is used as a performance measure of a classifier. However, in imbalanced situations due to data scarcity in the minority class, classification results have a tendency to favour the majority class. Thus, the average MCR is not an appropriate performance measure for a classifier $\mathcal{T}$. This motivated us to adapt the optimality definition of a classifier from [38] to our setting as follows.

Definition 1. *Suppose $\mathcal{T}$ is a classifier in a binary classification problem. The misclassification rates of $\mathcal{T}$, given the training data $\mathcal{D}_n$, are denoted by $\Pi_k^{\mathcal{T}}(\mathcal{D}_n), k = 1, 2$. Then,*

(i) $\mathcal{T}$ is asymptotically-strong optimal if $\Pi_k^{\mathcal{T}}(\mathcal{D}_n)/\Pi^{opt} \xrightarrow{p} 1$, $k = 1, 2$,
(ii) $\mathcal{T}$ is asymptotically-strong sub-optimal if $\Pi_k^{\mathcal{T}}(\mathcal{D}_n) - \Pi^{opt} \xrightarrow{p} 0$, $k = 1, 2$,
(iii) $\mathcal{T}$ is asymptotically-strong worst if $\Pi_k^{\mathcal{T}}(\mathcal{D}_n) \xrightarrow{p} \frac{1}{2}$, $k = 1, 2$,
(iv) $\mathcal{T}$ is asymptotically ignorant if $\min_{k=1,2} \Pi_k^{\mathcal{T}}(\mathcal{D}_n) \xrightarrow{p} 0$ and $\max_{k=1,2} \Pi_k^{\mathcal{T}}(\mathcal{D}_n) \xrightarrow{p} 1$.

Note that any classifier $\mathcal{T}$ satisfying either of the properties in parts (i)-(iii) of the above definition also satisfies the properties discussed in the corresponding parts of Definition 1 of [38] for a balanced case, but not vice versa. Part (iv) of the above definition occurs when a classifier completely ignores one of the classes, and more specifically the minority class. We now state our first result.

Theorem 2.1. *Suppose that the estimator $\widehat{\mathbf{\Sigma}}_n^{-1}$ in δ^{LDA} in (2.3) is replaced by $\mathbf{\Sigma}^{-1}$, and $\mathbf{\Sigma}$ is known. When $n_2 = o(n_1)$, such that $p/n_2 \to \infty$ and $\sqrt{\frac{n_2}{p}}\Delta_p^2 = o(1)$, as $n_1, n_2 \to \infty$, then the LDA is asymptotically ignorant, that is,*

$$\Pi_1^{\text{LDA}}(\mathcal{D}_n) \xrightarrow{p} 0 \quad , \quad \Pi_2^{\text{LDA}}(\mathcal{D}_n) \xrightarrow{p} 1.$$

This result implies that in the high-dimensional imbalanced cases, the MCR of the majority class tends to 0 which will be even better than the optimal value Π^{opt}, but the MCR of the minority class approaches 1 which is worse than a random guess. Note that the above result also holds in the case of $p/n_1 \to c$, for some finite constant $c \geq 0$. [19, 27] showed that, under certain conditions, the SVM ignores the minority class in high-dimensional imbalanced problems.

Remark 2.1. *When p is fixed and $n_2 = o(n_1)$, then the LDA is asymptotically-strong optimal.*

Remark 2.1 illustrates that in the fixed-dimensional case, the impact of imbalanced class sizes asymptotically vanishes and $\Pi_k^{\text{LDA}}(\mathcal{D}_n)$, $k = 1, 2$, converge to the optimal value Π^{opt}. Hence, by Theorem 2.1, Shao's results, and Remark 2.1, the effects of both dimension and imbalanced class sizes are responsible for ignorance of the minority class.

3. Proposed Method: Msplit hard-thresholding rule (Msplit-HR)

A common approach to deal with high-dimensionality in the LDA is to incorporate feature selection using a hard-thresholding rule (HR) based on a two-sample t-statistic as in [12]. More specifically, by ignoring the correlation among features, $\mathbf{\Sigma}$ is estimated by the diagonal matrix $\widehat{\mathbf{D}}_n = diag\{\hat{\sigma}_1^2, ..., \hat{\sigma}_p^2\}$, and the discriminant function is given by

$$\delta^{\text{HR}}(\mathbf{x}^*; \hat{\boldsymbol{\theta}}_n) = \sum_{j=1}^{p} r_j(\mathbf{x}^*; \hat{\boldsymbol{\theta}}_n)\ h_j(\hat{\boldsymbol{\theta}}_n), \tag{3.1}$$

where $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \widehat{\mathbf{D}}_n)$, $r_j(\mathbf{x}^*; \hat{\boldsymbol{\theta}}_n) = (\hat{\mu}_{dj}/\hat{\sigma}_j^2)(x_j^* - \hat{\mu}_{aj})$, and $h_j(\hat{\boldsymbol{\theta}}_n) = \mathbf{1}\{|t_j| > \tau_n\}$ is the thresholding operator based on the t-statistic

$$t_j = \frac{\hat{\mu}_{j2} - \hat{\mu}_{j1}}{\hat{\sigma}_j \sqrt{n/n_1 n_2}}. \tag{3.2}$$

Here $\hat{\mu}_{jk}$'s and $\hat{\sigma}_j^2$'s are the entries of $\hat{\boldsymbol{\mu}}_k$ and $\widehat{\mathbf{\Sigma}}_n$ in (2.4) and (2.5), respectively. The discriminant function of FAIR proposed by [12] for balanced problems belongs to the class of functions in (3.1). The authors select an optimal number of statistically most significant features, or equivalently the threshold value τ_n of t-statistic, by minimizing a common upper bound on its corresponding MCRs. However, for the case of general $\mathbf{\Sigma}$, such choice of τ_n does not necessarily result in an asymptotically optimal classifier [38]. Thus, for generality, in the rest of the paper, for any given sequence of τ_n, we refer to a classifier based on (3.1) as an HR unless otherwise is specified.

If indeed $\mathbf{\Sigma} = \mathbf{D}$, [3] showed that the HR in (3.1) based on a fixed threshold $\tau_n = \tau$, is asymptotically ignorant when $\rho = (n_1 - n_2)/(n_1 + n_2) > 0$ is fixed, as $n_1, n_2 \to \infty$. As stated after Theorem 3.1 below, it is interesting to note that under the imbalanced setting $n_2 = o(n_1)$ and by an appropriate choice of τ_n, the HR is indeed asymptotically-strong optimal. However, our simulations in Section 5 show an unsatisfactory finite-sample performance of the HR in the sense of both the MCR in the minority class and large difference between the two MCRs. We propose a new construction of the HR which outperforms (3.1) in finite-samples, while maintaining the same desirable large-sample properties, to be discussed below. To fix ideas, we first consider the imbalanced problem with a diagonal $\mathbf{\Sigma} = \mathbf{D}$. The general case of a non-diagonal $\mathbf{\Sigma}$ is discussed in Subsection 3.2, which is based on a feature screening technique. Note that under this case, the HR based on (3.1) is not optimal, as it ignores the correlation among the features.

3.1. Msplit-HR under a diagonal $\mathbf{\Sigma}$

As discussed in Section 2, the class specific MCRs of the optimal rule are equal, and are given in (2.2). Our numerical experiments show that, due to the imbalanced class sizes, HR performs well in majority class but underperforms in

minority class, though it has large-sample optimal property as discussed after Theorem 3.1 below. Thus, the idea in our work is to reduce the difference between two conditional MCRs of HR toward that of the optimal rule which is zero. More specifically, our main goal is to propose a new discriminant function aiming to reduce the difference between the MCRs of the HR,

$$|\Pi_1^{\text{HR}}(\mathcal{D}_n) - \Pi_2^{\text{HR}}(\mathcal{D}_n)| = |\Phi(\psi_{1,n}) - \Phi(\psi_{2,n})|,$$

where $\psi_{k,n} = \Psi_k^{\text{HR}}(\hat{\boldsymbol{\theta}}_n)/\sqrt{\Upsilon^{\text{HR}}(\hat{\boldsymbol{\theta}}_n)}$ and

$$\Psi_k^{\text{HR}}(\hat{\boldsymbol{\theta}}_n) = (-1)^{k+1} \sum_{j=1}^{p} r_j(\boldsymbol{\mu}_k; \hat{\boldsymbol{\theta}}_n) h_j(\hat{\boldsymbol{\theta}}_n) \ , \ \Upsilon^{\text{HR}}(\hat{\boldsymbol{\theta}}_n) = \sum_{j=1}^{p} (\hat{\mu}_{dj}/\hat{\sigma}_j^2)^2 \sigma_j^2 h_j(\hat{\boldsymbol{\theta}}_n)$$

for $k = 1, 2$. To understand the above difference, [3] studied distributional properties of the quantities $\psi_{k,n}, k = 1, 2$. They focused on reducing the so-called bias

$$B_n^{\text{HR}} = \mathbb{E}\big\{\Psi_1^{\text{HR}}(\hat{\boldsymbol{\theta}}_n) - \Psi_2^{\text{HR}}(\hat{\boldsymbol{\theta}}_n)\big\} \tag{3.3}$$

to zero, which results in decreasing the bias between $\psi_{1,n}$ and $\psi_{2,n}$ and consequently of that between MCRs. However, it turns out that due to the dependency between the random variables r_j and h_j, computing B_n^{HR} is not an easy task. [3] studied the origin of the bias and proposed methods for its correction. We instead propose a new construction of the HR that facilities the computation of such bias by adapting a sample-splitting strategy as follows.

The training sample of each class is randomly partitioned into two subsamples of sizes $n_k' = \lfloor n_k/2 \rfloor$. The two sub-samples are used for computing two quantities similar to the r_j and h_j in (3.1), for each $j = 1, \ldots, p$, and then the results are merged. To reduce the effect of randomness due to the data splitting, this process is repeated, say, $\mathcal{L}$ times. Our new discrimination function is then constructed by averaging over the HR-type discriminant functions based on each splitting. Thus, we chose the name Msplit-HR for our method. More specifically, at the ℓ-th data splitting, for each $\ell = 1, ..., \mathcal{L}$, the entire training data $\mathcal{D}_n$ is partitioned into two parts $\mathcal{D}_{n,\ell}^{(1)}$ and $\mathcal{D}_{n,\ell}^{(2)}$. The parameter estimates based on each sub-sample are distinguished by the superscripts $^{(1)}$ and $^{(2)}$, that is, $\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)}$ and $\hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}$. A new observation with a feature vector $\mathbf{x}^*$ is then classified using the discriminant function

$$\delta_0^{\text{Msplit-HR}}(\mathbf{x}^*; \hat{\boldsymbol{\theta}}_n) = \frac{1}{\mathcal{L}} \sum_{\ell=1}^{\mathcal{L}} \sum_{j=1}^{p} r_j(\mathbf{x}^*; \hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}) \ h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)}), \tag{3.4}$$

where $\hat{\boldsymbol{\theta}}_n = \{(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)}, \hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}) : \text{ for } \ell = 1, \ldots, \mathcal{L}\}$. Due to the statistical independence of the two random functions h_j and r_j in (3.4), for all $j = 1, \ldots, p$, calculation of the bias B_n for $\delta_0^{\text{Msplit-HR}}$ is straight forward, which is shown below. Recall $n_k' = \lfloor n_k/2 \rfloor$, and let $n' = n_1' + n_2'$ and $f_{n'} = n'/2 - 1$.

Proposition 3.1. *The bias B_n in (3.3) corresponding to $\delta_0^{Msplit\text{-HR}}$ is given by*

$$B_{0,n}^{Msplit\text{-HR}} = \mathbb{E}\big\{\Psi_{0,1}^{Msplit\text{-HR}}(\hat{\boldsymbol{\theta}}_n) - \Psi_{0,2}^{Msplit\text{-HR}}(\hat{\boldsymbol{\theta}}_n)\big\} = \frac{\bar{r}_n}{\mathcal{L}} \sum_{\ell=1}^{\mathcal{L}} \sum_{j=1}^{p} \mathbb{E}\{h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)})\},$$

where $\bar{r}_n = f_{n'}(\frac{1}{n_1'} - \frac{1}{n_2'})\frac{\Gamma(f_{n'}-1)}{\Gamma(f_{n'})}$, *and* $\Gamma(\cdot)$ *is the gamma function.*

Finally, using the above result, we propose the bias-corrected discriminant function

$$\delta^{\text{Msplit-HR}}(\mathbf{x}^*;\hat{\boldsymbol{\theta}}_n) = \frac{1}{\mathcal{L}}\sum_{\ell=1}^{\mathcal{L}}\sum_{j=1}^{p}\left\{r_j(\mathbf{x}^*;\hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}) - \frac{\bar{r}_n}{2}\right\}h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)}) \tag{3.5}$$

which has its bias $B_n^{\text{Msplit-HR}} = 0$. The term $\bar{r}_n$ is a function of $(n_2' - n_1')$ which is negative since $n_2 < n_1$. Hence, for any new feature vector $\mathbf{x}^*$, the resulting discriminant function (3.5) tends to be more positive compared to the rule in (3.4). This increases the chance (or probability) of classifying a new observation to the minority class, and hence improving the classification results for this class. In our simulations and the real-data analysis, we evaluate the performance of Msplit-HR based on the bias corrected function $\delta^{\text{Msplit-HR}}$. We now describe Algorithm 1 that summarizes the steps for computing (3.5).

Algorithm 1 : Computing the discriminant function $\delta^{\text{Msplit-HR}}$.

Require: Input $n_1' = \lfloor n_1/2 \rfloor, n_2' = \lfloor n_2/2 \rfloor, \mathbf{x}^*, \mathcal{L}, \bar{r}_n$ and τ_n.

1: **for** $\ell = 1, \ldots, \mathcal{L}$ **do**
2: Split $\mathcal{D}_n$ into $\mathcal{D}_{n,\ell}^{(1)}$ and $\mathcal{D}_{n,\ell}^{(2)}$
3: **for** $j = 1, \ldots, p$ **do**
4: Step1: Using $\mathcal{D}_{n,\ell}^{(1)}$ compute $h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)})$
5: Step2: Using $\mathcal{D}_{n,\ell}^{(2)}$ compute $r_j(\mathbf{x}^*;\hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}) - \frac{\bar{r}_n}{2}$
6: **end for**
7: **end for**
8: **return** $\delta^{\text{Msplit-HR}}(\mathbf{x}^*;\hat{\boldsymbol{\theta}}_n) = \frac{1}{\mathcal{L}}\sum_{\ell=1}^{\mathcal{L}}\sum_{j=1}^{p}\left\{r_j(\mathbf{x}^*;\hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}) - \frac{\bar{r}_n}{2}\right\}h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)})$.

In practice, a value of $\mathcal{L}$ is required to compute (3.5). Figure 1 shows the class-specific MCRs of (3.5) as a function of $\mathcal{L}$, corresponding to scenario **(i)** in our simulations in Section 5.1. It can be seen that a value of $\mathcal{L}$ between 20 to 30 provides a satisfactory performance of Msplit-HR. We used $\mathcal{L} = 30$ in our numerical experiments.

The following results show the asymptotic behaviour of $\delta^{\text{Msplit-HR}}$. First, we state Lemma 3.1 that provides conditions under which the t-statistic (3.2) used in the thresholding operator h_j in $\delta^{\text{Msplit-HR}}$ selects all the important features. Since $\mathcal{L}$ is fixed, the result of the lemma holds for all $\ell = 1, \ldots, \mathcal{L}$.

Lemma 3.1. *Assume that the mean difference vector* $\boldsymbol{\mu}_d = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ *is sparse. Let* $\mathcal{S} = \{j : \mu_{dj} \neq 0\}$ *be the the corresponding active set with the cardinality* $s = |\mathcal{S}|$, *and define* $d_{0,n} = \min_{j\in\mathcal{S}}|\mu_{dj}|$. *Under Conditions (C1) and (C2) in Appendix A, if* $\tau_n = O(\sqrt{n_2}d_{0,n})$, $\log s = o(n_2 d_{0,n}^2)$, $\log(p-s) = o(\tau_n^2)$, $n_2 = o(n_1)$, *and* $\sqrt{n_2}d_{0,n} \to \infty$, *as* $n_1, n_2 \to \infty$, *then*

$$(a)\ \Pr\left(\bigcap_{j\notin\mathcal{S}}\{|t_j| \leq \tau_n\}\right) \to 1; \quad (b)\quad \Pr\left(\bigcap_{j\in\mathcal{S}}\{|t_j| > \tau_n\}\right) \to 1.$$

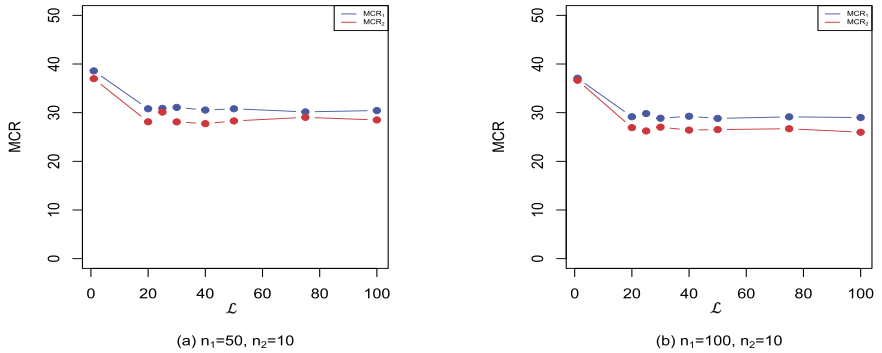

FIG 1. *Effect of the number of sample-splits* $\mathcal{L}$ *on Msplit-*HR *performance for the Simulation setting* ***(i)*** *and* $p = 1000$.

In the above Lemma, if $d_{0,n} = d_0$, for some constant $d_0 > 0$, then $\tau_n = O(\sqrt{n_2})$ and $\log p = o(n_2)$. On the other hand, if $d_{0,n} \sim n_2^{-\gamma}\alpha_{n_2}$, for $0 < \gamma < 1$ and some $\alpha_{n_2} \to \infty$, such that $d_{0,n}$ declines to zero and $\sqrt{n_2}d_{0,n} \to \infty$, then we have $\tau_n = O(n_2^{1/2-\gamma}\alpha_{n_2})$ and $\log p = o(n_2^{1-2\gamma}\alpha_{n_2}^2)$. Therefore, in both cases the divergence rate of the dimension p is smaller than that of the minority class size n_2, as opposed to the balanced case where $\log p = o(n)$, that is, a larger dimension p allowance.

Theorem 3.1. *Suppose that the conditions of Lemma 3.1 are satisfied. Let* $\kappa_n = \max\{\Delta_p^{-1}\sqrt{s/n_2}\ ,\ \sqrt{\log p/n_1}\}$. *For any fixed* $\mathcal{L}$,

*(a) the MCRs of Msplit-*HR *are given by*

$$\Pi_k^{Msplit\text{-HR}}(\mathcal{D}_n) = \Phi\Big(-\frac{1}{2}\Delta_p(1 + O_p(\kappa_n))\Big), \quad k = 1, 2$$

(b) if $s\Delta_p^2 = o(n_2)$ *and* $\Delta_p^2\sqrt{\log p/n_1} = o(1)$, *the Msplit-*HR *is asymptotically-strong optimal.*

Note that the result of Theorem 3.1 also holds for the HR. Part (b) of the theorem implies that the growth rates of both the sparsity size s and the discriminative power Δ_p are controlled by the minority class size n_2.

3.2. Msplit-HR under a general $\boldsymbol{\Sigma}$

When the dimension p is large compared to the sample size n, the sample covariance matrix in (2.5) is ill-conditioned. To deal with the singularity issue, many existing methods in the literature involve a feature selection strategy. In what follows, we use a variable screening method [14, 29] to select a subset of features x_j's that have the highest discriminative power.

At the ℓ-th data splitting stage of Msplit-HR, we consider the mean difference estimators $\hat{\boldsymbol{\mu}}_{d,\ell}^{(1)} = \hat{\boldsymbol{\mu}}_{2,\ell}^{(1)} - \hat{\boldsymbol{\mu}}_{1,\ell}^{(1)}$, which are computed based on the training sub-samples $\mathcal{D}_{n,\ell}^{(1)}$, for $\ell = 1, ..., \mathcal{L}$. For a given threshold parameter τ_n, we select those features x_j whose indices belong to the set $\mathcal{S}_{n,\ell}^{(1)} = \{1 \le j \le p : |\hat{\mu}_{dj,\ell}^{(1)}| > \tau_n\}$, where $\hat{\mu}_{dj,\ell}^{(1)}$ is the j-th entry of $\hat{\boldsymbol{\mu}}_{d,\ell}^{(1)}$.

For any p-dimensional feature vector $\mathbf{x}^*$, we define the discriminant function

$$\delta_0^{\text{Msplit-HR}}(\mathbf{x}^*; \hat{\boldsymbol{\theta}}_n) = \frac{1}{\mathcal{L}} \sum_{\ell=1}^{\mathcal{L}} \tilde{\boldsymbol{\mu}}_{d,\ell}^\top \widetilde{\boldsymbol{\Sigma}}_{n,\ell}^{-1} \, (\mathbf{x}_\ell^* - \tilde{\boldsymbol{\mu}}_{a,\ell}), \tag{3.6}$$

where $\hat{\boldsymbol{\theta}}_n$ is the vector of corresponding parameter estimates, and $\mathbf{x}_\ell^* = (x_j^* : j \in \mathcal{S}_{n,\ell}^{(1)})^\top$ are sub-vectors of the full feature vector $\mathbf{x}^*$. Furthermore, for all $\ell = 1, \ldots, \mathcal{L}$, we have $\tilde{\boldsymbol{\mu}}_{d,\ell} = \tilde{\boldsymbol{\mu}}_{2,\ell} - \tilde{\boldsymbol{\mu}}_{1,\ell}$, $\tilde{\boldsymbol{\mu}}_{a,\ell} = (\tilde{\boldsymbol{\mu}}_{1,\ell} + \tilde{\boldsymbol{\mu}}_{2,\ell})/2$, such that $\tilde{\boldsymbol{\mu}}_{k,\ell} = (\hat{\mu}_{jk,\ell}^{(2)} : j \in \mathcal{S}_{n,\ell}^{(1)})^\top$ for $k = 1, 2$, and $\widetilde{\boldsymbol{\Sigma}}_{n,\ell} = [\hat{\sigma}_{jj',\ell}^{(2)} : j, j' \in \mathcal{S}_{n,\ell}^{(1)}]$ are respectively the sub-vectors and sub-matrices of the sample means and covariance matrix given in (2.4) and (2.5). Note that for the existence of $\widetilde{\boldsymbol{\Sigma}}_{n,\ell}^{-1}$, for all $\ell = 1, 2, \ldots, \mathcal{L}$, we include at most $(n' - 2)$ features in each $\mathcal{S}_{n,\ell}^{(1)}$.

As discussed in Subsection 3.1, the data splitting technique facilitates computation of the bias B_n (3.3) corresponding to (3.6).

Proposition 3.2. *If* $|\mathcal{S}_{n,\ell}^{(1)}| < n' - 3$, *for all* $\ell = 1, .., \mathcal{L}$, *then*

$$B_{0,n}^{\textit{Msplit-HR}} = \mathbb{E}\{\Psi_{0,1}^{\textit{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) - \Psi_{0,2}^{\textit{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n)\} = \frac{1}{\mathcal{L}} \sum_{\ell=1}^{\mathcal{L}} \mathbb{E}\{\bar{r}_{n,\ell}\},$$

where

$$\bar{r}_{n,\ell} = \left(\frac{1}{n_1'} - \frac{1}{n_2'}\right) \frac{n' - 2}{n' - 3 - |\mathcal{S}_{n,\ell}^{(1)}|} \times |\mathcal{S}_{n,\ell}^{(1)}|. \tag{3.7}$$

Finally, our bias-corrected discriminant function is

$$\delta^{\text{Msplit-HR}}(\mathbf{x}^*; \hat{\boldsymbol{\theta}}_n) = \frac{1}{\mathcal{L}} \sum_{\ell=1}^{\mathcal{L}} \{\tilde{\boldsymbol{\mu}}_{d,\ell}^\top \tilde{\boldsymbol{\Sigma}}_{n,\ell}^{-1} (\mathbf{x}_\ell^* - \tilde{\boldsymbol{\mu}}_{a,\ell}) - \frac{\bar{r}_{n,\ell}}{2}\} \tag{3.8}$$

which has its bias $B_n^{\text{Msplit-HR}} = 0$. The term $\bar{r}_{n,\ell}$ as a function of $(n_2' - n_1')$ makes the corrected discriminant function (3.8) more positive compared to the rule in (3.6). This increases the probability of classifying a new observation to the minority class, and hence improving the results for this class. Algorithm 2 below summarize the steps for computing in (3.8).

Figure 2 shows the class-specific MCRs of (3.8) as a function of $\mathcal{L}$, corresponding to scenario **(iv)** in our simulations in Section 5.2. Based on these results, we used $\mathcal{L} = 30$ in our numerical experiments.

The following lemma shows that the variable screening method used to obtain the selection sets $\mathcal{S}_{n,t}^{(1)}$ have a so-called strong screening consistency property, as

Algorithm 2 Computing the discriminant function in (3.8)

Require: Input $n_1' = \lfloor n_1/2 \rfloor, n_2' = \lfloor n_2/2 \rfloor, \mathbf{x}^*, \mathcal{L}$, and τ_n.

for $\ell = 1, \ldots, \mathcal{L}$ **do**
  Split $\mathcal{D}_n$ into $\mathcal{D}_{n,\ell}^{(1)}$ and $\mathcal{D}_{n,\ell}^{(2)}$
  Using $\mathcal{D}_{n,\ell}^{(1)}$, obtain $\mathcal{S}_{n,\ell}^{(1)} = \{1 \le j \le p : |\hat{\mu}_{dj,\ell}^{(1)}| > \tau_n\}$ and compute $\bar{r}_{n,\ell}$ in (3.7)
  if $|\mathcal{S}_{n,\ell}^{(1)}| < n_1' + n_2' - 3$ **then**
    Using $\mathcal{S}_{n,\ell}^{(1)}$ and $\mathcal{D}_{n,\ell}^{(2)}$, compute $\tilde{\boldsymbol{\mu}}_{d,\ell}^\top \tilde{\boldsymbol{\Sigma}}_{n,\ell}^{-1}(\mathbf{x}_\ell^* - \tilde{\boldsymbol{\mu}}_{a,\ell})$
  else
    Step 1: Select the first $(n_1' + n_2' - 4)$ features in $\mathcal{S}_{n,\ell}^{(1)}$ with highest value of $|\hat{\mu}_{dj,\ell}^{(1)}|$
    Step 2: Using $\mathcal{D}_{n,\ell}^{(2)}$ and the selected features in Step 1, compute $\tilde{\boldsymbol{\mu}}_{d,\ell}^\top \tilde{\boldsymbol{\Sigma}}_{n,\ell}^{-1}(\mathbf{x}_\ell^* - \tilde{\boldsymbol{\mu}}_{a,\ell})$
  end if
end for
return $\delta^{\text{Msplit-HR}}(\mathbf{x}^*; \hat{\boldsymbol{\theta}}_n) = \frac{1}{\mathcal{L}} \sum_{\ell=1}^{\mathcal{L}} \{\tilde{\boldsymbol{\mu}}_{d,\ell}^\top \tilde{\boldsymbol{\Sigma}}_{n,\ell}^{-1}(\mathbf{x}_\ell^* - \tilde{\boldsymbol{\mu}}_{a,\ell}) - \frac{\bar{r}_{n,\ell}}{2}\}$.

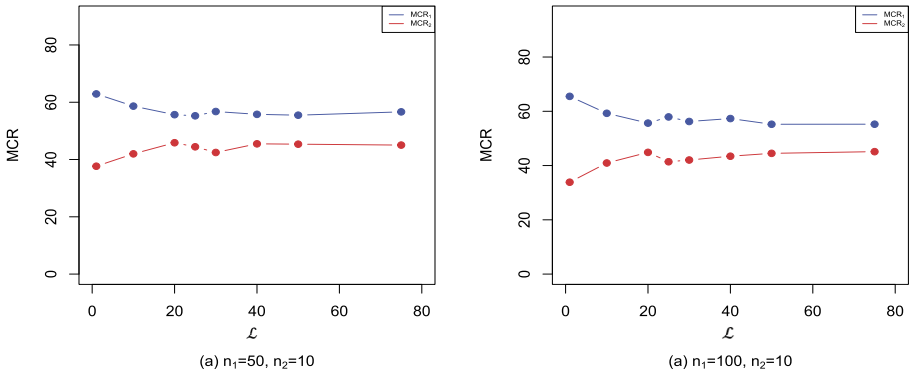

FIG 2. *Effect of the number of sample-splits $\mathcal{L}$ on Msplit-HR performance for the Simulation setting **(iv)** and $p = 500$.*

discussed in [29]. We then establish the asymptotic optimality of $\delta^{\text{Msplit-HR}}$ in Theorem 3.2.

Lemma 3.2. *Let $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$, and define the active set $\mathcal{S} = \{1 \le j \le p : \beta_j \neq 0\}$ with its cardinality denoted by $|\mathcal{S}|$. Furthermore, let $d_{0,n} = \min_{j \in \mathcal{S}} |\mu_{dj}|$ and $m_{\max} = c_1(\max_{j\in\mathcal{S}} \beta_j^2)|\mathcal{S}|/d_{0,n}^2$, for some constant $c_1 > 0$ such that $m_{\max} \ge |\mathcal{S}|$. Under Condition (C2) in Appendix A, if $\tau_n \asymp d_{0,n}$, $\log p = o(n_2 d_{0,n}^2)$, $n_2 = o(n_1)$, and $\sqrt{n_2} d_{0,n} \to \infty$, as $n_1, n_2 \to \infty$, for any $\ell = 1, ..., \mathcal{L}$, we have that*

$$(a)\ \Pr\left(\mathcal{S}_{n,\ell}^{(1)} \supset \mathcal{S}\right) \to 1 \quad ; \quad (b)\ \Pr\left(|\mathcal{S}_{n,\ell}^{(1)}| \le m_{\max}\right) \to 1.$$

Part *(a)* implies that that for large sample sizes n, with probability tending to one, all the active features will be included in the selection sets $\mathcal{S}_{n,\ell}^{(1)}$, for each $\ell = 1, 2, \ldots, \mathcal{L}$. Part *(b)* shows that the size of each set $\mathcal{S}_{n,\ell}^{(1)}$ is of order $m_{\max}$. These properties are obtained under the conditions that the divergence rate of

the dimension p is lower than that of the minority class size n_2.

Theorem 3.2. *Suppose that the conditions of Lemma 3.2 are satisfied. Let* $\kappa'_n = \max\{\Delta_p^{-1}\sqrt{m_{\max}/n_2}\ ,\ m_{\max}\sqrt{\log p/n_1}\}$. *If* $m_{\max}\sqrt{\log p/n_1} = o(1)$, *then for any fixed* $\mathcal{L}$,

*(a) the MCRs of Msplit-*HR *are given by*

$$\Pi_k^{Msplit\text{-HR}}(\mathcal{D}_n) = \Phi\bigg(-\frac{1}{2}\Delta_p(1+O_p(\kappa'_n))\bigg), \quad k=1,2$$

(b) if $\Delta_p^2 m_{\max} = o(n_2)$ *and* $\Delta_p^2 m_{\max}\sqrt{\log p/n_1} = o(1)$, *then the Msplit-*HR *is asymptotically-strong optimal.*

Condition $\Delta_p^2 m_{\max} = o(n_2)$ in the above theorem implies that the maximum size of the selection sets $\mathcal{S}_{n,\ell}$, that is $m_{\max}$, is affected by the minority class size n_2. Note that the results of the theorem also holds for the pairwise sure independence screening of [29] in the imbalanced binary cases, as well as in the balanced cases which was not studied before.

4. Two existing high-dimensional variants of LDA

In this section, we investigate conditions under which two well-known sparse variants of the LDA obtain certain optimality properties under the imbalanced setting.

4.1. Sparse LDA (SLDA)

This method, proposed by [38], uses thresholding-type estimators for both the mean-difference vector $\boldsymbol{\mu}_d = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}$. In SLDA, a new feature vector $\mathbf{x}^*$ is allocated to Class 1 if and only if

$$\delta^{\text{SLDA}}(\mathbf{x}^*;\hat{\boldsymbol{\theta}}_n) = \tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}(\mathbf{x}^* - \hat{\boldsymbol{\mu}}_a) < 0,$$

where $\hat{\boldsymbol{\mu}}_a = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2$, and $(\widetilde{\boldsymbol{\Sigma}}_n, \tilde{\boldsymbol{\mu}}_d)$ are thresholded estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_d$, respectively, with the entries,

$$\tilde{\sigma}_{ij} \;=\; (1-2/n)\ \hat{\sigma}_{ij}\ \mathbf{1}\{(1-2/n)|\hat{\sigma}_{ij}| > t_n\},\ i,j=1,\ldots,p$$

$$\tilde{\mu}_{dj} \;=\; \hat{\mu}_{dj}\ \mathbf{1}\{|\hat{\mu}_{dj}| > a_n\},\ j=1,\ldots,p,$$

where $\hat{\sigma}_{ij}$ is (i,j)-th element of $\widehat{\boldsymbol{\Sigma}}_n$ in (2.5), and $\hat{\mu}_{dj}$ is the j-th entry of $\hat{\boldsymbol{\mu}}_d$ in (2.4). Further, $t_n = M_1\sqrt{\log p/n}$ with $M_1 > 0$, and $a_n = M_2(\log p/n)^\alpha$, $0 < \alpha < 1/2, M_2 > 0$.

[38] derived conditions under which the SLDA is optimal according to their Definition 1, when $p/n \to \infty$ and $n_1/n \to \pi$ with $0 < \pi < 1$, as $n \to \infty$. It

turns out that their conditions do not yield an optimal SLDA in the imbalanced case. In Theorem 4.1 below, we investigate conditions under which the SLDA is asymptotically-strong optimal under the imbalanced case. We then discuss and compare these conditions with those of [38] under the balanced case.

First, for ease of comparison, we recall some notations introduced in [38]. Let $\hat{q}_n$ be the number of features for which the value $|\hat{\mu}_{dj}|$ is greater than a_n. Further, let q_{n0} and q_n be the number of features for which the value of $|\mu_{dj}|$ is greater than ra_n and a_n/r, respectively, for some fixed constant $r > 1$. Also let $D_{g,p} = \sum_{j=1}^{p} \mu_{dj}^{2g}$, $0 \le g < 1$, and $C_{h,p} = \max_{1\le i\le p} \sum_{j=1}^{p} |\sigma_{ij}|^h$, $0 \le h < 1$, be the sparsity measures corresponding to $\boldsymbol{\mu}_d$ and $\boldsymbol{\Sigma}$, respectively. Here, 0^0 is defined to be 0. Furthermore, let $d_{n_1} = C_{h,p}({n_1}^{-1}\log p)^{(1-h)/2}$, and

$$b_{n_1} = \Delta_p^{-1} \max\left\{\Delta_p d_{n_1},\ \sqrt{a_n^{2(1-g)} D_{g,p}},\ \sqrt{q_n/n_2},\ \sqrt{C_{h,p} q_n/n_1}\right\},$$

$$b_{n_2} = \Delta_p^{-1} \max\left\{\Delta_p d_{n_1},\ \sqrt{a_n^{2(1-g)} D_{g,p}},\ \sqrt{C_{h,p} q_n/n_2}\right\},$$

where $\Delta_p^2 = \boldsymbol{\mu}_d^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_d$. Note that under the imbalanced setting $n_2 = o(n_1)$, we have $d_{n_1} \sim d_n$, where $d_n = C_{h,p}(n^{-1}\log p)^{(1-h)/2}$.

The following Lemma shows that the set $\{1 \le j \le p : |\hat{\mu}_{dj}| > a_n\}$ has indeed a sure screening property, which is essential in Theorem 4.1 for the assessment of SLDA.

Lemma 4.1. *Suppose that,*

$$(\log p)\,(n_1/\log p)^{2\alpha} = o(n_2), \tag{4.1}$$

and $n_2 = o(n_1)$, then as $n_1, n_2 \to \infty$,

(a) $\Pr\left(\bigcap_{j:|\mu_{dj}|>ra_n} \{|\hat{\mu}_{dj}| > a_n\}\right) \to 1,$

(b) $\Pr\left(\bigcap_{j:|\mu_{dj}|\le a_n/r} \{|\hat{\mu}_{dj}| \le a_n\}\right) \to 1,$

(c) $\Pr\left(q_{n0} \le \hat{q}_n \le q_n\right) \to 1.$

Condition (4.1) replaces the condition $\log p/n = o(1)$ in [38]. One implication of (4.1) is $\log p/n_2 = o(1)$, which shows the impact of the minority class size n_2 on the dimension allowance p.

Theorem 4.1. *Suppose that the conditions of Lemma 4.1, and Conditions (C2) and (C3) in Appendix A are satisfied. Then, as $n_1, n_2 \to \infty$,*

(a) the MCRs of SLDA *are given by*

$$\Pi_k^{\text{SLDA}}(\mathcal{D}_n) = \Phi\left(-\frac{1}{2}\Delta_p\{1 + O_p(b_{n_k})\}\right), \quad k = 1, 2.$$

(b) the SLDA *is asymptotically-strong optimal if*
i. Δ_p^2 is bounded, and $b_{n_2} = o(1)$, or
ii. $\Delta_p^2 \to \infty$, such that $\Delta_p^2 b_{n_2} = o(1)$ holds.

The difference between the above theorem and Theorem 3 of [38] appears in b_{n_2}. To simplify the comparison in this case as in [38], suppose that $\mathbf{\Sigma}$ is a diagonal matrix ($C_{0,p} = 1$), and let s be the number of nonzero (active) entries of the mean difference vector $\boldsymbol{\mu}_d$. If there are two constant $c_1, c_2 > 0$, such that $c_1 \leq |\mu_{dj}| \leq c_2$, for the active j's, then we have $q_n = s$. This implies that, by the Conditions (C2) and (C3), Δ_p^2 and $D_{0,p}$ are of order s. Now, in this case, if $s \to \infty$, according to Theorem 4.1-(b)-ii above, under condition (4.1), $\Delta_p^2 b_{n_2} = o(1)$ is equivalent to $s = o((n_1/\log p)^\alpha)$. This implies that under the imbalanced setting, the growth rate of the sparsity factor s is smaller than $\sqrt{n_2}$ and consequently is smaller than the growth rate of s in the balanced setting. Therefore, due to the data scarcity in the minority class (n_2) in the imbalanced setting, in order for the SLDA to be asymptotically-strong optimal more restrictive conditions are required on both the dimension p and the sparsity size s compared to the balanced case.

Next, we compare the optimality conditions of Msplit-HR and SLDA. The relation between these conditions for a general $\mathbf{\Sigma}$ is not straightforward, and thus to get some insight we consider a diagonal case. Suppose that $\mathbf{\Sigma}$ is diagonal ($C_{0,p} = 1$), and $g = 0$ such that $D_{0,p} = s = |\mathcal{S}|$, where $\mathcal{S} = \{1 \leq j \leq p : \mu_{dj} \neq 0\}$. By condition (4.1), we have $\log p = o(n_2)$ which implies the necessary conditions of Lemma 3.1 on (s, p), if $d_{0,n} = \min_{j\in\mathcal{S}} |\mu_{dj}| = d_0 > 0$ and $\tau_n = M\sqrt{n_2}$, for some constant $M > 0$. On the other hand, if $d_{0,n}$ decays, the same conclusion holds when $a_n = O(d_{0,n})$ and $\tau_n = M\sqrt{n_2}d_{0,n}$. Furthermore, by (4.1) the conditions of Theorem 4.1-(b) are equivalent to $s\Delta_p^2(\log p/n_1)^{2\alpha} = o(1)$ implying $s\Delta_p^2 = o(n_2)$ which is required for the optimality of Msplit-HR. Therefore, the conditions of Theorem 4.1 for SLDA on the dimension p and the sparsity size s are more restrictive than those in Theorem 3.1 for Msplit-HR. In terms of feature selection, Lemma 3.2-(b) provides an upper bound $m_{\max} = o(n_2\Delta_p^{-2})$ on the size of the set of selected features by Msplit-HR, whereas the SLDA allows the number of nonzero estimators of μ_{dj}'s or $\sigma_{l,j}$'s to be much larger than the class sizes to ensure optimality of the classifier, see [38]. Therefore, the number of selected features by SLDA could be potentially larger than the class sizes which we have also observed in our numerical study in Section 5.

4.2. Regularized optimal affine discriminant (ROAD)

This method, proposed by [13], is constructed based on a sparse estimate of $\mathbf{w} = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d$, unlike the SLDA which uses sparse estimates of $\boldsymbol{\mu}_d$ and $\mathbf{\Sigma}$, separately. The ROAD assigns $\mathbf{x}^*$ to Class 1 if and only if

$$\delta^{\text{ROAD}}(\mathbf{x}^*; \hat{\boldsymbol{\theta}}_n, c) = \hat{\mathbf{w}}_c^\top(\mathbf{x}^* - \hat{\boldsymbol{\mu}}_a) < 0, \tag{4.2}$$

where $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \widehat{\mathbf{\Sigma}}_n)$, $\hat{\boldsymbol{\mu}}_a = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2$, and

$$\hat{\mathbf{w}}_c \in \arg \min_{\|\mathbf{w}\|_1 \leq c,\ \mathbf{w}^\top\hat{\boldsymbol{\mu}}_d = 1} \mathbf{w}^\top \widehat{\mathbf{\Sigma}}_n \mathbf{w} \tag{4.3}$$

with $\hat{\boldsymbol{\mu}}_d = \hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1$, and $(\hat{\boldsymbol{\mu}}_k, \widehat{\mathbf{\Sigma}}_n)$ are the estimates in (2.4)-(2.5). Note that in (4.3) the smaller the c, the sparser the solution $\hat{\mathbf{w}}_c$, and as $c \to \infty$ the solution

is equivalent to the regular weight $\mathbf{w}_c \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$. [13] studied the asymptotic difference between the average MCR of the ROAD and its oracle version for which the true values of $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ are used in (4.3). However, as discussed in Section 2.2, under the imbalanced setting the average MCR is not an appropriate performance measure for a classifier. Therefore, in the following theorem, we study the class-wise MCRs of the ROAD.

Theorem 4.2. *Let* $s_c = \|\boldsymbol{w}_c\|_0$, $s_c^{(1)} = \|\boldsymbol{w}_c^{(1)}\|_0$ *and* $\hat{s}_c = \|\hat{\boldsymbol{w}}_c\|_0$, *where* $\boldsymbol{w}_c$, $\boldsymbol{w}_c^{(1)}$, *and* $\hat{\boldsymbol{w}}_c$ *are respectively the solutions of* (4.3) *when* $(\boldsymbol{\mu}_d, \boldsymbol{\Sigma})$, $(\hat{\boldsymbol{\mu}}_d, \boldsymbol{\Sigma})$ *and* $(\hat{\boldsymbol{\mu}}_d, \widehat{\boldsymbol{\Sigma}}_n)$ *are used. Furthermore, let* $\Pi_k^{\text{ROAD}}(\mathcal{D}_n; c)$ *be the MCR of Class* $k = 1, 2$, *associated with* ROAD, *and* $\Pi_k^{orc}(c)$ *denotes its oracle value. Under Condition (C2) in Appendix A, if* $n_2 = o(n_1)$ *and* $\log p = o(n_2)$, *then as* $n_1, n_2 \to \infty$,

$$\Pi_k^{\text{ROAD}}(\mathcal{D}_n; c) - \Pi_k^{orc}(c) = O_p(e_n), \quad k = 1, 2, \tag{4.4}$$

where $e_n = \max\left\{c^2(\log p)/n_1 \ , \ \sqrt{(\log p)/n_2} \times \sqrt{\max\{s_c, \ s_c^{(1)}, \ \hat{s}_c\}}\right\}$.

By Theorem 4.2, a necessary condition for convergency of the MCRs of ROAD to their oracle values is that the sparsity size s_c of the vector $\mathbf{w}_c$ and the dimension p are controlled by the minority class size n_2 (similar to the SLDA), which in turn shows the effect of imbalanced class sizes on the performance of ROAD.

In general, the conditions of Theorem 4.2 do not guarantee the optimality of ROAD according to Definition 1. [13] showed that when the penalty parameter c is chosen as $c \geq \Delta_p^{-2} \parallel \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d \parallel_1$, then $\mathbf{w}_c \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$ and the oracle MCRs $\Pi_k^{orc}(c)$ reduce to those of the optimal rule in (2.2). Hence, by Definition 1, for such c's, Theorem 4.2 shows that ROAD is asymptotically-strong sub-optimal as long as $e_n \to 0$. Furthermore, ROAD becomes asymptotically-strong optimal if Δ_p is bounded. The condition $e_n \to 0$ is the same as $\log p = o(n_1/c^2)$ and $\log p = o(n_2/s_{\max})$, where $s_{\max} = \max\{s_c, \ s_c^{(1)}, \ \hat{s}_c\}$. Note that, the larger the c, the larger the quantities s_c, $\hat{s}_c$ and $s_c^{(1)}$, and hence more restrictions on (n_1, n_2, p) compared to those in Theorem 4.2, and the conditions of Msplit-HR. In our numerical study, we observe that the performance of ROAD in terms of MCR_2 improves for lower dimensions.

5. Simulation study

In this section, we assess the finite-sample performance of Msplit-HR and several binary classification methods using simulations. We consider two settings of diagonal and general covariance matrix $\boldsymbol{\Sigma}$ under the model $\mathbf{X}|Y = k \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, $k = 1, 2$.

5.1. Diagonal $\boldsymbol{\Sigma}$

We compare the following methods: the bias adjusted independence (BAI) and leave-one-out independence rules (LOUI) [3], diagonal ROAD method (DROAD)

[13], the bias corrected LDA (BLDA) [20], the HR and its under-sampling version (US-HR), and our proposed method Msplit-HR. Note that the aforementioned methods use the knowledge of a diagonal $\mathbf{\Sigma}$. In our comparison, we also include a bias-corrected support vector machines proposed by [27] coupled with an under-sampling method (US-BCSVM). In regards to over-sampling techniques such as the SOMTE, [3] and [8] showed that such techniques deduce larger differences between the MCRs in high-dimensional imbalanced problems. For example, we examined the performance of HR and BCSVM coupled with SMOTE (under both diagonal and general $\mathbf{\Sigma}$) and since their performances were not satisfactory, we did not report the results here.

We implemented the methods using R software. The DROAD results are based on the authors' MATLAB codes available on their website [1]. Our computations are carried out on a computer with an AMD Opteron(tm) Processor 6174 CPU 2.2GHz.

The above methods involve certain tuning (threshold) parameters that need to be chosen using data-driven methods. We chose best threshold parameters in BLDA, BAI and LOUI by a grid search using the techniques outlined by the authors. As in [20], an F-statistic is used to select the important features in BLDA method. In both HR and Msplit-HR, we choose the tuning parameter τ by minimizing MCR of the minority class based on a leave-one-out cross validation.

We consider the binary classification problem $\mathbf{X}|(Y=k) \sim N_p(\boldsymbol{\mu}_k, \mathbf{D}), k = 1, 2$, and $\mathbf{D} = \text{diag}\{\sigma_1^2, ..., \sigma_p^2\}$. We generated training data with different class sizes n_1 and n_2, and test data sets of size 50 in both classes. We considered two dimensions $p = 1000, 3000$, and class-wise sample sizes $(n_1, n_2) = (25, 5)$, $(50, 10)$, $(100, 10)$ for the training data. The simulation results are based on 100 randomly generated data sets, and the two parameter settings:

- **(i)** $\boldsymbol{\mu}_1 = (1, 1, \mathbf{0}_{p-2})^\top$, $\boldsymbol{\mu}_2 = (2, 2.2, \mathbf{0}_{p-2})^\top$, $\sigma_1^2 = 1.5^2$, $\sigma_2^2 = 0.75^2$, and $\sigma_j^2 = 1$, for $j = 3, ..., p$.
- **(ii)** $\boldsymbol{\mu}_1 = (\mathbf{1}_9, \mathbf{0}_{p-9})^\top$, $\boldsymbol{\mu}_2 = (2 * \mathbf{1}_4, 2.5 * \mathbf{1}_3, 3 * \mathbf{1}_2, \mathbf{0}_{p-9})^\top$, $\sigma_j^2 = 10$, for $j = 1, ..., 4$, $\sigma_j^2 = 2.25^2$, for $j = 5, 6, 7$, $\sigma_j^2 = 1.5^2$, $j = 8, 9$, and $\sigma_j^2 = 1$, for $j = 10, ..., p$.

The number (s) of active features x_j's that distinguish the two classes, and also the value of Δ_p in the two settings are respectively $s = 2, \Delta_p^2 = 3$ and $s = 9, \Delta_p^2 = 8.7$. Since the signal strength is measured by Δ_p, setting **(i)** has a weaker signal than **(ii)**. Under these settings, the value of the optimal MCR, Π^{opt} in (2.2), are respectively 19.32% and 7%. Also, the active features have different marginal signal values $|\mu_{dj}|/\sigma_j$, in each of the settings.

The performance measures used to compare different methods are: per-class misclassification rates (MCR_1, MCR_2), and the geometric mean (GM) of the MCRs. The results reported in the tables are average and standard deviations (in parentheses) of the measures over 100 generated samples. We also reported median number of true selected features, denoted by A, and falsely selected features denoted by N, respectively. For the new method Msplit-HR, similar to

[1] <https://github.com/statcodes/ROAD>

the stability selection technique of [24], the selected features for each simulated sample are those with a relative frequency more than 50%, that is the set $\mathcal{S}_n = \{j : \frac{f_j}{\mathcal{L}} \geq 0.5\}$, where f_j is selection frequency of j-th feature among $\mathcal{L}$ splits.

5.1.1. Discussion of the results

The results for the cases $(n_1, n_2, p) = (25, 5, 1000)$, $(50, 10, 1000)$ and $(100, 10, 1000)$ are given in Table 1. The results corresponding to dimension $p = 3000$ are given in Table 2.

From Table 1, under both settings **(i)** and **(ii)**, we can see that DROAD, HR, and BLDA have smaller error rates in the majority class (MCR_1) compared to the other methods, but the differences between their MCR_1 and MCR_2 are larger. The class-wise error rates corresponding to US-HR and US-BCSVM have smaller differences than those of DROAD, HR, and BLDA. Furthermore, the US-HR outperforms US-BCSVM, DROAD, HR, and BLDA in terms of MCR_2. Under setting **(i)**, Msplit-HR outperforms all the other methods in terms of MCR_2; for example, its MCR_2 is better than the next best method LOUI up to about 8%, depending on class sizes (n_1, n_2) and dimension p, while having balanced results for both classes. In setting **(ii)**, Msplit-HR behaves similarly to LOUI and BAI, with its MCR_2 better than LOUI and BAI respectively up to about 3% and 7%. Note that in **(i)**, we have a weaker signal strength (Δ_p^2) and fewer number of active features (s) than **(ii)**, which matches the conditions of Theorem 3.1 for Msplit-HR on controlling the size of $s\Delta_p^2$. In other words, we can see that the weaker the signal, the better the performance of Msplit-HR in terms of MCRs in both classes. On the other hand, from the columns A and N of Table 1, Msplit-HR tends to select fewer number of inactive (noise) features compared to the two its competitors BAI and LOUI. In BCSVM, the bias caused by dimension is corrected by using all features in the model and therefore this method does not perform any feature selection.

Table 2 consists of the results for dimension $p = 3000$. As expected, the class-specific MCRs of all the methods increase compared to $p = 1000$. Msplit-HR outperforms all the other techniques in terms of MCR_2 while having balanced misclassification rates. For example, the MCR_2 of Msplit-HR is smaller than the next best method LOUI up to about 7%. In addition, we observe that Msplit-HR has better performance than BAI and LOUI even in setting **(ii)** in which they have comparable performance for $p = 1000$.

We now assess the computational efficiency of the different methods. For a fixed threshold, the computational complexity of BAI and LOUI is $O(n^2p)$ and that of all the other methods is $O(np)$. In our simulations, the threshold (or tuning) parameter in each method was chosen using a cross validation criterion. Table 3 provides the average computational time (in seconds) taken by each method to complete per-sample results. Note that since US-BCSVM does not involve any feature selection step, as expected, this method is among the faster methods discussed here. It can be seen that the HR and BLDA, followed by US-HR and US-BCSVM, are the fastest among all the methods we considered, but they

TABLE 1
*Classification results for the simulation settings **(i)-(ii)** with a diagonal $\mathbf{\Sigma}$ and $p = 1000$.*

(n_1, n_2)	Setting	Methods	$MCR_1\%$	$MCR_2\%$	$GM\%$	A	N
(25,5)	**(i)**	US-BCSVM	$48.96_{(13.99)}$	$47.46_{(14.38)}$	$46.33_{(5.99)}$	2	998
		DROAD	$2.62_{(7.45)}$	$93.14_{(15.91)}$	$5.45_{(11.37)}$	2	364
		HR	$15.96_{(13.02)}$	$67.3_{(26.43)}$	$26.25_{(15.45)}$	1	2
		US-HR	$44.34_{(16.45)}$	$45.14_{(16.49)}$	$42.73_{(8.91)}$	1	143.5
		BLDA	$14.72_{(11.05)}$	$70.16_{(22.47)}$	$28.04_{(11.75)}$	1	5
		BAI	$38.86_{(15.16)}$	$48.2_{(17.23)}$	$41.15_{(10.06)}$	1	75.5
		LOUI	$41.46_{(18.14)}$	$43.9_{(19.43)}$	$39.05_{(11.75)}$	1	20.5
		Msplit-HR	$42.66_{(16.97)}$	$40.04_{(16.65)}$	$39.12_{(11.01)}$	1	2
(25,5)	**(ii)**	US-BCSVM	$46.42_{(14.04)}$	$41.22_{(12.27)}$	$41.99_{(5.75)}$	9	991
		DROAD	$5.46_{(8.63)}$	$58.48_{(30.27)}$	$9.84_{(10.24)}$	6	17
		HR	$12.38_{(10.52)}$	$55.78_{(27.55)}$	$20.80_{(12.40)}$	1	2
		US-HR	$39.34_{(14.75)}$	$35.48_{(15.43)}$	$35.13_{(9.11)}$	4	124
		BLDA	$11.06_{(8.74)}$	$57.48_{(26.60)}$	$20.85_{(11.30)}$	2	3
		BAI	$30.72_{(13.94)}$	$35.06_{(16.32)}$	$30.53_{(10.12)}$	3	33
		LOUI	$29.86_{(14.54)}$	$31.24_{(16.37)}$	$27.95_{(10.83)}$	3	36.5
		Msplit-HR	$32.2_{(15.57)}$	$28.22_{(15.44)}$	$27.61_{(9.91)}$	1	3.5
(50,10)	**(i)**	US-BCSVM	$47.88_{(10.19)}$	$44.56_{(10.78)}$	$45.25_{(5.71)}$	2	998
		DROAD	$6.30_{(9.10)}$	$75.28_{(30.25)}$	$11.23_{(11.61)}$	2	68
		HR	$19.36_{(7.47)}$	$40.82_{(20.99)}$	$26.15_{(7.68)}$	1	1
		US-HR	$34.22_{(14.05)}$	$32.66_{(14.34)}$	$32.12_{(10.72)}$	1	0
		BLDA	$18.26_{(8.82)}$	$48.68_{(21.24)}$	$25.27_{(8.81)}$	1	3
		BAI	$31.94_{(14.39)}$	$36.92_{(15.19)}$	$32.71_{(10.52)}$	1	11
		LOUI	$29.28_{(12.21)}$	$34.12_{(17.04)}$	$29.99_{(10.39)}$	1	8.5
		Msplit-HR	$30.22_{(12.66)}$	$26.68_{(13.42)}$	$26.99_{(9.75)}$	1	0
(50,10)	**(ii)**	US-BCSVM	$41.72_{(9.03)}$	$38.02_{(10.21)}$	$38.93_{(5.23)}$	9	991
		DROAD	$5.60_{(6.04)}$	$30.72_{(19.67)}$	$9.39_{(6.22)}$	7	17.5
		HR	$11.02_{(7.04)}$	$25.42_{(15.69)}$	$14.59_{(6.84)}$	2	0
		US-HR	$22.84_{(10.14)}$	$19.04_{(8.49)}$	$19.74_{(7.23)}$	1	0
		BLDA	$11.72_{(6.70)}$	$24.36_{(15.71)}$	$14.80_{(6.13)}$	2	0
		BAI	$17.6_{(8.76)}$	$19.8_{(11.79)}$	$17.18_{(8.07)}$	3	3
		LOUI	$16.72_{(8.55)}$	$19.16_{(10.99)}$	$16.55_{(7.39)}$	3	3.5
		Msplit-HR	$19.22_{(9.58)}$	$17.82_{(9.07)}$	$17.08_{(6.84)}$	2	0
(100,10)	**(i)**	US-BCSVM	$47.96_{(10.08)}$	$44.1_{(10.50)}$	$45.09_{(5.42)}$	2	998
		DROAD	$2.60_{(5.25)}$	$85.74_{(22.67)}$	$6.28_{(9.67)}$	2	494
		HR	$19.96_{(8.53)}$	$34.82_{(19.40)}$	$24.31_{(8.09)}$	1	0
		US-HR	$34.08_{(13.17)}$	$30.52_{(13.07)}$	$31.14_{(9.29)}$	1	0
		BLDA	$16.84_{(7.60)}$	$45.48_{(22.81)}$	$25.08_{(8.01)}$	1	2
		BAI	$28.86_{(12.15)}$	$33.64_{(16.62)}$	$29.72_{(9.85)}$	1	7
		LOUI	$26.26_{(11.03)}$	$32.48_{(16.76)}$	$27.81_{(9.26)}$	1	6
		Msplit-HR	$27.94_{(12.09)}$	$24.84_{(13.11)}$	$24.95_{(8.93)}$	1	0
(100,10)	**(ii)**	US-BCSVM	$41.66_{(9.67)}$	$37.38_{(10.88)}$	$38.51_{(6.02)}$	9	991
		DROAD	$3.22_{(4.23)}$	$37.96_{(20.38)}$	$6.57_{(6.07)}$	8	31.5
		HR	$10.02_{(6.20)}$	$22.14_{(12.49)}$	$13.11_{(5.77)}$	3	0
		US-HR	$20.64_{(10.65)}$	$18.98_{(10.47)}$	$18.30_{(7.76)}$	1	0
		BLDA	$10.44_{(6.26)}$	$22.28_{(14.29)}$	$12.96_{(5.83)}$	3	0
		BAI	$16.44_{(9.70)}$	$17.08_{(10.19)}$	$15.49_{(7.89)}$	3.5	2
		LOUI	$15.02_{(8.37)}$	$15.94_{(9.32)}$	$14.13_{(6.42)}$	3	2
		Msplit-HR	$16.56_{(8.98)}$	$14.38_{(7.88)}$	$14.04_{(5.36)}$	3	0

TABLE 2
*Classification results for Simulation settings **(i)-(ii)** with a diagonal $\boldsymbol{\Sigma}$ and $p = 3000$.*

(n_1, n_2)	Setting	Methods	$MCR_1\%$	$MCR_2\%$	$GM\%$	A	N
(25,5)	**(i)**	US-BCSVM	$50.84_{(13.78)}$	$47.32_{(13.75)}$	$47.32_{(5.03)}$	2	2998
		DROAD	$3.14_{(7.66)}$	$94.40_{(12.91)}$	$6.71_{(13.44)}$	1	529
		HR	$14.92_{(12.68)}$	$73.64_{(24.84)}$	$26.10_{(16.85)}$	0	1.5
		US-HR	$46.38_{(16.21)}$	$46.86_{(16.10)}$	$44.25_{(7.23)}$	1	255
		BLDA	$13.6_{(11.63)}$	$76.44_{(23.64)}$	$26.03_{(14.53)}$	1	5
		BAI	$38.04_{(15.86)}$	$50.92_{(17.61)}$	$41.77_{(9.48)}$	1	107.5
		LOUI	$41.02_{(18.00)}$	$47.04_{(18.60)}$	$41.16_{(10.20)}$	1	42
		Msplit-HR	$43.06_{(19.61)}$	$44.08_{(19.25)}$	$40.09_{(9.18)}$	1	3
(25,5)	**(ii)**	US-BCSVM	$47.12_{(13.88)}$	$45.32_{(13.29)}$	$44.37_{(5.11)}$	9	2991
		DROAD	$5.54_{(8.59)}$	$60.58_{(28.94)}$	$10.24_{(11.52)}$	6	16
		HR	$14.04_{(12.40)}$	$62.44_{(27.65)}$	$23.41_{(14.78)}$	1	1
		US-HR	$43.78_{(15.90)}$	$41.52_{(15.33)}$	$40.35_{(8.22)}$	3	251.5
		BLDA	$11.04_{(10.41)}$	$65.64_{(27.20)}$	$20.50_{(13.70)}$	1	3
		BAI	$33.04_{(15.35)}$	$41.68_{(17.60)}$	$34.84_{(10.61)}$	3	79.5
		LOUI	$31.64_{(16.28)}$	$41.84_{(18.91)}$	$33.63_{(10.90)}$	3	78
		Msplit-HR	$37.78_{(17.62)}$	$36.32_{(17.68)}$	$34.18_{(10.36)}$	1	3
(50,10)	**(i)**	US-BCSVM	$48.46_{(11.56)}$	$47.98_{(11.55)}$	$47.04_{(5.29)}$	2	998
		DROAD	$5.7_{(9.27)}$	$81.02_{(24.78)}$	$11.01_{(13.54)}$	2	77
		HR	$18.68_{(8.29)}$	$40.88_{(24.32)}$	$24.61_{(9.61)}$	1	0
		US-HR	$35.62_{(13.25)}$	$36.66_{(14.57)}$	$34.87_{(10.01)}$	1	0
		BLDA	$17.16_{(8.35)}$	$48.18_{(24.08)}$	$25.74_{(8.52)}$	1	2
		BAI	$32.08_{(11.86)}$	$37.42_{(17.28)}$	$33.32_{(11.03)}$	1	12.5
		LOUI	$31.1_{(12.33)}$	$34.82_{(17.41)}$	$31.48_{(11.18)}$	1	9.5
		Msplit-HR	$32.3_{(12.81)}$	$30.98_{(16.05)}$	$30.35_{(11.34)}$	1	0
(50,10)	**(ii)**	US-BCSVM	$44.08_{(10.75)}$	$44.16_{(10.42)}$	$43.04_{(4.85)}$	9	991
		DROAD	$5.12_{(5.21)}$	$32.40_{(17.42)}$	$9.24_{(6.25)}$	1	25.50
		HR	$12.98_{(8.14)}$	$28.7_{(18.25)}$	$17.14_{(8.18)}$	2	0
		US-HR	$26.8_{(12.13)}$	$24.72_{(12.17)}$	$24.48_{(8.81)}$	1	0
		BLDA	$12.7_{(6.88)}$	$29.1_{(19.32)}$	$16.70_{(7.20)}$	2	1
		BAI	$20.24_{(10.71)}$	$22.44_{(13.68)}$	$19.55_{(9.60)}$	3	4
		LOUI	$19.02_{(10.22)}$	$22.74_{(13.48)}$	$19.45_{(9.35)}$	3	9
		Msplit-HR	$21.4_{(11.07)}$	$19.2_{(10.14)}$	$18.93_{(7.47)}$	2	0
(100,10)	**(i)**	US-BCSVM	$48.3_{(10.85)}$	$48.58_{(11.83)}$	$47.32_{(5.50)}$	2	998
		DROAD	$1.80_{(4.23)}$	$88.66_{(20.70)}$	$4.49_{(8.31)}$	2	861.50
		HR	$18.42_{(8.24)}$	$42.38_{(24.65)}$	$25.08_{(9.02)}$	1	0.50
		US-HR	$36.94_{(14.20)}$	$37.1_{(13.93)}$	$35.91_{(10.47)}$	1	0
		BLDA	$15.64_{(8.56)}$	$50.18_{(26.11)}$	$24.00_{(9.47)}$	1	2
		BAI	$29.92_{(10.65)}$	$39.64_{(16.46)}$	$33.37_{(10.74)}$	1	14.5
		LOUI	$27.04_{(10.72)}$	$36.04_{(17.31)}$	$29.95_{(10.71)}$	1	7
		Msplit-HR	$31.46_{(11.87)}$	$29.1_{(15.09)}$	$29.14_{(11.13)}$	1	0
(100,10)	**(ii)**	US-BCSVM	$44.52_{(10.96)}$	$44.74_{(10.91)}$	$43.52_{(5.09)}$	9	991
		DROAD	$3.28_{(4.47)}$	$38.90_{(20.50)}$	$6.74_{(6.38)}$	1	31.5
		HR	$10.18_{(6.08)}$	$27.96_{(18.17)}$	$14.07_{(6.17)}$	2	0
		US-HR	$24.28_{(11.97)}$	$24.48_{(12.20)}$	$22.97_{(8.76)}$	1	0
		BLDA	$10.06_{(6.06)}$	$28.26_{(18.13)}$	$14.25_{(6.17)}$	2	1
		BAI	$17.32_{(9.01)}$	$20.94_{(14.02)}$	$17.39_{(8.37)}$	3	3
		LOUI	$16.08_{(8.97)}$	$21.32_{(13.90)}$	$17.06_{(8.32)}$	3	3
		Msplit-HR	$18.64_{(9.41)}$	$18.04_{(11.09)}$	$16.84_{(8.28)}$	2	0

are outperformed by the other methods in terms of the error rate in the minority class. In addition, while BAI and LOUI's performances in terms of the error rates in the minority class are comparable to our proposed method Msplit-HR; the former are slower in terms of computational time.

TABLE 3
Average computational time (in seconds) taken by a method to complete per-sample results: Simulation setting ***(i)****.*

(n_1, n_2, p)	US-BCSVM	DROAD	HR	US-HR	BLDA	BAI	LOUI	Msplit-HR
(25,5,1000)	2.8	21.73	0.9	4.66	1.05	6	6.39	9.27
(50,10,1000)	5.12	30.77	1.47	19.98	3.53	58	260	92
(100,10,1000)	4.76	35.00	5.43	42.22	11.20	421	365	185
(25,5,3000)	7.5	97.58	1.13	9.05	1.75	14.72	13.83	19.63
(50,10,3000)	12.38	146.17	4.40	62.54	12.24	225	219	282
(100,10,3000)	10.67	141.82	19.90	169.97	29.34	1517	2294	1200

5.2. General $\boldsymbol{\Sigma}$

We considered the same binary classification problem as in Section 5.1, i.e. $\mathbf{X}|(Y = k) \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), k = 1, 2$, but with a general non-diagonal $\boldsymbol{\Sigma}$. We generated training data with different class sizes n_1 and n_2, and test data sets of sizes 50 in both classes. The simulation results are based on 100 randomly generated data sets. The parameter settings are:

(iii) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2^\top = (1, 0.5 * \mathbf{1}_5^\top, 0.1 * \mathbf{1}_5^\top, \mathbf{0}_{p-11}^\top)$, $(\boldsymbol{\Sigma})_{ij} = 0.8$, for $i \neq j$, $(\boldsymbol{\Sigma})_{ii} = 4$, for $i = 1, ..., p$ and $\Delta_p^2 = 0.71$.

(iv) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2^\top = (1, \mathbf{0}_4^\top, 0.1, \mathbf{0}_{p-6}^\top)$, $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & & \mathbf{0} \\ & \boldsymbol{\Sigma}_2 & \\ \mathbf{0} & & \ddots \end{bmatrix}$, where $(\boldsymbol{\Sigma}_1)_{ij} = 0.3$, and $(\boldsymbol{\Sigma}_2)_{ij} = 0.8$, for $i \neq j$, $(\boldsymbol{\Sigma}_1)_{ii} = (\boldsymbol{\Sigma}_2)_{ii} = 1$, for $i = 1, ..., 5$ and $\Delta_p^2 = 1.27$.

In what follows, using the same performance measures described in Section 5.1, we compare these methods: FAIR, SLDA, ROAD, Msplit-HR, a binary version of the pairwise sure independent screening (PSIS) method by [29], bias adjusted ROAD (BA-ROAD) and leave-one-out ROAD (LOU-ROAD) by [3], and US-BCSVM mentioned in Section 5.1. For the FAIR, ROAD, BA-ROAD, LOU-ROAD, we used the techniques based on cross-validation described in the related papers for selecting tuning parameters. We applied the bi-section method of [22] for tuning parameter selection in SLDA by minimizing the MCR of the minority class (called $\text{SLDA}_{\text{MCR}_2}$, in the tables).

All the aforementioned methods provide sparse estimates, say $\hat{\boldsymbol{\beta}}$, of the vector $\boldsymbol{\beta} = (\beta_j : 1 \leq j \leq p)^\top = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$ by either plugging in particular sparse estimates of $\boldsymbol{\mu}_d$ and $\boldsymbol{\Sigma}$, or by directly finding sparse estimate of $\boldsymbol{\beta}$. Thus, in our simulation results for each method, we also report the number of j's for which $\hat{\beta}_j \neq 0$, denoted by S in the tables. For Msplit-HR, we report the cardinality of the set $\mathcal{S}_n = \{1 \leq j \leq p : \frac{f_j}{\mathcal{L}} \geq 0.5\}$, where f_j is the selection frequency

corresponding to index j over the splits $\ell = 1, ..., \mathcal{L}$. Table 4 contains the simulation results for $(n_1, n_2, p) = (25, 5, 200)$, $(50, 10, 200)$ and $(100, 10, 200)$, and the results for the dimension $p = 500$ are given in the Table 5.

5.2.1. Discussion of the results

From Tables 4 and 5, under both settings **(iii)** and **(iv)**, we can see that FAIR, SLDA, PSIS and ROAD tend to classify more observations to the majority class, and resulting in large differences between the two MCRs. Overall, the techniques US-BCSVM, BA-ROAD, LOU-ROAD and Msplit-HR perform better than FAIR, SLDA, PSIS and ROAD in terms of MCR_2 and the geometric mean. For the setting **(iii)**, in the case $(n_1, n_2) = (25, 5)$, Msplit-HR outperforms others, and in the cases, $(n_1, n_2) = (50, 10)$ and $(100, 10)$, the US-BCSVM and LOU-ROAD have better performance than others; for example, when $(n_1, n_2) = (100, 10)$, LOU-ROAD outperforms Msplit-HR about 4%. For the setting **(iv)**, Msplit-HR outperforms all the other techniques in terms of MCR_2; for example outperforms BC-SVM and LOU-ROAD respectively up to about 10% and 12% depending on the values of (n_1, n_2, p). Moreover, this performance of Msplit-HR is based on a much smaller set of selected features compared to its competitors. In summary, Msplit-HR has better performance in the setting **(iv)** which includes more features with weak signals than **(iii)**.

Next, we assess the computational efficiency of different methods by studying the average computational time (in seconds) taken by each method to complete per-sample results, which are given in Table 6. We can see that PSIS is the fastest method followed by FAIR and US-BCSVM. However, as seen above, these methods do not perform well in terms of the MCRs. As mentioned before, US-BCSVM is computationally fast, since it does not involve any feature selection step. The SLDA is slower than the Msplit-HR when the dimension p is increased from $p = 200$ to 500. On the other hand, Msplit-HR is computationally more efficient than its two competitors BA-ROAD and LOU-ROAD. Note that, for a fixed value of tuning parameter, the computational complexity of BA-ROAD and LOU-ROAD is $O(n^2p^2)$, and that of Msplit-HR is $O(np^2)$. Therefore, even without a tuning selection procedure, our technique has lower computational cost.

In summary, given the difficulty of the imbalanced problem, our current simulation study shows that (considering all the three factors: misclassification rates, feature selection, and computational efficiency) Msplit-HR has a good performance compared to the methods discussed here, and is yet another reliable technique for high-dimensional imbalanced problems.

6. Real-data analysis

We now demonstrate the performance of different methods on two real data sets. [2]

[2]Both data sets are publicly available from the R package *datamicroarray* [37], and are available at <https://github.com/>.

TABLE 4
*Classification results for the simulation settings **(iii)-(iv)** with a general $\mathbf{\Sigma}$ and $p = 200$.*

(n_1, n_2)	Setting	Methods	$MCR_1\%$	$MCR_2\%$	$GM\%$	S
(25,5)	**(iii)**	US-BCSVM	$46.26_{(15.97)}$	$51.22_{(15.50)}$	$46.20_{(6.30)}$	200
		FAIR	$23.22_{(9.64)}$	$78.56_{(10.26)}$	$40.53_{(8.04)}$	6.87
		$SLDA_{MCR_2}$	$42.04_{(14.38)}$	$57.38_{(14.87)}$	$47.02_{(6.55)}$	147.07
		PSIS	$31.56_{(9.44)}$	$66.98_{(10.51)}$	$45.02_{(6.21)}$	1
		ROAD	$15.47_{(8.03)}$	$82.67_{(8.87)}$	$34.16_{(6.89)}$	26.17
		BA-ROAD	$48.20_{(12.51)}$	$48.91_{(12.36)}$	$46.83_{(4.55)}$	56.45
		LOU-ROAD	$48.07_{(12.27)}$	$49.03_{(12.40)}$	$46.97_{(2.55)}$	54.09
		Msplit-HR	$53.58_{(15.06)}$	$45.76_{(16.23)}$	$47.12_{(6.65)}$	4
(25,5)	**(iv)**	US-BCSVM	$49.96_{(15.76)}$	$46_{(15.44)}$	$45.39_{(5.62)}$	200
		FAIR	$20.96_{(8.00)}$	$76.38_{(10.82)}$	$38.83_{(6.85)}$	8.07
		$SLDA_{MCR_2}$	$37.18_{(14.30)}$	$60.74_{(14.84)}$	$45.31_{(7.84)}$	124.39
		PSIS	$30.02_{(9.15)}$	$62.52_{(16.16)}$	$42.17_{(8.64)}$	1
		ROAD	$15.77_{(7.24)}$	$78.93_{(11.66)}$	$33.94_{(6.09)}$	24.38
		BA-ROAD	$46.52_{(15.06)}$	$46.63_{(16.65)}$	$43.92_{(7.17)}$	48.53
		LOU-ROAD	$46.34_{(15.79)}$	$46.22_{(17.97)}$	$43.17_{(7.63)}$	47.26
		Msplit-HR	$52.32_{(18.66)}$	$43.02_{(18.51)}$	$43.77_{(7.92)}$	4.5
(50,10)	**(iii)**	US-BCSVM	$46.34_{(11.67)}$	$48.88_{(12.77)}$	$46.27_{(5.25)}$	200
		FAIR	$28.98_{(8.96)}$	$69.1_{(8.95)}$	$43.96_{(6.76)}$	6
		$SLDA_{MCR_2}$	$44.5_{(11.65)}$	$55.84_{(12.40)}$	$48.57_{(5.37)}$	195
		PSIS	$37.96_{(8.26)}$	$60.72_{(8.67)}$	$47.47_{(5.54)}$	1
		ROAD	$19.98_{(8.79)}$	$77.54_{(9.02)}$	$38.07_{(7.11)}$	44
		BA-ROAD	$48.36_{(14.40)}$	$48.50_{(14.15)}$	$45.99_{(9.29)}$	53.50
		LOU-ROAD	$47.38_{(11.90)}$	$49.14_{(12.19)}$	$46.99_{(6.34)}$	53
		Msplit-HR	$50.76_{(13.85)}$	$47.64_{(12.15)}$	$47.65_{(6.11)}$	3
(50,10)	**(iv)**	US-BCSVM	$47.88_{(11.98)}$	$48.42_{(12.55)}$	$46.76_{(5.53)}$	200
		FAIR	$23.98_{(8.65)}$	$64.5_{(12.53)}$	$38.33_{(7.57)}$	6
		$SLDA_{MCR_2}$	$37.5_{(13.23)}$	$54.58_{(16.77)}$	$43.61_{(9.53)}$	189.5
		PSIS	$32.16_{(8.79)}$	$51.04_{(17.23)}$	$39.64_{(9.48)}$	1
		ROAD	$21.68_{(9.25)}$	$64.44_{(18.71)}$	$35.45_{(7.40)}$	19.50
		BA-ROAD	$37.82_{(13.06)}$	$44.74_{(15.80)}$	$39.00_{(10.23)}$	26
		LOU-ROAD	$39.74_{(12.51)}$	$42.46_{(13.50)}$	$39.80_{(8.27)}$	27
		Msplit-HR	$44.62_{(13.96)}$	$40.1_{(14.25)}$	$40.77_{(8.48)}$	1
(100,10)	**(iii)**	US-BCSVM	$46.54_{(11.31)}$	$48.46_{(12.50)}$	$46.19_{(5.04)}$	200
		FAIR	$26.08_{(7.63)}$	$70.62_{(7.94)}$	$42.27_{(6.27)}$	6.68
		$SLDA_{MCR_2}$	$47.24_{(13.75)}$	$54.18_{(12.48)}$	$49.09_{(6.46)}$	169.26
		PSIS	$36.06_{(8.41)}$	$63_{(8.95)}$	$46.52_{(6.62)}$	1.01
		ROAD	$11.16_{(6.67)}$	$86.04_{(8.46)}$	$29.49_{(7.64)}$	71.70
		BA-ROAD	$44.50_{(13.57)}$	$50.94_{(13.67)}$	$45.40_{(8.71)}$	85.41
		LOU-ROAD	$44.62_{(10.18)}$	$42.44_{(9.37)}$	$42.79_{(6.12)}$	66.22
		Msplit-HR	$50.42_{(15.15)}$	$46.46_{(14.43)}$	$46.22_{(6.99)}$	11.45
(100,10)	**(iv)**	US-BCSVM	$47.76_{(12.15)}$	$47.26_{(12.50)}$	$46.03_{(5.56)}$	200
		FAIR	$22_{(7.97)}$	$67.54_{(11.80)}$	$37.63_{(7.61)}$	8.03
		$SLDA_{MCR_2}$	$34.2_{(11.71)}$	$50.54_{(17.12)}$	$40.03_{(8.67)}$	96.77
		PSIS	$31.36_{(8.13)}$	$47.58_{(18.24)}$	$37.74_{(9.75)}$	1
		ROAD	$11.16_{(6.67)}$	$86.04_{(8.46)}$	$29.49_{(7.64)}$	71.70
		BA-ROAD	$44.50_{(13.57)}$	$50.96_{(13.67)}$	$45.39_{(8.71)}$	85.41
		LOU-ROAD	$44.12_{(12.27)}$	$50.54_{(11.72)}$	$45.84_{(5.82)}$	97
		Msplit-HR	$45.86_{(17.28)}$	$37.6_{(14.92)}$	$39.19_{(8.68)}$	9.25

TABLE 5
*Classification results for the simulation settings **(iii)**-**(iv)** with a general $\mathbf{\Sigma}$ and $p = 500$.*

(n_1, n_2)	Setting	Methods	$MCR_1\%$	$MCR_2\%$	$GM\%$	S
(25,5)	**(iii)**	US-BCSVM	$49.18_{(17.81)}$	$50.46_{(17.69)}$	$46.39_{(8.47)}$	500
		FAIR	$20.48_{(8.72)}$	$79_{(8.91)}$	$38.79_{(8.29)}$	8.91
		$\text{SLDA}_{\text{MCR}_2}$	$44.84_{(15.58)}$	$54.56_{(16.11)}$	$46.97_{(5.57)}$	350.18
		PSIS	$23.22_{(9.64)}$	$75.68_{(10.26)}$	$40.53_{(8.05)}$	6
		ROAD	$12.01_{(8.27)}$	$87.16_{(8.70)}$	$30.21_{(8.08)}$	30.79
		BA-ROAD	$44.63_{(13.74)}$	$53.71_{(14.30)}$	$45.01_{(9.74)}$	57.49
		LOU-ROAD	$45.82_{(12.19)}$	$52.32_{(12.41)}$	$47.40_{(2.84)}$	69.18
		Msplit-HR	$55.96_{(16.92)}$	$44.58_{(17.43)}$	$46.77_{(7.11)}$	3
(25,5)	**(iv)**	US-BCSVM	$48.9_{(15.18)}$	$47.66_{(13.96)}$	$46.17_{(5.45)}$	500
		FAIR	$15.1_{(8.91)}$	$84.08_{(8.87)}$	$33.15_{(10.97)}$	13.38
		$\text{SLDA}_{\text{MCR}_2}$	$41.04_{(15.56)}$	$58.3_{(15.45)}$	$46.52_{(7.29)}$	315.96
		PSIS	$30.26_{(9.52)}$	$65.9_{(13.01)}$	$43.60_{(7.70)}$	1
		ROAD	$12.50_{(8.10)}$	$85.07_{(10.66)}$	$30.67_{(7.51)}$	27.82
		BA-ROAD	$48.07_{(13.10)}$	$48.40_{(14.23)}$	$46.20_{(6.71)}$	60.06
		LOU-ROAD	$48.47_{(13.37)}$	$47.48_{(14.94)}$	$45.97_{(5.10)}$	59.19
		Msplit-HR	$53.06_{(17.66)}$	$43.16_{(17.54)}$	$44.50_{(8.83)}$	2
(50,10)	**(iii)**	US-BCSVM	$48.46_{(12.44)}$	$48.64_{(14.25)}$	$46.94_{(5.82)}$	500
		FAIR	$26.1_{(8.65)}$	$72.12_{(8.35)}$	$42.48_{(6.57)}$	9
		$\text{SLDA}_{\text{MCR}_2}$	$43.76_{(12.28)}$	$55.18_{(11.98)}$	$47.78_{(5.51)}$	493.5
		PSIS	$35.44_{(8.02)}$	$63.02_{(9.33)}$	$46.67_{(5.37)}$	1
		ROAD	$14.10_{(9.43)}$	$83.70_{(11.63)}$	$31.80_{(9.44)}$	59.50
		BA-ROAD	$48.32_{(14.03)}$	$49.66_{(12.68)}$	$47.15_{(7.68)}$	64.50
		LOU-ROAD	$48.18_{(13.51)}$	$48.26_{(12.33)}$	$46.71_{(6.01)}$	68
		Msplit-HR	$50.06_{(15.67)}$	$48.26_{(14.52)}$	$46.91_{(5.78)}$	1
(50,10)	**(iv)**	US-BCSVM	$49.04_{(10.87)}$	$49.28_{(11.96)}$	$48.01_{(5.56)}$	500
		FAIR	$17.38_{(6.91)}$	$76.14_{(10.84)}$	$35.41_{(7.35)}$	13.5
		$\text{SLDA}_{\text{MCR}_2}$	$38.58_{(13.19)}$	$51.14_{(14.79)}$	$42.98_{(9.05)}$	473
		PSIS	$32.02_{(8.58)}$	$54.56_{(17.60)}$	$40.97_{(9.69)}$	1
		ROAD	$15.46_{(9.32)}$	$73.56_{(19.75)}$	$30.97_{(7.64)}$	38.50
		BA-ROAD	$42.20_{(13.08)}$	$44.26_{(15.04)}$	$41.40_{(8.97)}$	38
		LOU-ROAD	$42.60_{(13.97)}$	$43.16_{(14.41)}$	$41.25_{(8.55)}$	47
		Msplit-HR	$46.4_{(14.65)}$	$40.7_{(15.10)}$	$41.55_{(8.81)}$	1
(100,10)	**(iii)**	US-BCSVM	$47.96_{(11.92)}$	$49.98_{(13.98)}$	$47.47_{(5.67)}$	500
		FAIR	$23.91_{(8.31)}$	$74.6_{(7.48)}$	$41.28_{(7.55)}$	8.22
		$\text{SLDA}_{\text{MCR}_2}$	$44.28_{(12.54)}$	$54.34_{(12.45)}$	$47.61_{(5.27)}$	403.99
		PSIS	$33.32_{(7.73)}$	$65.48_{(8.83)}$	$46.24_{(5.87)}$	1.01
		ROAD	$4.60_{(3.47)}$	$94.60_{(4.99)}$	$19.02_{(8.42)}$	96.49
		BA-ROAD	$45.40_{(13.39)}$	$51.16_{(14.24)}$	$45.90_{(8.21)}$	105.05
		LOU-ROAD	$46.24_{(8.52)}$	$43.12_{(10.44)}$	$43.99_{(6.11)}$	103.05
		Msplit-HR	$51.76_{(14.45)}$	$46_{(14.62)}$	$46.61_{(6.55)}$	5.67
(100,10)	**(iv)**	US-BCSVM	$48.72_{(11.81)}$	$48.94_{(12.06)}$	$47.65_{(5.53)}$	500
		FAIR	$15.1_{(7.41)}$	$79.42_{(9.89)}$	$33.09_{(8.62)}$	17.18
		$\text{SLDA}_{\text{MCR}_2}$	$36.98_{(12.42)}$	$53.58_{(14.94)}$	$43.12_{(9.01)}$	226.33
		PSIS	$30.28_{(7.57)}$	$54.96_{(18.29)}$	$39.94_{(9.12)}$	1.01
		ROAD	$4.64_{(3.47)}$	$94.60_{(4.99)}$	$19.02_{(8.42)}$	96.49
		BA-ROAD	$45.40_{(13.39)}$	$51.16_{(14.24)}$	$45.90_{(8.21)}$	105.05
		LOU-ROAD	$45.64_{(13.15)}$	$49.18_{(12.87)}$	$45.79_{(5.82)}$	107.80
		Msplit-HR	$43.52_{(12.21)}$	$44.02_{(14.34)}$	$42.38_{(8.26)}$	5.74

TABLE 6
*Average computational time (in seconds) taken by a method to complete per-sample results: Simulation setting **(iv)**.*

(n_1, n_2, p)	US-BCSVM	FAIR	$\text{SLDA}_{\text{MCR}_2}$	PSIS	ROAD	BA-ROAD	LOU-ROAD	Msplit-HR
(25,5,200)	1.91	1.75	8.75	0.28	50.23	110.34	59.93	11.42
(50,10,200)	1.48	4.73	25.71	4.11	66.00	192.10	189.86	39.31
(100,10,200)	1.86	5.56	66.46	3.06	120.93	468.02	443.23	114.44
(25,5,500)	3.00	29.3	80.05	0.41	204.52	254.64	184.99	16.4
(50,10,500)	3.01	27.09	234.35	3.00	272.15	483.82	477.45	62.23
(100,10,500)	3.58	29.66	451.75	3.61	219.95	974.34	1084.75	176.65

TABLE 7
Classification results for Breast Cancer data set. S denotes the median number of selected features.

$\boldsymbol{\Sigma}$	Methods	$MCR_1\%$	$MCR_2\%$	$GM\%$	S
Diagonal	DROAD	$19.62_{(10.20)}$	$46.31_{(13.38)}$	$28.72_{(7.77)}$	201.50
	HR	$16.82_{(6.70)}$	$46.72_{(10.24)}$	$27.08_{(5.98)}$	26
	US-HR	$20.67_{(7.58)}$	$39.79_{(9.78)}$	$27.93_{(6.10)}$	32
	BLDA	$16.8_{(6.14)}$	$45.59_{(10.86)}$	$26.78_{(5.63)}$	35
	BAI	$22.24_{(7.09)}$	$37_{(10.08)}$	$27.83_{(5.40)}$	99
	LOUI	$22.65_{(7.05)}$	$37.03_{(10.82)}$	$28.06_{(5.20)}$	83.5
	Msplit-HR	$20.96_{(7.00)}$	$39.56_{(10.74)}$	$27.83_{(5.19)}$	6
General	US-BCSVM	$19.78_{(5.74)}$	$34.79_{(10.01)}$	$25.50_{(4.24)}$	1500
	FAIR	$16.24_{(5.54)}$	$45.41_{(9.31)}$	$26.43_{(4.95)}$	22
	$\text{SLDA}_{\text{MCR}_2}$	$22.91_{(12.32)}$	$47.76_{(12.19)}$	$31.58_{(9.17)}$	1500
	PSIS	$27.47_{(14.62)}$	$46.17_{(15.30)}$	$33.77_{(9.85)}$	1
	ROAD	$19.51_{(10.03)}$	$47.41_{(13.81)}$	$28.96_{(7.26)}$	25
	BA-ROAD	$22.16_{(5.95)}$	$38.83_{(9.72)}$	$28.62_{(4.24)}$	51.50
	LOU-ROAD	$22.16_{(5.90)}$	$38.10_{(9.84)}$	$28.37_{(4.32)}$	56.50
	Msplit-HR	$24.11_{(8.85)}$	$40.55_{(10.76)}$	$30.35_{(6.50)}$	5

The first data set, on breast cancer [17], consists of the expression profiles of 2905 genes for 168 patients of whom 111 patients with no event after diagnosis were labelled as "good" and the remaining 57 patients with early metastasis were labelled as "poor". In our analysis, we randomly split the data into training data of sizes 56 and 28 of respectively good cases (the majority Class 1) and poor cases (the minority Class 2). The rest of the data is used for testing. The classification results, under the assumptions of (a) uncorrelated and (b) correlated features, are given in Table 7. Under (a), the results suggest that BAI, LOUI, Msplit-HR, and US-HR have comparable performance, with BAI and LOUI performing slightly better than the other two in terms of the MCR of the minority class (MCR_2). Under (b), BA-ROAD, LOU-ROAD, and Msplit-HR perform similar in terms of the MCRs. US-BCSVM has smaller MCRs compared to the others but by using the set of all features as it is not able to perform any feature selection. Note that in both cases, Msplit-HR selects a much smaller number of features toward the classification task.

The second data set, on multiple-myeloma cancer [39], consists of the expression profiles of $12,2625$ genes for 173 patients with newly diagnosed multiple-myeloma, of whom 137 were with bone lytic lesions and the remaining 36 pa-

TABLE 8
Classification results for Myeloma Cancer data set. S denotes the median number of selected features.

$\boldsymbol{\Sigma}$	Methods	$MCR_1\%$	$MCR_2\%$	$GM\%$	S
Diagonal	DROAD	$26.03_{(11.29)}$	$49.33_{(10.30)}$	$34.93_{(8.58)}$	5
	HR	$25.94_{(11.60)}$	$57.78_{(11.92)}$	$37.43_{(8.47)}$	19
	US-HR	$41.6_{(9.74)}$	$41_{(12.75)}$	$40.17_{(6.66)}$	92.5
	BLDA	$25.58_{(9.10)}$	$53.28_{(11.23)}$	$35.89_{(6.36)}$	11
	BAI	$34.31_{(10.44)}$	$44.17_{(13.20)}$	$37.50_{(5.98)}$	30
	LOUI	$35.14_{(10.54)}$	$44.39_{(11.35)}$	$38.26_{(5.95)}$	27.5
	Msplit-HR	$38.18_{(13.68)}$	$41.94_{(14.37)}$	$37.89_{(7.51)}$	7
General	US-BCSVM	$53.78_{(27.56)}$	$39.44_{(28.32)}$	$46.06_{(18.47)}$	1500
	FAIR	$27.92_{(7.64)}$	$49.56_{(11.16)}$	$36.50_{(6.15)}$	14
	$\text{SLDA}_{\text{MCR}_2}$	$28.83_{(9.79)}$	$47.22_{(10.34)}$	$36.18_{(7.59)}$	13
	PSIS	$31.42_{(15.24)}$	$50.11_{(10.33)}$	$38.43_{(10.72)}$	1
	ROAD	$26.01_{(10.27)}$	$53.22_{(10.63)}$	$36.47_{(8.13)}$	7.50
	BA-ROAD	$34.01_{(13.75)}$	$43.17_{(13.92)}$	$35.52_{(9.38)}$	20
	LOU-ROAD	$33.74_{(9.63)}$	$42.78_{(10.67)}$	$38.05_{(6.43)}$	23.50
	Msplit-HR	$34.74_{(11.84)}$	$42.61_{(11.66)}$	$37.27_{(7.16)}$	6

tients were without bone lytic lesions. We randomly choose a training set containing 18 observations from patients labelled by MRI-no-lytic-lesion (the minority Class 2), and 72 observations from patients labelled by MRI-lytic-lesion (the majority Class 1). The rest of the data were used for testing. Table 8 contains the classification results under the aforementioned assumptions (a) and (b). Under (a), the results show that Msplit-HR and US-HR outperform the other methods in terms of the error rate in the minority class, MCR_2. In addition, Msplit-HR outperforms US-HR in terms of the error rate in the majority class, MCR_1. Under (b), the three methods BA-ROAD, LOU-ROAD, and Msplit-HR perform similar in terms of the MCRs. For this data set, the overall performances of the aforementioned three methods are better than US-BCSVM. Note that in both cases, Msplit-HR selects a smaller number of features toward the classification task.

To reduce the computational cost of each method, and by using a t-statistic, we screened the initial number of features in each of the above data sets by selecting a subset of $p = 1500$ genes.

7. Conclusion

In this paper, we have studied linear discriminant analysis (LDA) in high-dimensional imbalanced binary classification. To the best of our knowledge, this is the first work that rigorously investigates such problems which frequently arise in a wide range of applications.

First, we showed that in the aforementioned settings the standard LDA asymptotically ignores the so-called minority class. Second, using a multiple data splitting technique, we proposed a new method, called Msplit-HR, that obtains desirable large-sample properties. Third, we derived conditions under

which two well-known sparse versions of the LDA in our setting obtain certain desirable large-sample properties. We then examined the finite-sample performance of different methods via simulations and by analyzing two real data sets. In our simulations, the Msplit-HR either outperforms competing methods or has comparable performance in terms of misclassification rate in the minority class, while it has a lower computational cost.

The methodology (Msplit-HR) and theory developed in this paper are based on normal distribution for the feature vector $\mathbf{X}$. The normality is used for bias calculations in Propositions 3.1-3.2, and to establish feature selection consistency in Lemmas 3.1-3.2. On the other hand, [11] showed that feature selection methods based on mean-differences are sensitive to heavy-tailed distributions for $\mathbf{X}$, and they suggested transformation approaches in feature space which are more resistant to extreme observations from heavy-tailed distributions. Properties of such transformations with respect to our theoretical guidelines, and in general, extension of our results to non-normal models require further investigation and is a topic of future research.

If the covariance matrix differs between the two classes, i.e. $\mathbf{X}|(Y = k) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), k = 1, 2$, the optimal (Bayes) rule is the quadratic discriminant analysis (QDA). Our limited numerical experiment shows that the QDA in imbalanced high-dimensional problems behaves similarly to the LDA ignoring the minority class. A potential approach to alleviate the impact of imbalanced class sizes is to reduce the difference between MCRs of an empirical QDA toward that of the optimal rule. However, the main challenge is that none of the aforementioned MCRs have workable closed forms. [22] studied such differences for sparse QDA, and their results might be useful toward imbalanced problems in the context of QDA. This, however, requires a careful investigation and is a subject of future work.

Another possible future research direction is to investigate the possibility of extending the methodology and theory developed in this paper to imbalanced multi-class classification problems.

Appendix A: Technical lemmas

In this Appendix, we first state the technical conditions (C1)-(C3) required in our theoretical developments. Next, we state several lemmas that are used in the proofs of our main results. Lemmas A.1 and A.2 are from [6] and [38]. Lemmas A.3-A.5 are the results from other papers adapted to the imbalanced setting under our consideration. Lemma A.6 states an upper bound for the tail of Student's t-distribution.

Technical Conditions:

(C1) $\log p = o(n_1)$, where n_1 is the majority class size.
(C2) $0 < c_0^{-1} < \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) < c_0 < \infty$, for a constant $c_0 > 0$.
(C3) $0 < c_0^{-1} < \max_{j=1,\ldots,p} \mu_{dj}^2 < c_0 < \infty$, where $\boldsymbol{\mu}_d = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$.

Lemma A.1. *[6, Lemma A.3] Let $\boldsymbol{Z}_i$ be independent and identically random variables from $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\lambda_{\max}(\boldsymbol{\Sigma}) \leq \varepsilon_0^{-1} < \infty$. Then,*

$$P\left(\mid \sum_{i=1}^{n} (Z_{ij}Z_{ik} - \sigma_{jk}) \mid > n\nu \right) \leq C_1 \exp(-C_2 n\nu^2) \quad \text{for all } |\nu| \leq \delta$$

where σ_{jk}'s are entries of $\boldsymbol{\Sigma}$, and C_1, C_2, and δ depend on ϵ_0 only.

Lemma A.2. *[38, Lemma 1] Let ξ_n and ν_n be two sequence of positive numbers such that $\xi_n \to \infty$ and $\nu_n \to 0$ as $n \to \infty$. If $\lim_{n\to\infty} \xi_n \nu_n = \gamma$, where γ may be 0, positive or ∞, then*

$$\lim_{n\to\infty} \frac{\Phi(-\sqrt{\xi_n}(1-\nu_n))}{\Phi(-\sqrt{\xi_n})} = e^{\gamma}.$$

Lemma A.3. *Denote the sets*

$$\begin{aligned} U_\tau(h, c_0(p), M) &= \left\{ \boldsymbol{\Sigma} : \ \sigma_{ii} \leq M, \sum_{j=1}^{p} |\sigma_{ij}|^h \leq c_0(p), \ \forall i, \ 0 \leq h < 1 \right\}, \\ U_\tau(h, c_0(p), M, \epsilon_0) &= \left\{ \boldsymbol{\Sigma} : \ \boldsymbol{\Sigma} \in U_\tau(h, c_0(p), M), \lambda_{\min}(\boldsymbol{\Sigma}) \geq \epsilon_0 > 0 \right\}. \end{aligned}$$

Let $\widetilde{\boldsymbol{\Sigma}}_n$ be a thresholded version of the pooled sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ in (2.5), such that $\tilde{\sigma}_{ij} = (1-2/n)\hat{\sigma}_{ij} \boldsymbol{1}\{(1-2/n)|\hat{\sigma}_{ij}| > t_n\}$, with $t_n = M_1\sqrt{\frac{\log p}{n}}$ and some positive constant M_1. Then uniformly on $U_\tau(h, c_0(p), M)$, and for sufficiently large M_1, under the Condition (C3) and $n_2 = o(n_1)$, as $n_1, n_2 \to \infty$, then

$$\| \widetilde{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma} \| = O_p\left(c_0(p) \left(\log p / n_1 \right)^{\frac{1-h}{2}} \right),$$

and uniformly on $U_\tau(h, c_0(p), M, \epsilon_0)$,

$$\| \widetilde{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}^{-1} \| = O_p\left(c_0(p) \left(\log p / n_1 \right)^{\frac{1-h}{2}} \right).$$

Proof. The proof is a straight forward extension of Theorem 1 of [5] to imbalanced case, and thus omitted here. ■

Lemma A.4. *Let $\boldsymbol{X}_{ik} = (X_{i1k}, ..., X_{ipk})^\top$, for $i = 1, ..., n_k$, and $k = 1, 2$, be random samples from p-variate normal distribution with mean vector $\boldsymbol{0}$ and diagonal covariance matrix $\boldsymbol{D} = diag\{\sigma_1^2, ..., \sigma_p^2\}$. If the Conditions (C1) and (C2) are satisfied and $n_2 = o(n_1)$, then as $n_1, n_2 \to \infty$, we have*

$$\max_{1 \leq j \leq p} \mid \hat{\sigma}_j^2 - \sigma_j^2 \mid = O_p(\sqrt{(\log p)/n_1}),$$

where $\hat{\sigma}_j^2, j = 1, \ldots, p$, are the diagonal elements of the pooled sample variance $\widehat{\boldsymbol{\Sigma}}_n$ in (2.5).

Proof. Let $\overline{X}_{jk} = \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ijk}$, for $k = 1, 2$, $j = 1, ..., p$. We have,

$$\begin{aligned}
&\Pr\left(\max_{1\le j\le p} \mid \hat{\sigma}_j^2 - \sigma_j^2 \mid > \eta\right) \le \sum_{j=1}^{p} \Pr\left(\mid \hat{\sigma}_j^2 - \sigma_j^2 \mid > \eta\right) \\
\le\ & \sum_{k=1}^{2}\sum_{j=1}^{p} \Pr\left(\frac{1}{\sqrt{n_k}} \mid \sum_{i=1}^{n_k}(X_{ijk}^2 - \sigma_j^2) \mid > \frac{1}{\sqrt{n_k}}\frac{\eta}{4}(n_1 + n_2 - 2)\right) \\
+\ & \sum_{k=1}^{2}\sum_{j=1}^{p} \Pr\left(\mid n_k\overline{X}_{jk}^2 - \sigma_j^2 \mid > \frac{\eta}{4}(n_1 + n_2 - 2)\right) \\
\le\ & \sum_{k=1}^{2} pC_1 \exp\left\{ - C_2\frac{\eta^2}{16}\frac{(n-2)^2}{n_k}\right\} + pC_3\exp\left\{ - C_4\frac{\eta^2}{16}(n-2)^2\right\}
\end{aligned}$$

for $|\eta| < \delta$, where C_1, C_2, C_3, C_4, and δ are constants depending only on c_0. The last inequality follows from Lemma A.1. By taking $\eta = M\sqrt{\log p/n_1}$, for sufficiently large $M > 0$, under the imbalanced setting and the Condition (C1), the result holds. ■

Lemma A.5. *Under conditions of Lemma 3.2 and the imbalanced setting* $n_2 = o(n_1)$*, assume that* $m_{\max}\sqrt{\log p/n_1} = o(1)$*. Then for* $\ell = 1, \ldots, \mathcal{L}$*, as long as* $n_1, n_2 \to \infty$*,*

$$\| \widetilde{\mathbf{\Sigma}}_{n,\ell} - \mathbf{\Sigma}_\ell \| = O_p\left(m_{\max}\sqrt{\log p/n_1}\right).$$

where $\widetilde{\mathbf{\Sigma}}_{n,\ell} = [\hat{\sigma}^{(2)}_{jj',\ell} :\ j, j' \in \mathcal{S}^{(1)}_{n,\ell}]$ *and* $\mathbf{\Sigma}_\ell = [\sigma_{jj'} :\ j, j' \in \mathcal{S}^{(1)}_{n,\ell}]$.

Proof. Note that if $\mathbf{A} = [a_{jj'}]$ be a symmetric $p \times p$ matrix then $\| \mathbf{A} \| \le \max_{j'}\sum_{j=1}^{p}|a_{jj'}|$. Thus, the result is implied by

$$\Pr\left(\max_{j\in\mathcal{S}^{(1)}_{n,\ell}}\sum_{j'\in\mathcal{S}^{(1)}_{n,\ell}} |\hat{\sigma}^{(2)}_{jj',\ell} - \sigma_{jj'}| > \eta\right) \le \sum_{j,j'\in\mathcal{S}^{(1)}_{n,\ell}} \Pr\left(|\hat{\sigma}^{(2)}_{jj',\ell} - \sigma_{jj'}| > \frac{\eta}{m_{\max}}\right) \tag{A.1}$$

where $m_{\max} = c_1|\mathcal{S}|(\max_{j\in\mathcal{S}}\beta_j^2)/d_{0,n}^2$. The inequality follows from part (ii) of Lemma 3.2. Let $\boldsymbol{\mu}_{k,\ell} = [\mu_{jk} :\ j \in \mathcal{S}^{(1)}_{n,\ell}]$, $Z_{ijk,\ell} = X_{ijk,\ell} - \mu_{jk,\ell}$, and $\bar{Z}_{jk,\ell} = \sum_{i=1}^{n'_k} X_{ijk,\ell}/n'_k$, where $X_{ijk,\ell} \in \mathcal{D}^{(2)}_{n,\ell}$, for $i = 1, ..., n'_k$, $j = 1, ..., p$, $k = 1, 2$, and $\ell = 1, ..., \mathcal{L}$, where $\mathbf{X}_{ik,\ell} \sim N_p(\boldsymbol{\mu}_{k,\ell}, \mathbf{\Sigma}_\ell)$. For the first probability term in (A.1), we have

$$\begin{aligned}
&\Pr\left(|\hat{\sigma}^{(2)}_{jj',\ell} - \sigma_{jj'}| > \frac{\eta}{m_{\max}}\right) \le \\
&\sum_{k=1}^{2}\Pr\left(\left|\sum_{i=1}^{n'_k} Z_{ijk,\ell}Z_{ij'k,\ell} - n'_k\bar{Z}_{jk,\ell}\bar{Z}_{j'k,\ell} - (n'_k - 1)\sigma_{jj'}\right| > \frac{(n'-2)\eta}{m_{\max}}\right) \\
\le\ & \sum_{k=1}^{2}\left\{\Pr\left(\left|\sum_{i=1}^{n'_k} Z_{ijk,\ell}Z_{ij'k,\ell} - n'_k\sigma_{jj'}\right| > \frac{(n'-2)\eta}{m_{\max}}\right)\right.
\end{aligned}$$

$$+\Pr\left(|n_k'\bar{Z}_{jk,\ell}\bar{Z}_{j'k,\ell}-\sigma_{jj'}|>\frac{(n'-2)\eta}{m_{\max}}\right)\Bigg\}.$$

Finally, using Lemma A.1,

$$\begin{aligned}&\sum_{j,j'\in\mathcal{S}_{n,t}^{(2)}}\Pr\left(|\hat{\sigma}_{jj',\ell}^{(2)}-\sigma_{jj'}|>\frac{\eta}{m_{\max}}\right)\\ \leq\ &\sum_{k=1}^{2}C_1p^2\exp\left\{-C_2\frac{(n-2)^2\eta^2}{m_{\max}^2n_k}\right\}+C_1'p^2\exp\left\{-C_2'\frac{(n-2)^2\eta^2}{m_{\max}^2}\right\},\end{aligned}$$

where C_1, C_1', C_2, C_2' are some positive constants. If $m_{\max}\sqrt{\log p/n_1}=o(1)$ and by taking $\eta = M\times m_{\max}\sqrt{\log p/n_1}$, for sufficiently large $M>0$, the desired result is obtained. ■

Lemma A.6. *Suppose that T has the Student's t-distribution with $n>1$ degrees of freedom. Then, for any large constant $\tau>0$, we have*

$$\Pr(T>\tau)\leq\frac{c_n}{\tau}\frac{n}{n-1}\left(1+\frac{1}{n}\tau^2\right)^{-\frac{n-1}{2}},$$

where $c_n=\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}}$, and $\Gamma(.)$ is the gamma function.

Proof. For any $\tau>0$,

$$\begin{aligned}\Pr(T>\tau) &= \int_\tau^\infty\frac{c_n}{(1+\frac{x^2}{n})^{\frac{n+1}{2}}}dx<\int_\tau^\infty\frac{x}{\tau}\frac{c_n}{(1+\frac{x^2}{n})^{\frac{n+1}{2}}}dx\\ &= \frac{c_n}{\tau}\frac{n}{n-1}\left(1+\frac{1}{n}\tau^2\right)^{-\frac{n-1}{2}}.\end{aligned}$$

The result follows from the facts that $\tau>0$ and $\tau<x<\infty$. ■

Appendix B: Proofs of the main results

In this Appendix, we provide the proofs of Theorems 2.1-4.2.

Proof of Theorem 2.1. Let $\boldsymbol{\epsilon}_{ik}=\mathbf{X}_{ik}-\boldsymbol{\mu}_k$, for $i=1,...,n_k$, and $k=1,2$, where $\mathbf{X}_{ik}=(\mathbf{X}_i|Y_i=k)\sim N_p(\boldsymbol{\mu}_k,\boldsymbol{\Sigma})$, and the vectors $\bar{\boldsymbol{\epsilon}}_k=(\bar{\epsilon}_{1k},\bar{\epsilon}_{2k},...,\bar{\epsilon}_{pk})^\top$ with entires $\bar{\epsilon}_{jk}=\frac{1}{n_k}\sum_{i=1}^{n_k}\epsilon_{ijk}$. Also, recall $\Delta_p^2=\boldsymbol{\mu}_d^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$ and $\boldsymbol{\mu}_d=\boldsymbol{\mu}_2-\boldsymbol{\mu}_1$. The quantities $\Psi_1^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)$, $\Psi_2^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)$, and $\Upsilon^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)$ in (2.6) can be decomposed as

$$\begin{aligned}\Psi_1^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n) &= (\boldsymbol{\mu}_1-\hat{\boldsymbol{\mu}}_a)^\top\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}_2-\hat{\boldsymbol{\mu}}_1)\\ &= \frac{1}{2}(-\bar{\boldsymbol{\epsilon}}_2-\bar{\boldsymbol{\epsilon}}_1-\boldsymbol{\mu}_d)^\top\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\epsilon}}_2-\bar{\boldsymbol{\epsilon}}_1+\boldsymbol{\mu}_d)\\ &= \frac{1}{2}\left\{\bar{\boldsymbol{\epsilon}}_1^\top\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{\epsilon}}_1-\bar{\boldsymbol{\epsilon}}_2^\top\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{\epsilon}}_2-2\bar{\boldsymbol{\epsilon}}_2^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d-\boldsymbol{\mu}_d^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d\right\}\end{aligned}$$

$$= \frac{1}{2}\left\{\mathcal{I}_1 - \mathcal{I}_2 - 2\mathcal{I}_3 - \boldsymbol{\mu}_d^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d\right\},$$

$$\begin{aligned}
\Psi_2^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n) &= -(\boldsymbol{\mu}_2 - \hat{\boldsymbol{\mu}}_a)^\top \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \\
&= -\frac{1}{2}(-\bar{\boldsymbol{\epsilon}}_2 - \bar{\boldsymbol{\epsilon}}_1 + \boldsymbol{\mu}_d)^\top \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\epsilon}}_2 - \bar{\boldsymbol{\epsilon}}_1 + \boldsymbol{\mu}_d) \\
&= -\frac{1}{2}\left\{-\bar{\boldsymbol{\epsilon}}_2^\top \boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{\epsilon}}_2 + \bar{\boldsymbol{\epsilon}}_1^\top \boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{\epsilon}}_1 - 2\bar{\boldsymbol{\epsilon}}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d + \boldsymbol{\mu}_d^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d\right\} \\
&= \frac{1}{2}\left\{\mathcal{I}_2 - \mathcal{I}_1 + 2\mathcal{I}_4 - \boldsymbol{\mu}_d^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d\right\},
\end{aligned}$$

and

$$\begin{aligned}
\Upsilon^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n) &= (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \\
&= (\bar{\boldsymbol{\epsilon}}_2 - \bar{\boldsymbol{\epsilon}}_1 + \boldsymbol{\mu}_d)^\top \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\epsilon}}_2 - \bar{\boldsymbol{\epsilon}}_1 + \boldsymbol{\mu}_d) \\
&= (\bar{\boldsymbol{\epsilon}}_2 - \bar{\boldsymbol{\epsilon}}_1)^\top \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\epsilon}}_2 - \bar{\boldsymbol{\epsilon}}_1) + 2(\bar{\boldsymbol{\epsilon}}_2 - \bar{\boldsymbol{\epsilon}}_1)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d + \boldsymbol{\mu}_d^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d \\
&= \mathcal{I}_5 + 2\mathcal{I}_6 + \boldsymbol{\mu}_d^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d.
\end{aligned}$$

We first show that

$$\mathcal{I}_1 = \bar{\boldsymbol{\epsilon}}_1^\top \boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{\epsilon}}_1 = p/n_1 + o_p(\sqrt{p/n_1}).$$

Note that $\bar{\boldsymbol{\epsilon}}_1 \sim N_p(\mathbf{0}, n_1^{-1}\boldsymbol{\Sigma})$. By Chebyshev's inequality, for any $\tau > 0$,

$$\Pr\left(\sqrt{\frac{n_1}{p}} \mid \mathcal{I}_1 - \frac{p}{n_1} \mid > \tau\right) \leq \frac{1}{\tau^2} Var\{\mathcal{I}_1.\sqrt{n_1/p}\}.$$

This together with the fact that $Var\{\mathcal{I}_1.\sqrt{n_1/p}\} \to 0$, when $n_1, n_2 \to \infty$ such that $n_2 = o(n_1)$, implies that $\mathcal{I}_1 = p/n_1 + o_p(\sqrt{p/n_1})$. Similarly, we have

$$\mathcal{I}_2 = p/n_2 + o_p(\sqrt{p/n_2}) \quad , \quad \mathcal{I}_3 = \bar{\boldsymbol{\epsilon}}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d = O_p\left(\sqrt{\Delta_p^2/n_2}\right),$$

$$\mathcal{I}_4 = \bar{\boldsymbol{\epsilon}}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d = O_p\left(\sqrt{\Delta_p^2/n_1}\right),$$

$$\mathcal{I}_5 = (\bar{\boldsymbol{\epsilon}}_2 - \bar{\boldsymbol{\epsilon}}_1)^\top \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\epsilon}}_2 - \bar{\boldsymbol{\epsilon}}_1) = \sqrt{\frac{np}{n_1 n_2}} o_p(1) + \frac{np}{n_1 n_2},$$

and

$$\mathcal{I}_6 = (\bar{\boldsymbol{\epsilon}}_2 - \bar{\boldsymbol{\epsilon}}_1)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d = O_p\left(\sqrt{\frac{n}{n_1 n_2}\Delta_p^2}\right).$$

By combining the above results, we have

$$\frac{\Psi_1^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)}{\sqrt{\Upsilon^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)}} = \frac{\mathcal{I}_1 - \mathcal{I}_2 - 2\mathcal{I}_3 - \Delta_p^2}{2\left\{\mathcal{I}_5 + 2\mathcal{I}_6 + \Delta_p^2\right\}^{1/2}}$$

$$= \frac{\frac{p}{n_1} + o_p(\sqrt{p/n_1}) - \frac{p}{n_2} + o_p(\sqrt{p/n_2}) + O_p(\sqrt{\Delta_p^2/n_2}) - \Delta_p^2}{2\left\{\sqrt{\frac{np}{n_1 n_2}}o_p(1) + \frac{np}{n_1 n_2} + O_p(\sqrt{n\Delta_p^2/n_1 n_2}) + \Delta_p^2\right\}^{1/2}}$$

$$= \frac{-\sqrt{\frac{p}{n_2}}(1 - \frac{n_2}{n_1}) + o_p(\sqrt{n_2/n_1}) + O_p(\sqrt{\Delta_p^2/p}) - \sqrt{\frac{n_2}{p}}\Delta_p^2}{2\left\{1 + o_p(\sqrt{n_2/p}) + O_p(\sqrt{n_2\Delta_p^2}/p) + n_2\Delta_p^2/p\right\}^{1/2}}$$

and

$$\frac{\Psi_2^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)}{\sqrt{\Upsilon^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)}} = \frac{\mathcal{I}_2 - \mathcal{I}_1 + 2\mathcal{I}_4 - \Delta_p^2}{2\left\{\mathcal{I}_5 + 2\mathcal{I}_6 + \Delta_p^2\right\}^{1/2}}$$

$$= \frac{\frac{p}{n_2} - \frac{p}{n_1} + o_p(\sqrt{p/n_2}) + o_p(\sqrt{p/n_1}) + O_p(\sqrt{\Delta_p^2/n_1}) - \Delta_p^2}{2\left\{\sqrt{\frac{np}{n_1 n_2}}o_p(1) + \frac{np}{n_1 n_2} + O_p(\sqrt{n\Delta_p^2/n_1 n_2}) + \Delta_p^2\right\}^{1/2}}$$

$$= \frac{\sqrt{\frac{p}{n_2}}(1 - \frac{n_2}{n_1}) + o_p(\sqrt{n_2/n_1}) + O_p(\sqrt{n_2\Delta_p^2/pn_1}) - \sqrt{\frac{n_2}{p}}\Delta_p^2}{2\left\{1 + o_p(\sqrt{n_2/p}) + O_p(\sqrt{n_2\Delta_p^2}/p) + n_2\Delta_p^2/p\right\}^{1/2}}.$$

Since $\sqrt{\frac{n_2}{p}}\Delta_p^2 = o(1)$, as long as $n_1, n_2 \to \infty$, thus we obtain

$$\frac{\Psi_1^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)}{\sqrt{\Upsilon^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)}} \xrightarrow{p} -\infty, \qquad \frac{\Psi_2^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)}{\sqrt{\Upsilon^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)}} \xrightarrow{p} +\infty.$$

Hence, $\Pi_1^{\text{LDA}}(\mathcal{D}_n) \xrightarrow{p} 0$ and $\Pi_2^{\text{LDA}}(\mathcal{D}_n) \xrightarrow{p} 1$, which completes the proof. ■

Proof of Lemma 3.1. (a) Note that

$$\Pr\left(\bigcap_{j\notin\mathcal{S}}\{|t_j| \le \tau_n\}\right) = 1 - \Pr\left(\max_{j\notin\mathcal{S}}|t_j| \ > \tau_n\right).$$

By Lemma A.6 of the Appendix A, with $c_n = \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-2}{2})\sqrt{(n-2)\pi}}$, we have

$$\Pr\left(\max_{j\notin\mathcal{S}}|t_j| \ > \tau_n\right) \le \sum_{j\notin\mathcal{S}}\Pr\left(|t_j| > \tau_n\right)$$

$$\le (p-s)\frac{n-2}{n-3}\frac{c_n}{\tau_n}\left(1 + \frac{1}{n-2}\tau_n^2\right)^{-\frac{n-3}{2}}$$

$$:= u(n_1, n_2, p-s, \tau_n),$$

where $n = n_1 + n_2$. The last inequality follows from the upper bound described in Lemma A.6, for the tail of a Student's t-distributed random variable, with $n-2$ degrees of freedom. Since $n_2 = o(n_1)$ as $n_1, n_2 \to \infty$, we then obtain

$$u(n_1, n_2, p-s, \tau_n) \sim \frac{p-s}{\tau_n}\left(1 + \frac{1}{n_1}\tau_n^2\right)^{-n_1}$$

and hence, as $n_1 \to \infty$,

$$u(n_1, n_2, p-s, \tau_n) \sim \frac{p-s}{\tau_n} e^{-\tau_n^2}.$$

Since $\log(p-s) = o(\tau_n^2)$, therefore $\Pr(\max_{j \notin \mathcal{S}} |t_j| > \tau_n) \to 0$, and this completes the proof.

(b) Note that

$$\Pr\left(\bigcap_{j \in \mathcal{S}} \{|t_j| > \tau_n\}\right) = \Pr\left(\min_{j \in \mathcal{S}} |t_j| > \tau_n\right) = 1 - \Pr\left(\min_{j \in \mathcal{S}} |t_j| \le \tau_n\right).$$

Let $\widetilde{t}_j = t_j - \frac{\mu_{dj}}{\hat{\sigma}_j \sqrt{n/n_1 n_2}}$. We have

$$\begin{aligned} \Pr\left(\min_{j \in \mathcal{S}} |t_j| \le \tau_n\right) &= \Pr\left(\max_{j \in \mathcal{S}} |\tilde{t}_j| \ge \min_{j \in \mathcal{S}} \frac{|\mu_{dj}|}{\hat{\sigma}_j \sqrt{n/(n_1 n_2)}} - \tau_n\right) \\ &\le \sum_{j \in \mathcal{S}} \Pr\left(|\tilde{t}_j| \ge \min_{j \in \mathcal{S}} \frac{|\mu_{dj}|}{\hat{\sigma}_j \sqrt{n/(n_1 n_2)}} - \tau_n\right). \end{aligned}$$

Also by Lemma A.4 and under the Condition (C2),

$$\min_{j \in \mathcal{S}} \frac{|\mu_{dj}|}{\hat{\sigma}_j \sqrt{n/(n_1 n_2)}} = d_{0,n}(1 + o_p(1)).$$

Hence,

$$\begin{aligned} &\Pr\left(\min_{j \in \mathcal{S}} |t_j| \le \tau_n\right) \le \sum_{j \in \mathcal{S}} \Pr\left(|\tilde{t}_j| > \frac{d_{0,n}}{\sqrt{n/n_1 n_2}}(1 + o_p(1)) - \tau_n\right) \\ \le\ & s\, \frac{n-2}{n-3}\, \frac{c_n}{\frac{d_{0,n}(1+o_p(1))}{\sqrt{n/n_1 n_2}} - \tau_n} \left(1 + \frac{1}{n-2}\left[\frac{d_{0,n}(1+o_p(1))}{\sqrt{n/n_1 n_2}} - \tau_n\right]^2\right)^{-\frac{n-3}{2}} \\ :=\ & u(n_1, n_2, s, d_{0,n}, \tau_n), \end{aligned}$$

where the last inequality follows from Lemma A.6, when $\tau_n = O(\sqrt{n_2} d_{0,n})$. Since $\sqrt{n_2} d_{0,n} \to \infty$, $\log s = o(n_2 d_{0,n}^2)$, and $n_2 = o(n_1)$, then as $n_1, n_2 \to \infty$, we have $\sqrt{\frac{n}{n_1 n_2}} \sim \frac{1}{\sqrt{n_2}}$, $\frac{n-2}{n-3} \sim 1$ and $c_n = \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-2}{2})\sqrt{(n-2)\pi}} \to \frac{1}{\sqrt{2\pi}}$. Therefore,

$$u(n_1, n_2, s, d_{0,n}, \tau_n) \to 0$$

and it completes the proof. ■

Proof of Theroem 3.1. (a) The class-specific misclassification rates of Msplit-HR in (3.5) are given by

$$\Pi_k^{\text{Msplit-HR}}(\mathcal{D}_n) = \Phi\left(\frac{\Psi_k^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n)}{\sqrt{\Upsilon^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n)}}\right), \quad k = 1, 2,$$

where

$$\Psi_k^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) = \frac{(-1)^{k+1}}{\mathcal{L}} \sum_{\ell=1}^{\mathcal{L}} \sum_{j=1}^{p} \left\{ r_j(\boldsymbol{\mu}_k; \hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}) - \frac{\bar{r}_n}{2} \right\} h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)}),$$

$$\Upsilon^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{\mathcal{L}^2} \sum_{\ell=1}^{\mathcal{L}} \sum_{j=1}^{p} \sigma_j^2 \left(\hat{\mu}_{dj,\ell}^{(2)} / \hat{\sigma}_{j,\ell}^{(2),2} \right)^2 h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)}).$$

By Lemma 3.1, if $\sqrt{n_2} d_{0,n} \to 0$, $\tau_n = O(\sqrt{n_2} d_{0,n})$, $\log(p-s) = o(\tau_n^2)$, and $\log s = o(n_2 d_{0,n})$, as $n_1, n_2 \to \infty$, then

$$\max_{j \in \mathcal{S}} \left| h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)}) - 1 \right| \xrightarrow{p} 0 \quad , \quad \max_{j \notin \mathcal{S}} h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)}) \xrightarrow{p} 0.$$

Using these results, for any $\epsilon > 0$, we have, for $k = 1, 2$,

$$\Pr\left(\left| \sum_{j \notin \mathcal{S}} r_j(\boldsymbol{\mu}_k; \hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}) h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)}) \right| > \epsilon \right) \leq \Pr\left(\max_{j \notin \mathcal{S}} h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)}) > \epsilon \right) \xrightarrow{p} 0,$$

and consequently,

$$\Psi_k^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) = \frac{(-1)^{k+1}}{\mathcal{L}} \sum_{\ell=1}^{\mathcal{L}} \sum_{j \in \mathcal{S}} \left\{ r_j(\boldsymbol{\mu}_k; \hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}) - \frac{\bar{r}_n}{2} \right\} (1 + o_p(1)), \quad k = 1, 2.$$

Similarly, we have

$$\Upsilon^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{\mathcal{L}^2} \sum_{\ell=1}^{\mathcal{L}} \sum_{j \in \mathcal{S}} \sigma_j^2 \left(\hat{\mu}_{dj,\ell}^{(2)} / \hat{\sigma}_{j,\ell}^{(2),2} \right)^2 (1 + o_p(1)).$$

Let $\bar{\epsilon}_{jk,\ell}^{(2)} = \hat{\mu}_{jk,\ell}^{(2)} - \mu_{jk}$, $\mathcal{I}_{k,\ell} = \sum_{j \in \mathcal{S}} (\bar{\epsilon}_{jk,\ell}^{(2)} / \sigma_j)^2$, for $k = 1, 2$, and $\mathcal{I}_{3,\ell} = \sum_{j \in \mathcal{S}} (\bar{\epsilon}_{j2,\ell}^{(2)} \mu_{dj} / \sigma_j^2)$, for each $\ell = 1, .., \mathcal{L}$. By the result of Lemma A.4 in the Appendix A, we have

$$\sum_{j \in \mathcal{S}} r_j(\boldsymbol{\mu}_{j1}, \hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}) h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)}) = \frac{1}{2} \left\{ \mathcal{I}_{1,\ell} - \mathcal{I}_{2,\ell} - 2\mathcal{I}_{3,\ell} - \Delta_p^2 \right\} \left(1 + O_p(\sqrt{\log p / n_1}) \right), \tag{B.1}$$

$$\sum_{j \in \mathcal{S}} r_j(\boldsymbol{\mu}_{j2}, \hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}) h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)}) = \frac{1}{2} \left\{ \mathcal{I}_{1,\ell} - \mathcal{I}_{2,\ell} - 2\mathcal{I}_{4,\ell} + \Delta_p^2 \right\} \left(1 + O_p(\sqrt{\log p / n_1}) \right), \tag{B.2}$$

where $\Delta_p^2 = \sum_{j \in \mathcal{S}} (\mu_{dj}^2 / \sigma_j^2)$. Now, for $\eta > 0$, and $k = 1, 2$

$$\Pr\left(|\mathcal{I}_{k,\ell}| > \eta \right) \leq \frac{s}{n_k \eta}, \tag{B.3}$$

by taking $\eta = M.s/n_k$, for sufficiently large $M > 0$, then $\mathcal{I}_{k,\ell} = O_p(s/n_k)$, for $k = 1, 2..$ By Cauchy-Schwartz inequality, we have $\mathcal{I}_{3,\ell} = O_p(\Delta_p\sqrt{s/n_2})$ and $\mathcal{I}_{4,\ell} = O_p(\Delta_p\sqrt{s/n_1})$. In addition, we have $\sum_{j=1}^{p} \bar{r}_n h_j(\hat{\boldsymbol{\theta}}^{(2)}_{n,\ell}) = o_p(s/n_2)$. By combining these results in (B.1)-(B.2), we arrive at

$$\Psi_k^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) = O_p(s/n_2) + O_p\left(\Delta_p\sqrt{s/n_2}\right) - \frac{1}{2}\Delta_p^2 + O_p\left(\Delta_p^2\sqrt{\log p/n_1}\right), \tag{B.4}$$

for $k = 1, 2$. Let $\mathcal{I}_{5,\ell} = \sum_{j\in\mathcal{S}}(\bar{\epsilon}_{j2,\ell} - \bar{\epsilon}_{j1,\ell})^2/\sigma_j^2$, $\mathcal{I}_{6,\ell} = \sum_{j\in\mathcal{S}}(\bar{\epsilon}_{j2,\ell} - \bar{\epsilon}_{j1,\ell})\mu_{dj}/\sigma_j^2$ for each $\ell = 1, .., \mathcal{L}$. Similar to (B.3), we result $\mathcal{I}_{5,\ell} = O_p(s/n_2)$ and also $\mathcal{I}_{6,\ell} = O_p(\Delta_p\sqrt{s/n_2})$. Therefore

$$\begin{aligned}\Upsilon^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) &= \frac{1}{\mathcal{L}^2}\sum_{\ell=1}^{\mathcal{L}}\left\{\mathcal{I}_{5,\ell} + 2\mathcal{I}_{6,\ell} + \Delta_p^2\right\} \\ &= O_p(s/n_2) + O_p\left(\Delta_p\sqrt{s/n_2}\right) + \Delta_p^2 + O_p\left(\Delta_p^2\sqrt{\log p/n_1}\right). \end{aligned} \tag{B.5}$$

By combining (B.4) and (B.5), we have, for $k = 1, 2$,

$$\begin{aligned}&\Pi_k^{\text{Msplit-HR}}(\mathcal{D}_n) \\ &= \Phi\left(\frac{O_p(s/n_2) + O_p(\Delta_p\sqrt{s/n_2}) - \frac{1}{2}\Delta_p^2 + O_p\left(\Delta_p^2\sqrt{\frac{\log p}{n_1}}\right)}{\left\{O_p(s/n_2) + O_p(\Delta_p\sqrt{s/n_2}) + \Delta_p^2 + O_p\left(\Delta_p^2\sqrt{\frac{\log p}{n_1}}\right)\right\}^{1/2}}\right) \\ &= \Phi\left(-\frac{1}{2}\Delta_p\{1 + O_p(\kappa_n)\}\right),\end{aligned}$$

where $\kappa_n = \max\{\Delta_p^{-1}\sqrt{s/n_2},\ \sqrt{\log p/n_1}\}$.

(b) When $\Delta_p \to \infty$, by Lemma A.2, if $\Delta_p^2\kappa_n = o(1)$, then Msplit-HR is asymptotically-strong optimal and the result follows. The condition $\Delta_p^2\kappa_n = o(1)$ is equivalent to $\Delta_p^2\sqrt{\log p/n_1} = o(1)$, and $\Delta_p^2 s = o(n_2)$. ■

Proof of Lemma 3.2. We follow a similar line of proof as in [29, Theorem 1], to show the results of both parts (a) and (b), under the imbalanced setting.

(a) It is enough to show that for any $\ell = 1, ..., \mathcal{L}$, as $n_1, n_2 \to \infty$,

$$\Pr\left(\mathcal{S} \not\subset \mathcal{S}^{(1)}_{n,\ell}\right) \to 0.$$

Suppose that there exist an index j in $\mathcal{S}$ for which $j \notin \mathcal{S}^{(1)}_{n,\ell}$. Thus, $|\mu_{dj}| \geq d_{0,n}$ and $|\hat{\mu}^{(1)}_{dj,\ell}| < \tau_n$, where $d_{0,n} = \min_{j\in\mathcal{S}}|\mu_{dj}|$. It results in $|\hat{\mu}^{(1)}_{dj,\ell} - \mu_{dj}| > d_{0,n} - \tau_n$. By conditions $\tau_n \asymp d_{0,n}$ and $\lambda_{\max}(\boldsymbol{\Sigma}) < c_0$, and for some constants $C_1, C_2 > 0$, we have

$$\Pr\left(\mathcal{S} \not\subset \mathcal{S}^{(1)}_{n,\ell}\right) \leq \sum_{j=1}^{p}\Pr\left(|\hat{\mu}^{(1)}_{dj,\ell} - \mu_{dj}| > d_{0,n} - \tau_n\right)$$

$$\leq \quad C_1\left(\frac{p}{d_{0,n}-\tau_n}\right)\sqrt{\frac{n_1'+n_2'}{n_1'n_2'}}\ \exp\left\{-C_2\frac{n_1'n_2'(d_{0,n}-\tau_n)^2}{n_1'+n_2'}\right\}.$$

The last term tends to zero, since $\log p = o(n_2 d_{0,n}^2)$ and $n_2 = o(n_1)$, and thus the result follows.

(b) By condition $\lambda_{\max}(\boldsymbol{\Sigma}) < c_0$, we have

$$\boldsymbol{\mu}_d^\top \boldsymbol{\mu}_d = \boldsymbol{\mu}_d^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d \leq \lambda_{\max}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)\boldsymbol{\beta}^\top\boldsymbol{\beta} \leq c_0 \times |\mathcal{S}| \times \max_{j\in\mathcal{S}}\beta_j^2. \quad \text{(B.6)}$$

Let $\mathcal{S}^* = \{j : |\mu_{dj}| > d_{0,n}/r\}$, for some constant $r > 1$. Thus, $\boldsymbol{\mu}_d^\top\boldsymbol{\mu}_d \geq |\mathcal{S}^*|d_{0,n}^2/r^2$. This together with (B.6), result in $|\mathcal{S}^*| \leq C_3|\mathcal{S}|\max_{j\in\mathcal{S}}\beta_j^2/d_{0,n}^2 \doteq m_{\max}$, for constant $C_3 > 0$. The result in part (b) follows by proving that, $|\mathcal{S}_{n,\ell}^{(1)}| < |\mathcal{S}^*|$, with probability tending to one, for any $\ell = 1, ..., \mathcal{L}$. If there exists an index j in $\mathcal{S}_{n,\ell}^{(1)}$ for which $j \notin \mathcal{S}^*$, thus $|\hat{\mu}_{dj,\ell}^{(1)}| > \tau_n$ and $|\mu_{dj}| < d_{0,n}/r$ and consequently, $|\hat{\mu}_{dj,\ell}^{(1)} - \mu_{dj}| > \tau_n - d_{0,n}/r$. Therefore, by condition $\tau_n \asymp d_{0,n}$ and for constants $C_4, C_5 > 0$

$$\begin{aligned}\Pr\left(|\mathcal{S}_{n,\ell}^{(1)}| \geq |\mathcal{S}^*|\right) \quad &\leq \quad \Pr\left(\mathcal{S}_{n,\ell}^{(1)} \not\subset \mathcal{S}^*\right) \leq \sum_{j=1}^{p}\Pr\left(|\hat{\mu}_{dj,\ell}^{(1)} - \mu_{dj}| > \tau_n - d_{0,n}/r\right)\\ &\leq \quad C_4\frac{p}{\tau_n - d_{0,n}}\sqrt{\frac{n_1'+n_2'}{n_1'n_2'}}\exp\left\{-C_5\frac{(\tau_n-d_{0,n})^2 n_1'n_2'}{n_1'+n_2'}\right\}.\end{aligned}$$

The last term tends to zero, as $\log p = o(n_2 d_{0,n}^2)$ and $\sqrt{n_2}d_{0,n} \to \infty$. ■

Proof of Theroem 3.2. **(a)** The misclassification rates of Msplit-HR in (3.8), are given as

$$\Pi_k^{\text{Msplit-HR}}(\mathcal{D}_n) = \Phi\left(\frac{\Psi_k^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n)}{\sqrt{\Upsilon^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n)}}\right), \quad k = 1, 2$$

where

$$\begin{aligned}\Psi_k^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) \quad &= \quad \frac{(-1)^k}{\mathcal{L}}\sum_{\ell=1}^{\mathcal{L}}\{\tilde{\boldsymbol{\mu}}_{d,\ell}^\top\tilde{\boldsymbol{\Sigma}}_{n,\ell}^{-1}(\tilde{\boldsymbol{\mu}}_{a,\ell} - \boldsymbol{\mu}_{k,\ell}) - \frac{\bar{r}_{n,\ell}}{2}\},\\ \bar{r}_{n,\ell} \quad &= \quad \left(\frac{1}{n_1'} - \frac{1}{n_2'}\right)\frac{n'-2}{n'-3-|\mathcal{S}_{n,\ell}^{(1)}|} \times |\mathcal{S}_{n,\ell}^{(1)}|,\\ \Upsilon^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) \quad &= \quad \frac{1}{\mathcal{L}^2}\sum_{\ell=1}^{\mathcal{L}}\{\tilde{\boldsymbol{\mu}}_{d,\ell}^\top\tilde{\boldsymbol{\Sigma}}_{n,\ell}^{-1}\boldsymbol{\Sigma}_\ell\tilde{\boldsymbol{\Sigma}}_{n,\ell}^{-1}\tilde{\boldsymbol{\mu}}_{d,\ell}\}.\end{aligned}$$

By Lemma A.5, we obtain

$$\tilde{\boldsymbol{\mu}}_{d,\ell}^\top\tilde{\boldsymbol{\Sigma}}_{n,\ell}^{-1}\boldsymbol{\Sigma}_\ell\tilde{\boldsymbol{\Sigma}}_{n,\ell}^{-1}\tilde{\boldsymbol{\mu}}_{d,\ell} = \tilde{\boldsymbol{\mu}}_{d,\ell}^\top\boldsymbol{\Sigma}_\ell^{-1}\tilde{\boldsymbol{\mu}}_{d,\ell}\left(1 + O_p(m_{\max}\sqrt{\log p/n_1})\right). \quad \text{(B.7)}$$

We consider the following decomposition

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_{d,\ell}^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}\tilde{\boldsymbol{\mu}}_{d,\ell} &= (\tilde{\boldsymbol{\mu}}_{d,\ell}-\boldsymbol{\mu}_{d,\ell})^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}(\tilde{\boldsymbol{\mu}}_{d,\ell}-\boldsymbol{\mu}_{d,\ell})+2(\tilde{\boldsymbol{\mu}}_{d,\ell}-\boldsymbol{\mu}_{d,\ell})^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}\boldsymbol{\mu}_{d,\ell}\\ &+ \boldsymbol{\mu}_{d,\ell}^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}\boldsymbol{\mu}_{d,\ell}=\mathcal{A}_1+2\mathcal{A}_2+\mathcal{A}_3\end{aligned}$$

Now by Lemma 3.2 and Markov's inequality, also using the Condition (C2) in the Appendix A, we have for a constant $C_1>0$,

$$\Pr\left((\tilde{\boldsymbol{\mu}}_{d,\ell}-\boldsymbol{\mu}_{d,\ell})^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}(\tilde{\boldsymbol{\mu}}_{d,\ell}-\boldsymbol{\mu}_{d,\ell})>\eta\right)\leq\frac{C_1}{\eta}\frac{n}{n_1n_2}m_{\max}.$$

If $\eta=M\frac{m_{\max}}{n_2}$, then for large $M>0$, $\mathcal{A}_1=O_p(m_{\max}/n_2)$. By Cauchy-Schwartz inequality, $\mathcal{A}_2^2\leq(\tilde{\boldsymbol{\mu}}_{d,\ell}-\boldsymbol{\mu}_{d,\ell})^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}(\tilde{\boldsymbol{\mu}}_{d,\ell}-\boldsymbol{\mu}_{d,\ell})\mathcal{A}_3$. Hence $\mathcal{A}_2=O_p(\sqrt{m_{\max}/n_2})\mathcal{A}_3^{1/2}$. Therefore, by combining these results we have

$$\tilde{\boldsymbol{\mu}}_{d,\ell}^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}\tilde{\boldsymbol{\mu}}_{d,\ell}=O_p(m_{\max}/n_2)+O_p(\sqrt{m_{\max}/n_2})\sqrt{\mathcal{A}_3}+\mathcal{A}_3 \tag{B.8}$$

Now for $\Psi_1^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n)$, we have

$$\tilde{\boldsymbol{\mu}}_{d,\ell}^{\top}\tilde{\boldsymbol{\Sigma}}_{n,\ell}^{-1}(\boldsymbol{\mu}_{1,\ell}-\tilde{\boldsymbol{\mu}}_{a,\ell})=\tilde{\boldsymbol{\mu}}_{d,\ell}^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}(\boldsymbol{\mu}_{1,\ell}-\tilde{\boldsymbol{\mu}}_{a,\ell})\left(1+O_p(m_{\max}\sqrt{\log p/n_1})\right) \tag{B.9}$$

We decompose it as

$$\begin{aligned}2\tilde{\boldsymbol{\mu}}_{d,\ell}^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}(\boldsymbol{\mu}_{1,\ell}-\tilde{\boldsymbol{\mu}}_{a,\ell}) &= (\tilde{\boldsymbol{\mu}}_{1,\ell}-\boldsymbol{\mu}_{1,\ell})^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}(\tilde{\boldsymbol{\mu}}_{1,\ell}-\boldsymbol{\mu}_{1,\ell})\\ &- (\tilde{\boldsymbol{\mu}}_{2,\ell}-\boldsymbol{\mu}_{2,\ell})^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}(\tilde{\boldsymbol{\mu}}_{2,\ell}-\boldsymbol{\mu}_{2,\ell})\\ &= 2(\tilde{\boldsymbol{\mu}}_{2,\ell}-\boldsymbol{\mu}_{2,\ell})^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}\boldsymbol{\mu}_{d,\ell}-\boldsymbol{\mu}_{d,\ell}^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}\boldsymbol{\mu}_{d,\ell}\\ &= \mathcal{B}_1-\mathcal{B}_2-2\mathcal{B}_3-\mathcal{A}_3\end{aligned}$$

Similar to the proof of $\mathcal{A}_1$, we have $\mathcal{B}_1=O_p(m_{\max}/n_1)$, and $\mathcal{B}_2=O_p(m_{\max}/n_2)$. Also similar to $\mathcal{A}_2$, we have $\mathcal{B}_3=O_p(\sqrt{m_{\max}/n_2})\sqrt{\mathcal{A}_3}$. Hence,

$$\tilde{\boldsymbol{\mu}}_{d,\ell}^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}(\boldsymbol{\mu}_{1,\ell}-\tilde{\boldsymbol{\mu}}_{a,\ell})=O_p(\frac{m_{\max}}{n_1})+O_p(\frac{m_{\max}}{n_2})+O_p(\sqrt{m_{\max}/n_2})\mathcal{A}_3^{1/2}-\frac{1}{2}\mathcal{A}_3 \tag{B.10}$$

We recall that $\Delta_p^2=\boldsymbol{\mu}_d^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d=\boldsymbol{\beta}^{\top}\boldsymbol{\mu}_d$ and $\mathcal{S}_{n,\ell}^{(1)}=\{j:|\hat{\boldsymbol{\mu}}_{dj,\ell}^{(1)}|>\tau_n\}$. For each $\ell=1,..,\mathcal{L}$, and any $\eta>0$

$$\begin{aligned}&\Pr\left(|\boldsymbol{\mu}_{d,\ell}^{\top}\boldsymbol{\Sigma}_{\ell}^{-1}\boldsymbol{\mu}_{d,\ell}-\Delta_p^2|>\eta\right)=\Pr\left(|\sum_{j\in\mathcal{S}_{n,\ell}^{(1)}}\beta_j\mu_{dj}-\sum_{j'\in\mathcal{S}}\beta_{j'}\mu_{dj'}|>\eta\right)\\ =\;&\Pr\left(\Big|\sum_{j\in\mathcal{S}_{n,\ell}^{(1)},j\notin\mathcal{S}}\beta_j\mu_{dj}+\sum_{j\in\mathcal{S}_{n,\ell}^{(1)},j\in\mathcal{S}}\beta_j\mu_{dj}-\sum_{j'\in\mathcal{S},j'\in\mathcal{S}_{n,\ell}^{(1)}}\beta_{j'}\mu_{dj'}-\sum_{j'\in\mathcal{S},j'\notin\mathcal{S}_{n,\ell}^{(1)}}\beta_{j'}\mu_{dj'}\Big|>\eta\right)\\ =\;&\Pr\left(|\sum_{j\in\mathcal{S},j\notin\mathcal{S}_{n,\ell}^{(1)}}\beta_j\mu_{dj}|>\eta\right)\leq\sum_{j=1}^{p}\Pr\left(j\in\mathcal{S}\text{ and }j\notin\mathcal{S}_{n,\ell}^{(1)}\right)\end{aligned}$$

By part (i) of Lemma 3.2, the last term tends to zero, as $n_1, n_2 \to \infty$. Therefore, $\mathcal{A}_3 = \Delta_p^2 + o_p(1)$. By combining this result together with (B.7)-(B.10), also with $\bar{r}_{n,t} = O_p(m_{\max}/n_2)$, we result

$$\begin{aligned}\Pi_1^{\text{Msplit-HR}}(\mathcal{D}_n) &= \Phi\left(\frac{-\Delta_p}{2}\left\{1 + O_p\left(\Delta_p^{-1}\sqrt{\frac{m_{\max}}{n_2}}\right) + O_p\left(m_{\max}\sqrt{\frac{\log p}{n_1}}\right)\right\}\right)\\ &= \Phi\left(\frac{-\Delta_p}{2}(1 + O_p(\kappa_n'))\right),\end{aligned}$$

We can show the same result for $\Pi_2^{\text{Msplit-HR}}(\mathcal{D}_n)$.

(b) When $\Delta_p \to \infty$, the result follows from Lemma A.2 by condition $\Delta_p^2\kappa_n' = o(1)$. ■

Proof of Lemma 4.1. (a) Recall the sequence $a_n = M_2(\log p/n)^\alpha$, with $0 < \alpha < 1/2$ and $M_2 > 0$. Let c_1, c_2 be some positive constants. Inspired by the proof of Lemma 2 of [38], we have

$$\begin{aligned}&\Pr\left(\bigcap_{\{j:|\mu_{dj}|>ra_n\}}\{|\hat{\mu}_{dj}| > a_n\}\right) \geq 1 - \sum_{j=1}^{p}\Pr\left(|\hat{\mu}_{dj} - \mu_{dj}| > a_n(r-1)\right)\\ &\geq 1 - 2\sum_{j=1}^{p}\Phi\left(\frac{-a_n(r-1)}{\sigma_j\sqrt{n/n_1n_2}}\right)\\ &\geq 1 - pc_1\exp\left\{-\left(\frac{\log p}{n}\right)^{2\alpha}.\frac{n_1n_2}{n}c_2\right\}. \qquad \text{(B.11)}\end{aligned}$$

Since $(\log p/n_2)(n_1/\log p)^{2\alpha} = o(1)$ and $n_2 = o(n_1)$, as $n_1, n_2 \to \infty$, (B.11) tends to 1, and the result of part (a) holds.

(b) Similar to part (a), for some positive constants c_1, c_2, we have

$$\Pr\left(\bigcap_{\{j:|\mu_{dj}|\leq a_n/r\}}\{|\hat{\mu}_{dj}| \leq a_n\}\right) \geq 1 - pc_1\exp\left\{-\left(\frac{\log p}{n}\right)^{2\alpha}\frac{n_1n_2}{n}c_2\right\}.$$

This together with $(\log p/n_2)(n_1/\log p)^{2\alpha} = o(1)$ and $n_2 = o(n_1)$, prove that the right hand side of the above inequality tends to 1, as $n_1, n_2 \to \infty$.

(c) The result follows from parts (a) and (b). ■

Proof of Theorem 4.1. (a) The misclassification rates of SLDA in Class $k = 1, 2$, are given as

$$\Pi_k^{\text{SLDA}}(\mathcal{D}_n) = \Phi\left(\frac{(-1)^k\tilde{\boldsymbol{\mu}}_d^\top\widetilde{\boldsymbol{\Sigma}}_n^{-1}(\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) - \tilde{\boldsymbol{\mu}}_d^\top\widetilde{\boldsymbol{\Sigma}}_n^{-1}\hat{\boldsymbol{\mu}}_d/2}{\sqrt{\tilde{\boldsymbol{\mu}}_d^\top\widetilde{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{\Sigma}\widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d}}\right).$$

Recall $d_{n_1} = C_{h,p}(n_1^{-1}\log p)^{(1-h)/2}$, where $C_{h,p} = \max_{1\leq i\leq p}\sum_{j=1}^{p}|\sigma_{ij}|^h$ for some $0 \leq h < 1$. It follows from Lemma A.3 in the Appendix A that

$$\tilde{\boldsymbol{\mu}}_d^\top\widetilde{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{\Sigma}\widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d = \tilde{\boldsymbol{\mu}}_d^\top\widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d\{1 + O_p(d_{n_1})\}.$$

Let $\Delta_p^2 = \boldsymbol{\mu}_d^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$, $\mathcal{J}_1 = (\tilde{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)^\top \boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)$ and $\mathcal{J}_2 = 2\boldsymbol{\mu}_d^\top \boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)$. Now,

$$\tilde{\boldsymbol{\mu}}_d^\top \boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\mu}}_d = \mathcal{J}_1 + \mathcal{J}_2 + \Delta_p^2$$

Following by the proof of Theorem 1 of [38], we have

$$\mathcal{J}_1 \leq c_0 \left\{ \parallel \tilde{\boldsymbol{\mu}}_{d1} - \boldsymbol{\mu}_{d1} \parallel^2 + \parallel \boldsymbol{\mu}_{d0} \parallel^2 \right\},$$

where $\tilde{\boldsymbol{\mu}}_d^\top = (\tilde{\boldsymbol{\mu}}_{d1}^\top, \mathbf{0}^\top)$, $\boldsymbol{\mu}_d^\top = (\boldsymbol{\mu}_{d1}^\top, \boldsymbol{\mu}_{d0}^\top)$, and $\tilde{\boldsymbol{\mu}}_{d1}$ and $\boldsymbol{\mu}_{d1}$ are two vectors of dimension $\hat{q}$, whose elements correspond to those features x_js for which $|\hat{\mu}_{dj}| > a_n$. By condition (4.1), we have $\parallel \tilde{\boldsymbol{\mu}}_{d1} - \boldsymbol{\mu}_{d1} \parallel^2 = O_p(q_n/n_2)$, $\parallel \boldsymbol{\mu}_{d0} \parallel^2 = O_p\left(D_{g,p}.a_n^{2(1-g)}\right)$, and $\mathcal{J}_1 = O_p(k_{n_2})$, where $k_{n_2} = \max\{\frac{q_n}{n_2},\ D_{g,p}a_n^{2(1-g)}\}$. Consequently, by condition (4.1),

$$\mathcal{J}_2 = (\tilde{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d \leq \Delta_p \sqrt{\parallel \tilde{\boldsymbol{\mu}}_{d1} - \boldsymbol{\mu}_{d1} \parallel^2 + \parallel \boldsymbol{\mu}_{d0} \parallel^2} = \Delta_p\ O_p(\sqrt{k_{n_2}}).$$

Therefore in the denominator of $\Pi_k^{\text{SLDA}}(\mathcal{D}_n)$, we have

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{\Sigma}\widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d &= \left\{ O_p(k_{n_2}) + \Delta_p O_p(\sqrt{k_{n_2}}) + \Delta_p^2 \right\} O_p(d_{n_1}) \\ &= \left\{ O_p\left(\sqrt{k_{n_2}/\Delta_p^2}\right) + 1 \right\} \Delta_p^2\ O_p(d_{n_1}). \end{aligned} \tag{B.12}$$

Now, the numerator of $\Pi_k^{\text{SLDA}}(\mathcal{D}_n)$ can be decomposed as

$$\begin{aligned} & (-1)^k \tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}(\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) - \frac{1}{2}\hat{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d \\ = \ & (-1)^k \tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}(\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) - \frac{1}{2}(\hat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d - \frac{1}{2}(\boldsymbol{\mu}_d - \tilde{\boldsymbol{\mu}}_d)^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d \\ - \ & \frac{1}{2}\tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d \\ = \ & \mathcal{J}_3 + \mathcal{J}_4 + \mathcal{J}_5 - \frac{1}{2}\tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d \\ = \ & \sqrt{\tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d} \left\{ O_p\left(\sqrt{q_n/n_k}\right) + O_p\left(\sqrt{q_n C_{h,p}/n_k}\right) \right\} \sqrt{1 + O_p(d_{n_1})} \\ + \ & \sqrt{\tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d} O_p(\sqrt{q_n/n_2}) \sqrt{1 + O_p(d_{n_1})} \\ + \ & \left\{ O_p(k_{n_2}) + \Delta_p O_p(\sqrt{k_{n_2}}) \right\} O_p(d_{n_1}) + \tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d. \end{aligned} \tag{B.13}$$

Again, by condition (4.1) we have

$$\begin{aligned} \mathcal{J}_3 &= \tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}(\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) \\ &= \sqrt{\tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d} \left\{ O_p\left(\sqrt{q_n/n_k}\right) + O_p\left(\sqrt{q_n C_{h,p}/n_k}\right) \right\} \sqrt{1 + O_p(d_{n_1})}, \end{aligned}$$

and

$$\mathcal{J}_4 = (\hat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d = \sqrt{\tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1}\tilde{\boldsymbol{\mu}}_d} O_p(\sqrt{q_n/n_2}).$$

Also, similar to the expression of $\mathcal{J}_1$, we have

$$\mathcal{J}_5 = (\boldsymbol{\mu}_d - \tilde{\boldsymbol{\mu}}_d)^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1} \tilde{\boldsymbol{\mu}}_d = \sqrt{\tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1} \tilde{\boldsymbol{\mu}}_d} O_p(\sqrt{k_{n_2}}) \sqrt{1 + O_p(d_{n_1})}.$$

finally, by combining (B.12) and (B.13) we arrive at

$$\begin{aligned} \Pi_k^{\text{SLDA}}(\mathcal{D}_n) &= \Phi\left(\frac{(-1)^k \tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1} (\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)}{\sqrt{\tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1} \boldsymbol{\Sigma} \widetilde{\boldsymbol{\Sigma}}_n^{-1} \tilde{\boldsymbol{\mu}}_d}} - \frac{1}{2} \frac{\hat{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1} \tilde{\boldsymbol{\mu}}_d}{\sqrt{\tilde{\boldsymbol{\mu}}_d^\top \widetilde{\boldsymbol{\Sigma}}_n^{-1} \boldsymbol{\Sigma} \widetilde{\boldsymbol{\Sigma}}_n^{-1} \tilde{\boldsymbol{\mu}}_d}} \right) \\ &= \Phi\left(-\frac{1}{2} \Delta_p \left\{ O_p\left(\Delta_p^{-1} \sqrt{q_n C_{h,p} / n_k} \right) + O_p\left(\sqrt{k_{n_2} / \Delta_p^2} \right) + 1 + O_p(d_{n_1}) \right\} \right) \\ &= \Phi\left(-\frac{1}{2} \Delta_p \left\{ 1 + O_p(b_{n_k}) \right\} \right) \quad , \quad k = 1, 2 \end{aligned}$$

as claimed, where

$$b_{n_k} = \max\left\{ d_{n_1}, \ \frac{\sqrt{k_{n_2}}}{\Delta_p}, \ \frac{1}{\Delta_p} \sqrt{\frac{q_n}{n_k} C_{h,p}} \right\}.$$

(b)-i. If Δ_p is bounded, then $\Delta_p^2 b_{n_k} \to 0$ is equivalent to $b_{n_2} \to 0$, which imply $\Pi_k^{\text{SLDA}}(\mathcal{D}_n) / \Pi^{\text{opt}} \xrightarrow{p} 1$, for $k = 1, 2$.
(b)-ii. If $\Delta_p \to \infty$, by Lemma A.2 in the Appendix A, when $\Delta_p^2 b_{n_2} \to 0$, and consequently $\Delta_p^2 b_{n_1} \to 0$, we have $\Pi_k^{\text{SLDA}}(\mathcal{D}_n) / \Pi^{\text{opt}} \xrightarrow{p} 1$, for $k = 1, 2$. ■

Proof of Theorem 4.2. The class-specific MCRs of the ROAD in (4.2) are given by

$$\Pi_k^{\text{ROAD}}(\mathcal{D}_n; c) = \Phi\left(\frac{(-1)^k \hat{\mathbf{w}}_c^\top (\hat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_k)}{(\hat{\mathbf{w}}_c^\top \boldsymbol{\Sigma} \hat{\mathbf{w}}_c)^{1/2}} \right), \ k = 1, 2.$$

The oracle versions of the MCRs, evaluated at the true parameter values of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_k$, are given by

$$\Pi_k^{\text{orc}}(c) = \Phi\left(\frac{-\mathbf{w}_c^\top \boldsymbol{\mu}_d}{2(\mathbf{w}_c^\top \boldsymbol{\Sigma} \mathbf{w}_c)^{1/2}} \right), \ k = 1, 2.$$

By the tail probability inequality

$$1 - \Phi(\tau) \leq \frac{1}{\tau \sqrt{2\pi}} \exp\{-\tau^2/2\}, \quad \tau > 0,$$

we have that, for $\eta_1 > 0$,

$$\Pr\left(\| \hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k \|_\infty > \eta_1 \right) \leq \sum_{j=1}^{p} \Pr\left(|\hat{\mu}_{jk} - \mu_{jk}| > \eta_1 \right) \leq C_1 p \exp\{-C_2 n_k \eta_1^2\}.$$

Thus, by choosing $\eta_1 = M_1 a_{n_k}$, for some $M_1 > 0$, we arrive at $\| \hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k \|_\infty = O_p(\sqrt{\log p/n_k})$. Also, by Lemma A.1 in the Appendix A, for $\eta_2 > 0$,

$$\begin{aligned}
& \Pr\left(\max_{j,l} |\hat{\sigma}_{j,l} - \sigma_{j,l}| > \eta_2 \right) \leq \\
\leq & \sum_{j,l} \sum_{k=1}^{2} \Pr\left(|\sum_{i=1}^{n_k} (X_{ijk} X_{ilk} - \sigma_{jl})| > (n-2)\eta_2/4 \right) \\
+ & \sum_{j,l} \sum_{k=1}^{2} \Pr\left(|n_k \hat{\mu}_{jk} \hat{\mu}_{lk} - \sigma_{jl})| > (n-2)\eta_2/4 \right) \\
\leq & p^2 C_1 \exp\{-C_2 (n-2)^2 \eta_2^2 / n_k\} + p^2 C_3 \exp\{-C_4 (n-2)^2 \eta_2^2\}.
\end{aligned}$$

Thus, by choosing $\eta_2 = M_2 \sqrt{\log p/n_1}$, for some $M_2 > 0$, we arrive at $\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_\infty = O_p(\sqrt{\log p/n_1})$. Using the Lipschitz property of the cumulative distribution function of standard normal, $\Phi(.)$, we have

$$\begin{aligned}
& \left| \Pi_2^{\text{ROAD}}(\mathcal{D}_n; c) - \Pi_2^{\text{orc}}(c) \right| \leq \left| \frac{-\hat{\mathbf{w}}_c^\top (\boldsymbol{\mu}_2 - \hat{\boldsymbol{\mu}}_a)}{(\hat{\mathbf{w}}_c^\top \boldsymbol{\Sigma} \hat{\mathbf{w}}_c)^{1/2}} - \frac{-\mathbf{w}_c^\top \boldsymbol{\mu}_d}{2(\mathbf{w}_c^\top \boldsymbol{\Sigma} \mathbf{w}_c)^{1/2}} \right| \\
= & \left| \frac{-\hat{\mathbf{w}}_c^\top (\boldsymbol{\mu}_2 - \hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_a)}{(\hat{\mathbf{w}}_c^\top \boldsymbol{\Sigma} \hat{\mathbf{w}}_c)^{1/2}} - \frac{-\mathbf{w}_c^\top \boldsymbol{\mu}_d}{2(\mathbf{w}_c^\top \boldsymbol{\Sigma} \mathbf{w}_c)^{1/2}} \right| \\
\leq & \left| \frac{\hat{\mathbf{w}}_c^\top (\boldsymbol{\mu}_2 - \hat{\boldsymbol{\mu}}_2)}{(\hat{\mathbf{w}}_c^\top \boldsymbol{\Sigma} \hat{\mathbf{w}}_c)^{1/2}} \right| + \left| \frac{\hat{\mathbf{w}}_c^\top \hat{\boldsymbol{\mu}}_d}{2(\hat{\mathbf{w}}_c^\top \boldsymbol{\Sigma} \hat{\mathbf{w}}_c)^{1/2}} - \frac{\mathbf{w}_c^\top \boldsymbol{\mu}_d}{2(\mathbf{w}_c^\top \boldsymbol{\Sigma} \mathbf{w}_c)^{1/2}} \right| \\
= & E_1 + E_2.
\end{aligned}$$

Now,

$$\begin{aligned}
E_1 = & \left| \frac{\hat{\mathbf{w}}_c^\top (\boldsymbol{\mu}_2 - \hat{\boldsymbol{\mu}}_2)}{(\hat{\mathbf{w}}_c^\top \boldsymbol{\Sigma} \hat{\mathbf{w}}_c)^{1/2}} \right| \leq \frac{\| \hat{\mathbf{w}}_c \|_1}{\| \hat{\mathbf{w}}_c \|_2 . c_0} \| \hat{\boldsymbol{\mu}}_2 - \boldsymbol{\mu}_2 \|_\infty \\
\leq & \sqrt{\| \hat{\mathbf{w}}_c \|_0} O_p(\sqrt{\log p/n_2}) = O_p(\sqrt{\hat{s}_c \log p/n_2})
\end{aligned}$$

and

$$\begin{aligned}
E_2 = & \left| \frac{\hat{\mathbf{w}}_c^\top \hat{\boldsymbol{\mu}}_d}{2(\hat{\mathbf{w}}_c^\top \boldsymbol{\Sigma} \hat{\mathbf{w}}_c)^{1/2}} - \frac{\mathbf{w}_c^\top \boldsymbol{\mu}_d}{2(\mathbf{w}_c^\top \boldsymbol{\Sigma} \mathbf{w}_c)^{1/2}} \right| \\
= & \left| \frac{\hat{\mathbf{w}}_c^\top \hat{\boldsymbol{\mu}}_d - \hat{\mathbf{w}}_c^\top \boldsymbol{\mu}_d + \hat{\mathbf{w}}_c^\top \boldsymbol{\mu}_d}{2(\hat{\mathbf{w}}_c^\top \boldsymbol{\Sigma} \hat{\mathbf{w}}_c)^{1/2}} - \frac{\mathbf{w}_c^\top \boldsymbol{\mu}_d}{2(\mathbf{w}_c^\top \boldsymbol{\Sigma} \mathbf{w}_c)^{1/2}} \right| \\
\leq & \left| \frac{\hat{\mathbf{w}}_c^\top (\hat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)}{2(\hat{\mathbf{w}}_c^\top \boldsymbol{\Sigma} \hat{\mathbf{w}}_c)^{1/2}} \right| + \left| \frac{\hat{\mathbf{w}}_c^\top \boldsymbol{\mu}_d}{2(\hat{\mathbf{w}}_c^\top \boldsymbol{\Sigma} \hat{\mathbf{w}}_c)^{1/2}} - \frac{\mathbf{w}_c^\top \boldsymbol{\mu}_d}{2(\mathbf{w}_c^\top \boldsymbol{\Sigma} \mathbf{w}_c)^{1/2}} \right| \\
\leq & \frac{\| \hat{\mathbf{w}}_c \|_1}{\| \hat{\mathbf{w}}_c \|_2} \frac{\| \hat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d \|_\infty}{\min_j \lambda_j} + E_3 \\
\leq & \sqrt{\| \hat{\mathbf{w}}_c \|_0} \frac{\| \hat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d \|_\infty}{c_0} + E_3 = O_p(\sqrt{\hat{s}_c \log p/n_2}) + E_3.
\end{aligned}$$

According to the same notations in [13], let $f_0(\mathbf{w}) = \mathbf{w}^\top \boldsymbol{\mu}_d/(\mathbf{w}^\top \boldsymbol{\Sigma}\mathbf{w})^{1/2}$, $f_1(\mathbf{w}) = \mathbf{w}^\top \hat{\boldsymbol{\mu}}_d/(\mathbf{w}^\top \boldsymbol{\Sigma}\mathbf{w})^{1/2}$, and $f_2(\mathbf{w}) = \mathbf{w}^\top \hat{\boldsymbol{\mu}}_d/(\mathbf{w}^\top \widehat{\boldsymbol{\Sigma}}\mathbf{w})^{1/2}$. By the proof of Theorem 1 of [13], we have

$$\begin{aligned} E_3 &= \left| \frac{\hat{\mathbf{w}}_c^\top \boldsymbol{\mu}_d}{2(\hat{\mathbf{w}}_c^\top \boldsymbol{\Sigma}\hat{\mathbf{w}}_c)^{1/2}} - \frac{\mathbf{w}_c^\top \boldsymbol{\mu}_d}{2(\mathbf{w}_c^\top \boldsymbol{\Sigma}\mathbf{w}_c)^{1/2}} \right| = \frac{1}{2}|f_0(\hat{\mathbf{w}}_c) - f_0(\mathbf{w}_c)| \\ &\leq |f_0(\hat{\mathbf{w}}_c) - f_1(\hat{\mathbf{w}}_c)| + |f_1(\hat{\mathbf{w}}_c) - f_2(\hat{\mathbf{w}}_c)| + |f_2(\hat{\mathbf{w}}_c) - f_0(\mathbf{w}_c)| \\ &= O_p(\sqrt{\hat{s}_c \log p/n_2}) + O_p(c^2\sqrt{\log p/n_1}) + O_p\left(\sqrt{\max\{s_c,\ s_c^{(1)}\}\log p/n_2}\right). \end{aligned}$$

Therefore, we have

$$E_1 + E_2 + E_3 = O_p(c^2\sqrt{\log p/n_1}) + O_p\left(\sqrt{\max\{s_c,\ s_c^{(1),\ \hat{s}_c}\}\log p/n_2}\right),$$

and finally

$$\left|\Pi_2^{\text{ROAD}}(\mathcal{D}_n; c) - \Pi_2^{\text{orc}}(c)\right| = O_p(c^2\sqrt{\log p/n_1}) + O_p\left(\sqrt{\max\{s_c,\ s_c^{(1),\ \hat{s}_c}\}\log p/n_2}\right)$$

Similarly, the same result holds for $|\Pi_1^{\text{ROAD}}(\mathcal{D}_n; c) - \Pi_1^{\text{orc}}(c)|$, and this completes the proof. ■

Appendix C: Remaining proofs

In this Appendix, we provide the proofs of our claim in Remark 2.1, and also the proofs of Propositions 3.1 and 3.2.

Proof of the Claim in Remark 2.1. Recall $\mathcal{I}_i$, $i = 1, ..., 6$, defined in Theorem 2.1. When p is fixed with respect to the sample size, and $n_2 = o(n_1)$, then as $n_1, n_2 \to \infty$, we have,

$$\begin{aligned} Var(\mathcal{I}_1) = \frac{2p}{n_1^2} \to 0\ ,\ Var(\mathcal{I}_2) = \frac{2p}{n_2^2} \to 0\ ,\ Var(\mathcal{I}_3) = \frac{\Delta_p^2}{n_2} \to 0\ , \\ Var(\mathcal{I}_4) = \frac{\Delta_p^2}{n_1} \to 0\ ,\ Var(\mathcal{I}_5) = \frac{n^2 p}{n_1^2 n_2^2} \to 0\ ,\ Var(\mathcal{I}_6) = \frac{\Delta_p^2}{n_2} \to 0. \end{aligned}$$

On the other hand, $\mathbb{E}(\mathcal{I}_1) = \frac{p}{n_1}$, $\mathbb{E}(\mathcal{I}_2) = \frac{p}{n_2}$, $\mathbb{E}(\mathcal{I}_3) = \mathbb{E}(\mathcal{I}_4) = 0$, $\mathbb{E}(\mathcal{I}_5) = \frac{np}{n_1 n_2}$, and $\mathbb{E}(\mathcal{I}_6) = 0$. Thus, by following the proof of Theorem 2.1, we have

$$\begin{aligned} \frac{\Psi_1^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)}{\sqrt{\Upsilon^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)}} &= \frac{\frac{p}{n_1} - \frac{p}{n_2} + o_p(1) - \Delta_p^2}{2\left\{\frac{np}{n_1 n_2} + o_p(1) + \Delta_p^2\right\}^{1/2}} = -\frac{1}{2}\Delta_p + o_p(1), \\ \frac{\Psi_2^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)}{\sqrt{\Upsilon^{\text{LDA}}(\hat{\boldsymbol{\theta}}_n)}} &= \frac{-\frac{p}{n_1} + \frac{p}{n_2} + o_p(1) - \Delta_p^2}{2\left\{\frac{np}{n_1 n_2} + o_p(1) + \Delta_p^2\right\}^{1/2}} = -\frac{1}{2}\Delta_p + o_p(1). \end{aligned}$$

Therefore, $\Pi_k^{\text{LDA}}(\mathcal{D}_n)/\Pi^{\text{opt}} \xrightarrow{p} 1$, for $k = 1, 2$. ■

Proof of Proposition 3.1. The MCRs of $\delta_0^{\text{Msplit-HR}}$ are given by

$$\Pi_{0,k}^{\text{Msplit-HR}}(\mathcal{D}_n) = \Phi\left(\frac{\Psi_{0,k}^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n)}{\sqrt{\Upsilon_0^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n)}}\right), \quad k = 1, 2,$$

where

$$\Psi_{0,k}^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) = (-1)^{(k+1)}\frac{1}{\mathcal{L}}\sum_{\ell=1}^{\mathcal{L}}\sum_{j=1}^{p} r_j(\boldsymbol{\mu}_k; \hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}) h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)}).$$

and

$$\Upsilon_0^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{\mathcal{L}^2}\sum_{\ell=1}^{\mathcal{L}}\sum_{j=1}^{p}\sigma_j^2\left(\hat{\mu}_{dj,\ell}^{(2)}/\hat{\sigma}_{j,\ell}^{(2),2}\right)^2 h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)}).$$

Now, due to the independence property of $\mathcal{D}_{n,\ell}^{(1)}$ and $\mathcal{D}_{n,\ell}^{(2)}$, for each replication t, we have,

$$\begin{aligned} B_{0,n}^{\text{Msplit-HR}} &= \mathbb{E}\{\Psi_{0,1}^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) - \Psi_{0,2}^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n)\} \\ &= \frac{1}{\mathcal{L}}\sum_{\ell=1}^{\mathcal{L}}\sum_{j=1}^{p}\mathbb{E}\left\{r_j(\boldsymbol{\mu}_1; \hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}) + r_j(\boldsymbol{\mu}_2; \hat{\boldsymbol{\theta}}_{n,\ell}^{(2)})\right\}\mathbb{E}\left\{h_j(\hat{\boldsymbol{\theta}}_{n,\ell}^{(1)})\right\}, \end{aligned}$$

where $r_j(\boldsymbol{\mu}_k; \hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}) = \hat{\mu}_{dj,\ell}^{(2)}(\mu_{jk,\ell} - \hat{\mu}_{aj,\ell}^{(2)})/\hat{\sigma}_{j,\ell}^{(2),2}$. Hence

$$\bar{r}_n = \mathbb{E}\left\{r_j(\boldsymbol{\mu}_1; \hat{\boldsymbol{\theta}}_{n,\ell}^{(2)}) + r_j(\boldsymbol{\mu}_2; \hat{\boldsymbol{\theta}}_{n,\ell}^{(2)})\right\} = (\frac{1}{n_1'} - \frac{1}{n_2'})\frac{\Gamma(f_{n'} - 1)}{\Gamma(f_{n'})}f_{n'},$$

where $f_{n'} = n'/2 - 1$. ■

Proof of Proposition 3.2. The MCRs of $\delta^{\text{Msplit-HR}}$ in (3.8) are given by

$$\Pi_{0,k}^{\text{Msplit-HR}}(\mathcal{D}_n) = \Phi\left(\frac{\Psi_{0,k}^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n)}{\sqrt{\Upsilon_0^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n)}}\right), \ k = 1, 2,$$

where

$$\Psi_{0,k}^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) = \frac{(-1)^k}{\mathcal{L}}\sum_{\ell=1}^{\mathcal{L}}\tilde{\boldsymbol{\mu}}_{d,\ell}^{\top}\widetilde{\boldsymbol{\Sigma}}_{\ell}^{-1}(\tilde{\boldsymbol{\mu}}_{a,\ell} - \boldsymbol{\mu}_k),$$

and

$$\Upsilon_0^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{\mathcal{L}^2}\sum_{\ell=1}^{\mathcal{L}}\tilde{\boldsymbol{\mu}}_{d,\ell}^{\top}\widetilde{\boldsymbol{\Sigma}}_{\ell}^{-1}\boldsymbol{\Sigma}_{\ell}\widetilde{\boldsymbol{\Sigma}}_{\ell}^{-1}\tilde{\boldsymbol{\mu}}_{d,\ell}.$$

Hence,

$$\mathbb{E}\{\Psi_{0,1}^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n) - \Psi_{0,2}^{\text{Msplit-HR}}(\hat{\boldsymbol{\theta}}_n)\}$$

$$
\begin{aligned}
&= \frac{1}{\mathcal{L}}\sum_{\ell=1}^{\mathcal{L}}\mathbb{E}\left\{\mathbb{E}\left\{\tilde{\boldsymbol{\mu}}_{d,\ell}^{\top}\widetilde{\boldsymbol{\Sigma}}_{\ell}^{-1}(\boldsymbol{\mu}_{1,\ell}-\tilde{\boldsymbol{\mu}}_{a,\ell})-\tilde{\boldsymbol{\mu}}_{d,\ell}^{\top}\widetilde{\boldsymbol{\Sigma}}_{\ell}^{-1}(\tilde{\boldsymbol{\mu}}_{a,\ell}-\boldsymbol{\mu}_{2,\ell})\middle|\mathcal{D}_{n,\ell}^{(1)}\right\}\right\}\\
&= \frac{1}{\mathcal{L}}\sum_{\ell=1}^{\mathcal{L}}\mathbb{E}\left\{\mathbb{E}\left\{\tilde{\boldsymbol{\mu}}_{d,\ell}^{\top}\widetilde{\boldsymbol{\Sigma}}_{\ell}^{-1}(\boldsymbol{\mu}_{1,\ell}+\boldsymbol{\mu}_{2,\ell}-2\tilde{\boldsymbol{\mu}}_{a,\ell})\right\}\right\}\\
&= \frac{1}{\mathcal{L}}\sum_{\ell=1}^{\mathcal{L}}\mathbb{E}\{\bar{r}_{n,\ell}\}.
\end{aligned}
$$

The second equation follows from the independence property of $\mathcal{D}_{n,\ell}^{(1)}$ and $\mathcal{D}_{n,\ell}^{(2)}$, for each ℓ. Under normal assumption for the distribution of features, the matrix $\widetilde{\boldsymbol{\Sigma}}_{\ell}^{-1}$ has the Inverse Wishart distribution with parameters $\boldsymbol{\Sigma}_{\ell}^{-1}$ and $n'-2$, where $\boldsymbol{\Sigma}_{\ell}$ is the covariance matrix corresponding to the features included in $\mathcal{S}_{n,\ell}^{(1)}$, Thus, if $n'-3>|\mathcal{S}_{n,\ell}^{(1)}|$, then $\mathbb{E}\{\widetilde{\boldsymbol{\Sigma}}_{\ell}^{-1}\}=\frac{n'-2}{n'-2-|\mathcal{S}_{n,\ell}^{(1)}|-1}\boldsymbol{\Sigma}_{\ell}^{-1}$, and

$$
\begin{aligned}
\bar{r}_{n,\ell} &= \mathbb{E}\{\tilde{\boldsymbol{\mu}}_{d,\ell}^{\top}\widetilde{\boldsymbol{\Sigma}}_{\ell}^{-1}(\boldsymbol{\mu}_{1,\ell}+\boldsymbol{\mu}_{2,\ell}-2\tilde{\boldsymbol{\mu}}_{a,\ell})\}\\
&= tr\left\{\boldsymbol{\Sigma}_{\ell}^{-1}\frac{n'-2}{n'-3-|\mathcal{S}_{n,\ell}^{(1)}|}\boldsymbol{\Sigma}_{\ell}(\frac{1}{n_1'}-\frac{1}{n_2'})\right\}\\
&= \frac{n'-2}{n'-3-|\mathcal{S}_{n,\ell}^{(1)}|}|\mathcal{S}_{n,\ell}^{(1)}|(\frac{1}{n_1'}-\frac{1}{n_2'}).
\end{aligned}
$$

and the result follows. ■

Acknowledgments

We would like to thank the editor Professor Domenico Marinucci, an associate editor, and two referees for their insightful comments and suggestions that improved the quality of this paper. We thank the National High Performance Computing Center (NHPCC) at Isfahan University of Technology for their computational support to conduct our numerical experiments. Arezou Mojiri is grateful to (late) Soroush Alimoradi and also Ali Rejali for their help and constant support during her graduate studies.

References

[1] Ahn, J. and J. Marron (2010). The maximal data piling direction for discrimination. Biometrika 97, 254–259. MR2594434

[2] Bach, M., A. Werner, J. Żywiec, and W. Pluskiewicz (2017). The study of under-and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. Information Sciences 384, 174–190.

[3] Bak, B. A. and J. L. Jensen (2016). High dimensional classifiers in the imbalanced case. Computational Statistics & Data Analysis 98, 46–59. MR3458521

[4] 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, 989–1010. MR2108040

[5] Bickel, P. J. and E. Levina (2008a). Covariance regularization by thresholding. Annals of Statistics 36, 2577–2604. MR2485008

[6] Bickel, P. J. and E. Levina (2008b). Regularized estimation of large covariance matrices. Annals of Statistics 36, 199–227. MR2387969

[7] Blagus, R. and L. Lusa (2010). Class prediction for high-dimensional class-imbalanced data. BMC Bioinformatics 11, 1–17.

[8] Blagus, R. and L. Lusa (2013). SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 14.

[9] Bolton, R. J. and D. J. Hand (2002). Statistical fraud detection: A review. Statistical science 17, 235–249. MR1963313

[10] Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.

[11] Delaigle, A. and P. Hall (2012). Effect of heavy tails on ultra high dimensional variable ranking methods. Statistica Sinica 22, 909–932. MR2987477

[12] Fan, J. and Y. Fan (2008). High dimensional classification using features annealed independence rules. Annals of Statistics 36, 2605—-2637. MR2485009

[13] 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, 745–771. MR2965958

[14] Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70, 849–911. MR2530322

[15] Feng, Y., M. Zhou, and X. Tong (2020). Imbalanced classification: an objective-oriented review. arXiv preprint arXiv:2002.04592.

[16] Gaynanova, I., M. Kolar, et al. (2015). Optimal variable selection in multi-group sparse discriminant analysis. Electronic Journal of Statistics 9, 2007–2034. MR3393602

[17] Gravier, E., G. Pierron, A. Vincent-Salomon, N. Gruel, V. Raynal, A. Savignoni, Y. De Rycke, J.-Y. Pierga, C. Lucchesi, F. Reyal, A. Fourquet, S. Roman-Roman, X. Radvanyi, François aand Sastre-Garau, B. Asselain, and O. Delattre (2010). A prognostic DNA signature for T1T2 node-negative breast cancer patients. Genes, Chromosomes and Cancer 49, 1125–1134.

[18] Guo, Y., T. Hastie, and R. Tibshirani (2006). Regularized linear discriminant analysis and its application in microarrays. Biostatistics 8, 86–100.

[19] Hall, P., J. S. Marron, and A. Neeman (2005). Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 427–444. MR2155347

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

MR2758497
[21] Iranmehr, A., H. Masnadi-Shirazi, and N. Vasconcelos (2019). Cost-sensitive support vector machines. Neurocomputing 343, 50–64.
[22] Li, Q. and J. Shao (2015). Sparse quadratic discriminant analysis for high dimensional data. Statistica Sinica 25, 457–473. MR3379082
[23] Li, Y., H. G. Hong, and Y. Li (2019). Multiclass linear discriminant analysis with ultrahigh-dimensional features. Biometrics 75, 1086–1097. MR4041813
[24] Meinshausen, N. and P. Bühlmann (2010). Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 417–473. MR2758523
[25] Meinshausen, N., L. Meier, and P. Bühlmann (2009). P-values for high-dimensional regression. Journal of the American Statistical Association 104, 1671–1681. MR2750584
[26] Nakayama, Y. (2020). Support vector machine and optimal parameter selection for high-dimensional imbalanced data. Communications in Statistics-Simulation and Computation, 1–16. MR4253829
[27] Nakayama, Y., K. Yata, and M. Aoshima (2017). Support vector machine and its bias correction in high-dimension, low-sample-size settings. Journal of Statistical Planning and Inference 191, 88–100. MR3679111
[28] Owen, A. B. (2007). Infinitely imbalanced logistic regression. Journal of Machine Learning Research 8, 761–773. MR2320678
[29] Pan, R., H. Wang, and R. Li (2016). Ultrahigh-dimensional multiclass linear discriminant analysis by pairwise sure independence screening. Journal of the American Statistical Association 111, 169–179. MR3494651
[30] Pang, H., T. Tong, and M. Ng (2013). Block-diagonal discriminant analysis and its bias-corrected rules. Statistical applications in genetics and molecular biology 12, 347–359. MR3101034
[31] Park, B.-J., S.-K. Oh, and W. Pedrycz (2013). The design of polynomial function-based neural network predictors for detection of software defects. Information Sciences 229, 40–57. MR3018718
[32] Qiao, X. and Y. Liu (2009). Adaptive weighted learning for unbalanced multicategory classification. Biometrics 65, 159–168. MR2665857
[33] Qiao, X., H. H. Zhang, Y. Liu, M. J. Todd, and J. S. Marron (2010). Weighted distance weighted discrimination and its asymptotic properties. Journal of the American Statistical Association 105, 401–414. MR2656058
[34] Qiao, X. and L. Zhang (2013). Distance-weighted support vector machine. arXiv preprint arXiv:1310.3003. MR3341331
[35] Qiao, X. and L. Zhang (2015). Flexible high-dimensional classification machines and their asymptotic properties. The Journal of Machine Learning Research 16, 1547–1572. MR3417790
[36] Ramaswamy, S., K. N. Ross, E. S. Lander, and T. R. Golub (2002). A molecular signature of metastasis in primary solid tumors. Nature genetics 33, 49.
[37] Ramey, J. (2016). Datamicroarray: collection of data sets for classification.
[38] Shao, J., Y. Wang, X. Deng, S. Wang, et al. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. Annals of

statistics 39, 1241–1265. MR2816353

[39] Tian, E., F. Zhan, R. Walker, E. Rasmussen, Y. Ma, B. Barlogie, and J. D. Shaughnessy Jr (2003). The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. New England Journal of Medicine 349, 2483–2494.

[40] 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, 6567–6572.

[41] Verbeke, W., K. Dejaeger, D. Martens, J. Hur, and B. Baesens (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research 218, 211–229.

[42] 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, 753–772. MR2867457

[43] Xie, J., M. Hao, W. Liu, and Y. Lin (2020). Fused variable screening for massive imbalanced data. Computational Statistics & Data Analysis 141, 94–108. MR3980510

[44] Zhu, M., W. Su, and H. A. Chipman (2006). Lago: A computationally efficient approach for statistical detection. Technometrics 48, 193–205. MR2277674

[45] Zong, W., G.-B. Huang, and Y. Chen (2013). Weighted extreme learning machine for imbalance learning. Neurocomputing 101, 229–242.