Journal of Machine Learning Research 22 (2021) 1-93 Submitted 6/20; Revised 1/21; Published 1/21

RaSE: Random Subspace Ensemble Classification

Ye Tian YE.T@COLUMBIA.EDU
Department of Statistics
Columbia University
New York, NY 10027, USA

Yang Feng YANG.FENG@NYU.EDU
Department of Biostatistics, School of Global Public Health
New York University
New York, NY 10003, USA

Editor: Jie Peng

Abstract

We propose a flexible ensemble classification framework, *Random Subspace Ensemble* (RaSE), for sparse classification. In the RaSE algorithm, we aggregate many weak learners, where each weak learner is a base classifier trained in a subspace optimally selected from a collection of random subspaces. To conduct subspace selection, we propose a new criterion, *ratio information criterion* (RIC), based on weighted Kullback-Leibler divergence. The theoretical analysis includes the risk and Monte-Carlo variance of the RaSE classifier, establishing the screening consistency and weak consistency of RIC, and providing an upper bound for the misclassification rate of the RaSE classifier. In addition, we show that in a high-dimensional framework, the number of random subspaces needs to be very large to guarantee that a subspace covering signals is selected. Therefore, we propose an iterative version of the RaSE algorithm and prove that under some specific conditions, a smaller number of generated random subspaces are needed to find a desirable subspace through iteration. An array of simulations under various models and real-data applications demonstrate the effectiveness and robustness of the RaSE classifier and its iterative version in terms of low misclassification rate and accurate feature ranking. The RaSE algorithm is implemented in the R package `RaSEn` on CRAN.

Keywords: Random Subspace Method, Ensemble Classification, Sparsity, Information Criterion, Consistency, Feature Ranking, High Dimensional Data

1. Introduction

Ensemble classification is a very popular framework for carrying out classification tasks, which typically combines the results of many weak learners to form the final classification. It aims at improving both the accuracy and stability of weak classifiers and usually leads to a better performance than the individual weak classifier (Rokach, 2010). Two prominent examples of ensemble classification are bagging (Breiman, 1996, 1999) and random forest (Breiman, 2001), which focused on decision trees and aimed to improve the performance by bootstrapping training data and/or randomly selecting the splitting dimension in different trees, respectively. Boosting is another example that converts a weak learner that performs only slightly better than random guessing into a strong learner achieving arbitrary accu-

racy (Freund and Schapire, 1995). Recently, several new ensemble ideas appeared. Blaser and Fryzlewicz (2016) aggregate decision trees with random rotations to get an ensemble classifier. In particular, it randomly rotates the feature space each time prior to fitting the decision tree. Random rotations make it possible for the tree-based classifier to arrive at oblique solutions, increasing the flexibility of the decision trees. As a variant, to make the ensembles favor simple base learners, Blaser and Fryzlewicz (2019) proposed a regularized random rotation algorithm. Another popular framework of ensemble classifiers is via random projection. Durrant and Kabán (2015) studied a random projection ensemble classifier with linear discriminant analysis (LDA) and developed its theoretical properties. Furthermore, Cannings and Samworth (2017) proposed a very general framework for random projection ensembles. Each weak learner first randomly projects the original feature space into a low-dimensional subspace, then the base classifier is trained in that new space. The choice of the base classifier is flexible, and it was shown that the random projection ensemble classifier performs competitively and has desirable theoretical properties. There are two key aspects of their framework. One is that since naïvely aggregating all projections might lead to a poor performance, they first select some "good" projections and only aggregate these ones. The other key idea is that they tune the decision threshold instead of applying the naïve majority vote. These two ideas will also appear in our framework (to be proposed). To make the random projection include more important features in the linear combinations, Mukhopadhyay and Dunson (2019) proposed a targeted random projection ensemble approach, which includes each variable with probability proportional to their marginal utilities.

Another example of ensemble classification is the random subspace method, which was first studied in the context of decision trees (Ho, 1998). As the name suggests, it randomly selects a feature subset and grows each tree within the chosen subspace. A similar idea is used in random forest when we restrict the splitting of each tree to a subset of features. The random subspace method is closely related to other aggregation-based approaches, including the bootstrap procedure for features (Boot and Nibbering, 2020). Also, as Cannings and Samworth (2017) pointed out, the random subspace method can be regarded as the random projection ensemble classification method when the projection space is restricted to be axis-aligned. Compared to other ensemble approaches, the random subspace method keeps the data structure via sticking to the original features, which can be helpful for interpretation and provide a direct way for feature ranking. It has been coupled with various base classifiers, including linear discriminant analysis (Skurichina and Duin, 2002), k-nearest neighbor classifier (Bay, 1998), and combined with other techniques such as boosting (García-Pedrajas and Ortiz-Boyer, 2008). A related approach is random partition (Ahn et al., 2007), where the whole space is partitioned into multiple parts. Bryll et al. (2003) introduced optimization ideas into the framework of the random subspace method, and selected optimal subspaces by evaluating how the corresponding fitted models performed on the training data. Despite these developments, most existing works do not have theoretical support, and the research on the link between random subspace and feature ranking is scarce to the best of our knowledge. Furthermore, the existing literature usually considers the ensemble of all generated random subspaces, which may not be a wise idea in the sparse classification scheme as many random subspaces will contain no signals. Our new ensemble framework on random subspaces is designed to tackle the sparse classification

problems with theoretical guarantees. Instead of naively aggregating all generated random subspaces, we divide them into groups and only keep the "optimally" performing subspace inside each group to construct the ensemble classifier.

Feature ranking and selection are of critical importance in many real-world applications. For example, in disease diagnosis, beyond getting an accurate prediction for patients, we are also interested in understanding how each feature contributes to our prediction, which can facilitate the advancement of medical science. It has been widely acknowledged that in many high-dimensional classification problems, we only have a handful of useful features, with the rest being irrelevant ones. This is sometimes referred to as the sparse classification problem, which we briefly review next. Bickel et al. (2004) showed that linear discriminant analysis (LDA) is equivalent to random guessing in the worst scenario when the sample size is smaller than the dimensionality. Exploiting the underlying sparsity plays a significant role in improving the performance of the classic methods, including the LDA and the quadratic discriminant analysis (QDA) (Mai et al., 2012; Jiang et al., 2018; Hao et al., 2018; Fan et al., 2012; Shao et al., 2011; Fan et al., 2015; Li and Shao, 2015). While those methods work well under their corresponding models, it is not clear how to conduct feature ranking with other types of base classifiers. In this work, we propose a flexible ensemble classification framework, named *Random Subspace Ensemble (RaSE)*, which can be combined with any base classifiers and provide feature ranking as a by-product.

RaSE is a flexible ensemble classification framework, the main mechanism of which is briefly described as below. Suppose the observation pair $(\boldsymbol{x}, y)$ takes values from $\mathcal{X} \times \{0, 1\}$, where $\mathcal{X}$ is an open subset of $\mathbb{R}^p$, p is a positive integer and y is the class label. Assume the training set consists of n observation pairs $\{(\boldsymbol{x_i}, y_i), i = 1, \ldots, n\}$. We use $C_n^{S-\mathcal{T}}(\boldsymbol{x}) \in \{0, 1\}$ to represent the prediction result of the classifier trained with only features in subset S when the base classifier is $\mathcal{T}$. For the j-th ($j \in \{1, \ldots, B_1\}$) weak learner, B_2 random subspaces $\{S_{jk}\}_{k=1}^{B_2}$ are generated and the optimal one S_{j*} is selected according to some criterion to be specified. Then this weak learner will be trained by using only the slice of training samples in this subspace S_{j*}. Finally the B_1 weak classifiers $C_n^{S_{1*}-\mathcal{T}}, \ldots, C_n^{S_{B_1*}-\mathcal{T}}$ are aggregated to form the decision function

$$C_n^{RaSE}(\boldsymbol{x}) = \mathbb{1}\left(\frac{1}{B_1}\sum_{j=1}^{B_1} C_n^{S_{j*}-\mathcal{T}}(\boldsymbol{x}) > \alpha\right),$$

where α is a threshold to be determined. This framework contributes to the research of ensemble method and feature ranking in the following aspects. First, it admits a flexible framework, in which any classification algorithm can serve as the base classifier. Some examples include the standard LDA, QDA, k-nearest neighbor classifier (kNN), support vector machines (SVM), and decision trees. Second, the ensemble process naturally implies a ranking of the importance of variables via the frequencies they appear in the B_1 subspaces $\{S_{j*}\}_{j=1}^{B_1}$. For several specific sparse classification problems, equipped with a new information criterion named *ratio information criterion (RIC)*, RaSE is shown to cover the minimal discriminative set (to be defined later) for each weak learner with high probability when B_2 and sample size n are sufficiently large.

Although the RaSE framework shares some similarities with the random projection (RP) framework of Cannings and Samworth (2017), there are several essential differences

between them. First, the key workhorse behind RaSE is to search for a desirable subspace that covers the signals, which makes it amenable for feature ranking and selection. The RP framework, on the other hand, is not naturally designed for feature ranking. Second, a key condition (Assumption 2 in Cannings and Samworth (2017)) to guarantee the success of RP implies that using the criterion for choosing the optimal random projection, each random projection does not deviate much from the optimal one with a non-zero probability that is independent of n and p, which may not be satisfied under the high-dimensional setting. RaSE, however, assumes a set of conditions that explicitly take into account the high-dimensionality, which leads to the screening consistency and weak consistency. Third, we propose a new information criterion RIC with its theoretical properties analyzed under the high-dimensional setting. Fourth, motivated by the stringent requirement of a large B_2 for the vanilla RaSE (see Sections 3.2 and 3.5), we propose the iterative RaSE, which relaxes the requirement on B_2 by taking advantage of the feature ranking in preceding steps.

The rest of this paper is organized as follows. In Section 2, we first introduce the RaSE algorithm, and discuss some important concepts, including minimal discriminative set and RIC. At the end of Section 2, an iterative version of the RaSE algorithm is presented. In Section 3, theoretical properties of RaSE and RIC are investigated, including the impact of B_1 on the risk and Monte-Carlo variance of RaSE classifier, the screening consistency and weak consistency of RIC, the upper bound of expected misclassification rate, and the theoretical analysis for iterative RaSE algorithm. In Section 4, we discuss several important computational issues in the RaSE algorithm, tuning parameter selection, and how to apply the RaSE framework for feature ranking. Section 5 focuses on numerical studies in terms of extensive simulations and several real data applications, through which we compare RaSE with various competing methods. The results frequently feature RaSE classifiers among the top-ranked methods and also shows its effectiveness in feature ranking. Finally, we summarize our contributions and point out a few potential directions for future work in Section 6. We present some additional results for empirical studies in Appendix A, and all proofs are relegated to Appendix B.

2. Methodology

Recall that we have n pairs of observations $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\} \overset{i.i.d.}{\sim} (\boldsymbol{x}, y) \in \mathcal{X} \times \{0, 1\}$, where $\mathcal{X}$ is an open subset of $\mathbb{R}^p$, p is a positive integer and $y \in \{0, 1\}$ is the class label. We use $S_{\text{Full}} = \{1, \ldots, p\}$ to represent the whole feature set. We assume the marginal densities of $\boldsymbol{x}$ for class 0 ($y = 0$) and 1 ($y = 1$) exist and are denoted as $f^{(0)}$ and $f^{(1)}$, respectively. The corresponding probability measures they induce are denoted as $\mathrm{P}^{(0)}$ and $\mathrm{P}^{(1)}$. Thus, the joint distribution of $(\boldsymbol{x}, y)$ can be described in the following mixture model

$$\boldsymbol{x}|y = y_0 \sim (1 - y_0)f^{(0)} + y_0 f^{(1)}, y_0 = 0, 1, \tag{1}$$

where y is a Bernoulli variable with success probability $\pi_1 = 1-\pi_0 \in (0, 1)$. For any subspace S, we use $|S|$ to denote its cardinality. Denote $\mathrm{P}^{\boldsymbol{x}}$ as the probability measure induced by the marginal distribution of $\boldsymbol{x}$, which is in fact $\pi_0\mathrm{P}^{(0)} + \pi_1\mathrm{P}^{(1)}$. When restricting to the feature subspace S, the corresponding marginal densities of class 0 and 1 are denoted as $f_S^{(0)}$ and $f_S^{(1)}$, respectively.

2.1 Random Subspace Ensemble Classification (RaSE)

Motivated by Cannings and Samworth (2017), to train each weak learner (e.g., the j-th one), B_2 independent random subspaces are generated as $S_{j1}, \ldots, S_{jB_2}$. Then, according to some specific criterion (to be introduced in Section 2.3), the optimal subspace S_{j*} is selected and the j-th weak learner is trained only in S_{j*}. Subsequently, B_1 such weak classifiers $\{C_n^{S_{j*}-\mathcal{T}}\}_{j=1}^{B_1}$ are obtained. Finally, we aggregate outputs of $\{C_n^{S_{j*}-\mathcal{T}}\}_{j=1}^{B_1}$ to form the final decision function by taking a simple average. The whole procedure can be summarized in the following algorithm.

Algorithm 1: Random subspace ensemble classification (RaSE)

Input: training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, new data $\boldsymbol{x}$, subspace distribution $\mathcal{D}$, criterion $\mathcal{C}$, integers B_1 and B_2, type of base classifier $\mathcal{T}$

Output: predicted label $C_n^{RaSE}(\boldsymbol{x})$, the selected proportion of each feature $\boldsymbol{\eta}$

1 Independently generate random subspaces $S_{jk} \sim \mathcal{D}, 1 \le j \le B_1, 1 \le k \le B_2$
2 **for** $j \leftarrow 1$ **to** B_1 **do**
3 | Select the optimal subspace S_{j*} from $\{S_{jk}\}_{k=1}^{B_2}$ according to $\mathcal{C}$ and $\mathcal{T}$
4 **end**
5 Construct the ensemble decision function $\nu_n(\boldsymbol{x}) = \frac{1}{B_1}\sum_{j=1}^{B_1} C_n^{S_{j*}-\mathcal{T}}(\boldsymbol{x})$
6 Set the threshold $\hat{\alpha}$ according to (2)
7 Output the predicted label $C_n^{RaSE}(\boldsymbol{x}) = \mathbb{1}(\nu_n(\boldsymbol{x}) > \hat{\alpha})$, the selected proportion of each feature $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^T$ where $\eta_l = B_1^{-1}\sum_{j=1}^{B_1}\mathbb{1}(l \in S_{j*}), l = 1, \ldots, p$

In Algorithm 1, the subspace distribution $\mathcal{D}$ is chosen as a *hierarchical uniform distribution* over the subspaces by default. Specifically, with D as the upper bound of the subspace size[1], we first generate the subspace size d from the uniform distribution over $\{1, \ldots, D\}$. Then, the subspace S_{11} follows the uniform distribution over $\{S \subseteq S_{\text{Full}} : |S| = d\}$. In practice, the subspace distribution can be adjusted if we have prior information about the data structure.

In Step 6 of Algorithm 1, we set the decision threshold to minimize the empirical classification error on the training set,

$$\hat{\alpha} = \arg\min_{\alpha \in [0,1]} [\hat{\pi}_0(1 - \hat{G}_n^{(0)}(\alpha)) + \hat{\pi}_1 \hat{G}_n^{(1)}(\alpha)], \tag{2}$$

where

$$\begin{aligned} n_r &= \sum_{i=1}^n \mathbb{1}(y_i = r), r = 0, 1, \\ \hat{\pi}_r &= \frac{n_r}{n}, r = 0, 1, \\ \hat{G}_n^{(r)}(\alpha) &= \frac{1}{n_r}\sum_{i=1}^n \mathbb{1}(y_i = r)\mathbb{1}(\nu_n(\boldsymbol{x}_i) \le \alpha), r = 0, 1, \end{aligned}$$

1. How to set D in practice will be discussed in Section 4.1.

$$\nu_n(\boldsymbol{x}_i) = \frac{1}{B_1}\sum_{j=1}^{B_1} \mathbb{1}(C_n^{S_{j*}}(\boldsymbol{x}_i) = 1).$$

2.2 Minimal Discriminative Set

For sparse classification problems, it is of significance to accurately separate signals from noises. Motivated by Kohavi et al. (1997) and Zhang and Wang (2011), we define the discriminative set and study some of its properties as follows.

Definition 1 *A feature subset S is called a **discriminative set** if y is conditionally independent with $\boldsymbol{x}_{S^c}$ given $\boldsymbol{x}_S$, where $S^c = S_{\text{Full}} \setminus S$. We call S a **minimal discriminative set** if it has minimal cardinality among all discriminative sets.*

Assumption 1 *The densities $f^{(0)}$ and $f^{(1)}$ have the same support a.s. with respect to $\mathrm{P}^{\boldsymbol{x}}$.*

Remark 2 *Note that the existence of densities and the common support requirement are not necessary for the definition of the minimal discriminative set and the RaSE framework. We focus on the continuous distribution purely for notation convenience. We will discuss this assumption again after introducing our information criterion in Definition 6.*

Proposition 3 *Under Assumption 1, we can characterize the discriminative set using the marginal density ratio due to the following two facts.*

(i) If S is a discriminative set, then

$$\frac{f^{(1)}(\boldsymbol{x})}{f^{(0)}(\boldsymbol{x})} = \frac{f_S^{(1)}(\boldsymbol{x}_S)}{f_S^{(0)}(\boldsymbol{x}_S)}$$

almost surely with respect to $\mathrm{P}^{\boldsymbol{x}}$.

(ii) If for a feature subset S, there exists a function $h : \mathbb{R}^{|S|} \to [0, +\infty]$ such that

$$\frac{f^{(1)}(\boldsymbol{x})}{f^{(0)}(\boldsymbol{x})} = h(\boldsymbol{x}_S)$$

almost surely with respect to $\mathrm{P}^{\boldsymbol{x}}$, then S is a discriminative set and

$$h(\boldsymbol{x}_S) = \frac{f_S^{(1)}(\boldsymbol{x}_S)}{f_S^{(0)}(\boldsymbol{x}_S)}$$

almost surely with respect to $\mathrm{P}^{\boldsymbol{x}}$.

In general, there may exist more than one minimal discriminative sets. For instance, if two features are exactly the same, then more than one minimal discriminative sets may exist since we cannot distinguish between them. To rule out this type of degenerate scenario, we impose the uniqueness assumption for the minimum discriminative set.

Assumption 2 *There is only one minimal discriminative set, which is denoted as S^*. In addition, all discriminative sets cover S^*.*

In the classification problem, we are often interested in the risk of a classifier C. With the 0-1 loss, the risk or the misclassification rate is defined as

$$R(C) = \mathrm{E}[\mathbb{1}(C(\boldsymbol{x}) \neq y)] = \mathrm{P}(C(\boldsymbol{x}) \neq y).$$

The Bayes classifier

$$C_{Bayes}(\boldsymbol{x}) = \begin{cases} 1, & \mathrm{P}(y=1|\boldsymbol{x}) > \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

is known to achieve the minimal risk among all classifiers (Devroye et al., 2013). If only features in S are used, it will provide us a "local" Bayes classifier $C^S_{Bayes}(\boldsymbol{x}_S)$ which achieves the minimal risk among all classifiers using only features in S. In general, there is no guarantee that $R(C^S_{Bayes}) = R(C_{Bayes})$. Fortunately, the equation holds when S is a discriminative set.

Proposition 4 *For any discriminative set S, it holds that*

$$R(C^S_{Bayes}) = R(C^{S^*}_{Bayes}) = R(C_{Bayes}).$$

This direct result illustrates that covering S^* is sufficient to obtain performance as good as the Bayes classifier.

To clarify the notions above, we next take the two-class Gaussian settings as examples and investigate what S^* is in each case.

Example 1 (LDA) *Suppose $f^{(0)} \sim N(\boldsymbol{\mu}^{(0)}, \Sigma), f^{(1)} \sim N(\boldsymbol{\mu}^{(1)}, \Sigma)$, where Σ is positive definite. The log-density ratio is*

$$\log\left(\frac{f^{(0)}(\boldsymbol{x})}{f^{(1)}(\boldsymbol{x})}\right) = C - (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)})^T \Sigma^{-1} \boldsymbol{x},$$

where C is a constant independent of $\boldsymbol{x}$. By Proposition 3, the minimal discriminative set $S^* = \{j : [\Sigma^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)})]_j \neq 0\}$.

This definition is equivalent to that in Mai et al. (2012) for the LDA case.

Example 2 (QDA) *Suppose $f^{(0)} \sim N(\boldsymbol{\mu}^{(0)}, \Sigma^{(0)})$, $f^{(1)} \sim N(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$, where $\Sigma^{(0)}$ and $\Sigma^{(1)}$ are positive definite matrices with $\Sigma^{(0)} \neq \Sigma^{(1)}$, then the log-density ratio is*

$$\log\left(\frac{f^{(0)}(\boldsymbol{x})}{f^{(1)}(\boldsymbol{x})}\right) = C + \frac{1}{2}\boldsymbol{x}^T\left[(\Sigma^{(1)})^{-1} - (\Sigma^{(0)})^{-1}\right]\boldsymbol{x} + \left[(\Sigma^{(0)})^{-1}\boldsymbol{\mu}^{(0)} - (\Sigma^{(1)})^{-1}\boldsymbol{\mu}^{(1)}\right]^T \boldsymbol{x}, \tag{4}$$

where C is a constant independent of $\boldsymbol{x}$. Let $S^*_l = \{j : [(\Sigma^{(0)})^{-1}\boldsymbol{\mu}^{(0)} - (\Sigma^{(1)})^{-1}\boldsymbol{\mu}^{(1)}]_j \neq 0\}$, $S^*_q = \{j : [(\Sigma^{(1)})^{-1} - (\Sigma^{(0)})^{-1}]_{ij} \neq 0, \exists i\}$. The elements in S^*_l are often called variables with main effects while elements in S^*_q are called variables with quadratic effects (Hao et al., 2018; Fan et al., 2015; Jiang et al., 2018). By Proposition 3, the minimal discriminative set $S^* = S^*_l \cup S^*_q$.

Proposition 5 *If $f^{(0)} \sim N(\boldsymbol{\mu}^{(0)}, \Sigma^{(0)}), f^{(1)} \sim N(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$, where $\Sigma^{(0)}$ and $\Sigma^{(1)}$ are positive definite matrices, then the following conclusions hold:*

(i) The minimal discriminative set S^* is unique;

(ii) For any discriminative set S, we have $S \supseteq S^*$;

(iii) Any set $S \supseteq S^*$ is a discriminative set. This conclusion also holds without the Gaussian assumption.

2.3 Ratio Information Criterion (RIC)

As discussed in Section 2.2, it is desirable to identify the minimal discriminative set S^* for the classifier to achieve a low misclassification rate. Hence in Algorithm 1, it is important to apply a proper criterion to select the "optimal" subspace. In the variable selection literature, a criterion enjoying the property of correctly selecting the minimal discriminative set with high probability is often referred to as a "consistent" one. For model (1), Zhang and Wang (2011) proved that BIC, in conjunction with a backward elimination procedure, is selection consistent in the Gaussian mixture case. However, the BIC they investigated involved the joint log-likelihood function for $(\boldsymbol{x}, y)$, which involves estimating high-dimensional covariance matrices that could be problematic when p is close to or larger than n without additional structural assumptions.

Here we propose a new criterion, which enjoys the weak consistency under the general setting of (1), based on Kullback-Leibler divergence (Kullback and Leibler, 1951). Two asymmetric Kullback-Leibler divergences for densities f and g are defined as

$$\mathrm{KL}(f||g) = \mathrm{E}_{\boldsymbol{x}\sim f}\left[\log\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\right)\right], \mathrm{KL}(g||f) = \mathrm{E}_{\boldsymbol{x}\sim g}\left[\log\left(\frac{g(\boldsymbol{x})}{f(\boldsymbol{x})}\right)\right],$$

where $E_{\boldsymbol{x}\sim f}$ represents taking expectation with respect to $\boldsymbol{x} \sim f$. In binary classification model (1), marginal probabilities can be crucial because imbalanced marginal distributions can significantly compromise the performance of most standard learning algorithms (He and Garcia, 2009). Therefore, we consider a weighted version of two KL divergences for the marginal distributions $f_S^{(0)}$ and $f_S^{(1)}$ with subspace S, i.e. $\pi_0\mathrm{KL}(f_S^{(0)}||f_S^{(1)})+\pi_1\mathrm{KL}(f_S^{(1)}||f_S^{(0)})$. Denote by $\hat{f}_S^{(0)}, \hat{f}_S^{(1)}, \hat{\pi}_0, \hat{\pi}_1$ the estimated version via MLEs of parameters, then it holds that

$$\hat{\pi}_0\widehat{\mathrm{KL}}(f_S^{(0)}||f_S^{(1)}) = n^{-1}\sum_{i=1}^{n}\mathbb{1}(y_i=0)\cdot\log\left[\frac{\hat{f}_S^{(0)}(\boldsymbol{x}_{i,S})}{\hat{f}_S^{(1)}(\boldsymbol{x}_{i,S})}\right],$$
$$\hat{\pi}_1\widehat{\mathrm{KL}}(f_S^{(1)}||f_S^{(0)}) = n^{-1}\sum_{i=1}^{n}\mathbb{1}(y_i=1)\cdot\log\left[\frac{\hat{f}_S^{(1)}(\boldsymbol{x}_{i,S})}{\hat{f}_S^{(0)}(\boldsymbol{x}_{i,S})}\right].$$

Now, we are ready to introduce the following new criterion named *ratio information criterion* (RIC), with a proper penalty term.

Definition 6 *For model* (1)*, the ratio information criterion (RIC) for feature subspace S is defined as*

$$\mathrm{RIC}_n(S) = -2[\hat{\pi}_0\widehat{\mathrm{KL}}(f_S^{(0)}||f_S^{(1)}) + \hat{\pi}_1\widehat{\mathrm{KL}}(f_S^{(1)}||f_S^{(0)})] + c_n\cdot\deg(S), \tag{5}$$

where c_n is a function of sample size n and $\deg(S)$ is the degree of freedom corresponding to the model with subspace S.

Remark 7 *Assumption 1 is necessary to make sure RIC is well-defined. To see this, note that the existence of both Kullback-Leibler divergences requires (1)* $f^{(0)}(\boldsymbol{x}) = 0 \Rightarrow f^{(1)}(\boldsymbol{x}) = 0$ *a.s. with respect to* $\mathrm{P}^{(1)}$*; (2)* $f^{(1)}(\boldsymbol{x}) = 0 \Rightarrow f^{(0)}(\boldsymbol{x}) = 0$ *a.s. with respect to* $\mathrm{P}^{(0)}$*. This is equivalent to Assumption 1 when* $\pi_0, \pi_1 > 0$*.*

Note that although AIC is also motivated by the Kullback-Leibler divergence, it aims to minimize the KL divergence between the estimated density and the true density (Burnham and Anderson, 1998). In our case, however, the goal is to maximize the KL divergence between the conditional densities under two classes to achieve a greater separation.

Next, let's work out a few familiar examples where explicit expressions exist for RIC.

Proposition 8 (RIC for the LDA model) *Suppose* $f^{(0)} \sim N(\boldsymbol{\mu}^{(0)}, \Sigma)$, $f^{(1)} \sim N(\boldsymbol{\mu}^{(1)}, \Sigma)$*, where* Σ *is positive definite. The MLEs of the parameters are*

$$\hat{\boldsymbol{\mu}}_S^{(r)} = \frac{1}{n_r} \sum_{i=1}^{n} \mathbb{1}(y_i = r) \boldsymbol{x}_{i,S}, r = 0, 1,$$
$$\hat{\Sigma}_{S,S} = \frac{1}{n} \sum_{i=1}^{n} \sum_{r=0}^{1} \mathbb{1}(y_i = r) \cdot (\boldsymbol{x}_{i,S} - \hat{\boldsymbol{\mu}}_S^{(r)})(\boldsymbol{x}_{i,S} - \hat{\boldsymbol{\mu}}_S^{(r)})^T,$$

Then we have

$$\mathrm{RIC}_n(S) = -(\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)})^T \hat{\Sigma}_{S,S}^{-1} (\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)}) + c_n(|S| + 1).$$

Proposition 9 (RIC for the QDA model) *Suppose* $f^{(0)} \sim N(\boldsymbol{\mu}^{(0)}, \Sigma^{(0)})$, $f^{(1)} \sim N(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$*, where* $\Sigma^{(0)}$*,* $\Sigma^{(1)}$ *are positive definite but not necessarily equal. The MLEs of the estimators are as follows.* $\{\hat{\boldsymbol{\mu}}_S^{(r)}, r = 0, 1\}$ *are the same as in Proposition 8, and*

$$\hat{\Sigma}_{S,S}^{(r)} = \frac{1}{n_r} \sum_{i=1}^{n} \mathbb{1}(y_i = r) \cdot (\boldsymbol{x}_{i,S} - \hat{\boldsymbol{\mu}}_S^{(r)})(\boldsymbol{x}_{i,S} - \hat{\boldsymbol{\mu}}_S^{(r)})^T, r = 0, 1.$$

Then we have

$$\begin{aligned}\mathrm{RIC}_n(S) = &-(\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)})^T [\hat{\pi}_1 (\hat{\Sigma}_{S,S}^{(0)})^{-1} + \hat{\pi}_0 (\hat{\Sigma}_{S,S}^{(1)})^{-1}] (\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)}) \\ &+ \mathrm{Tr}[((\hat{\Sigma}_{S,S}^{(1)})^{-1} - (\hat{\Sigma}_{S,S}^{(0)})^{-1})(\hat{\pi}_1 \hat{\Sigma}_{S,S}^{(1)} - \hat{\pi}_0 \hat{\Sigma}_{S,S}^{(0)})] + (\hat{\pi}_1 - \hat{\pi}_0)(\log |\hat{\Sigma}_{S,S}^{(1)}| - \log |\hat{\Sigma}_{S,S}^{(0)}|) \\ &+ c_n \cdot \left[\frac{|S|(|S| + 3)}{2} + 1 \right].\end{aligned}$$

Note that the primary term of RIC for the LDA case, is the Mahalanobis distance (McLachlan, 1999), which is closely related to the Bayes error of LDA classifier (Efron, 1975). And for the QDA case, the KL divergence components contain three terms. The first term is similar to the Mahalanobis distance, representing the contributions of linear signals to the classification model. And the second and third terms represent the contributions of quadratic signals.

Note that the KL divergence can also be estimated by non-parametric methods including the k-nearest neighbor distance (Wang et al., 2009; Ganguly et al., 2018), which may

sometimes lead to more robust estimates than the parametric ones in our numerical experiments. Specifically, consider two samples $\{\boldsymbol{x}_{i,S}^{(0)}\}_{i=1}^{n_0} \overset{i.i.d.}{\sim} f_S^{(0)}$ and $\{\boldsymbol{x}_{i,S}^{(1)}\}_{i=1}^{n_1} \overset{i.i.d.}{\sim} f_S^{(1)}$, and write $\rho_{k_0,0}(\boldsymbol{x}_{j,S}^{(0)})$ for the Euclidean distance between $\boldsymbol{x}_{j,S}^{(0)}$ and its k_0-th nearest neighbor in the sample $\{\boldsymbol{x}_{i,S}^{(0)}\}_{i=1}^{n_0} \backslash \{\boldsymbol{x}_{j,S}^{(0)}\}$, and write $\rho_{k_1,1}(\boldsymbol{x}_{j,S}^{(0)})$ for the Euclidean distance between $\boldsymbol{x}_{j,S}^{(0)}$ and its k_1-th nearest neighbor in the sample $\{\boldsymbol{x}_{i,S}^{(1)}\}_{i=1}^{n_1}$. Wang et al. (2009) and Ganguly et al. (2018) defined the following asymptotic unbiased estimator given k_0 and k_1:

$$\widehat{\mathrm{KL}}(f_S^{(0)}||f_S^{(1)}) = \frac{|S|}{n_0}\sum_{i=1}^{n_0}\log\left(\frac{\rho_{k_0,0}(\boldsymbol{x}_{i,S}^{(0)})}{\rho_{k_1,1}(\boldsymbol{x}_{i,S}^{(0)})}\right) + \log\left(\frac{n_1}{n_0-1}\right) + \Psi(k_0) - \Psi(k_1), \tag{6}$$

where Ψ denotes the diGamma function (Abramowitz and Stegun, 1948). Similarly we can obtain estimate $\widehat{\mathrm{KL}}(f_S^{(1)}||f_S^{(0)})$. Besides, Berrett and Samworth (2019) proposed a weighted estimator based on (6) and investigated its efficiency. We will compare the performance of RaSE when using the estimate in (6) to that using parametric methods in simulation.

Another line of research for classification is to study the conditional distribution of $y|\boldsymbol{x}$. For this setup, there has been a rich literature on various information type criteria that involves the conditional log-likelihood function. Akaike information criterion (AIC) (Akaike, 1973) was shown to be inconsistent. It was demonstrated that Bayesian information criterion (BIC) is consistent under certain regularity conditions (Rao and Wu, 1989). Chen and Chen (2008, 2012) modified the definition of conventional BIC to form the extended BIC (eBIC) for the high-dimensional setting where p grows at a polynomial rate of n. Fan and Tang (2013) proved the consistency of a generalized information criterion (GIC) for generalized linear models in ultra-high dimensional space.

2.4 Iterative RaSE

The success of the RIC proposed in Section 2.3 relies on the assumption that the minimal discriminative set S^* appears in some of the B_2 subspaces for each weak learner. For sparse classification problems, the size of S^* can be very small compared to p. When p is large, the probability of generating a subset that covers S^* is quite low according to the hierarchical uniform distribution for subspaces. It turns out by using the selected frequency of each feature in B_1 subspaces $\{S_{j*}\}_{j=1}^{B_1}$ from Algorithm 1, we can improve the RaSE algorithm by running the RaSE algorithm again with a new hierarchical distribution for subspaces. In particular, we first calculate the percentage vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^T$ representing the proportion of each feature appearing among B_1 subspaces $\{S_{j*}\}_{j=1}^{B_1}$, where $\eta_l = B_1^{-1}\sum_{j=1}^{B_1}\mathbb{1}(l \in S_{j*}), l = 1, \ldots, p$. The new hierarchical distribution is specified as follows. In the first step, we generate the subspace size d from the uniform distribution over $\{1, \ldots, D\}$ as before. Before moving on to the second step, note that each subspace S can be equivalently represented as $\boldsymbol{J} = (J_1, \ldots, J_p)^T$, where $J_l = \mathbb{1}(l \in S), l = 1, \ldots, p$. Then, we generate $\boldsymbol{J}$ from a restrictive multinomial distribution with parameter $(p, d, \tilde{\boldsymbol{\eta}})$, where $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_p)^T$, $\tilde{\eta}_l = \eta_l\mathbb{1}(\eta_l > C_0/\log p) + \frac{C_0}{p}\mathbb{1}(\eta_l \le C_0/\log p)$, and the restriction is that $J_l \in \{0, 1\}, l = 1, \ldots, p$. Here C_0 is a constant.

Intuitively, this strategy can be repeated to increase the probability that signals in S^* are covered in the subspaces we generate. It could also lead to an improved feature

ranking according to the updated proportion of each feature $\boldsymbol{\eta}$. This will be verified via multiple simulation experiments in Section 5. This iterative RaSE algorithm is summarized in Algorithm 2.

A related idea was introduced by Mukhopadhyay and Dunson (2019) to generate random projections with probabilities proportional to the marginal utilities. One major difference in RaSE is that the feature importance is determined via their joint contributions in the selected subspaces.

Algorithm 2: Iterative RaSE (RaSE_T)

Input: training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, new data $\boldsymbol{x}$, initial subspace distribution $\mathcal{D}^{(0)}$, criterion $\mathcal{C}$, integers B_1 and B_2, the type of base classifier $\mathcal{T}$, the number of iterations T

Output: predicted label $C_n^{RaSE}(\boldsymbol{x})$, the proportion of each feature $\boldsymbol{\eta}^{(T)}$

for $t \leftarrow 0$ **to** T **do**
  Independently generate random subspaces $S_{jk}^{(t)} \sim \mathcal{D}^{(t)}, 1 \le j \le B_1, 1 \le k \le B_2$
  for $j \leftarrow 1$ **to** B_1 **do**
    Select the optimal subspace $S_{j*}^{(t)}$ from $\{S_{jk}^{(t)}\}_{k=1}^{B_2}$ according to $\mathcal{C}$ and $\mathcal{T}$
  end
  Update $\boldsymbol{\eta}^{(t)}$ where $\eta_l^{(t)} = B_1^{-1} \sum_{j=1}^{B_1} \mathbb{1}(l \in S_{j*}^{(t)}), l = 1, \ldots, p$
  Update $\mathcal{D}^{(t)} \leftarrow$ restrictive multinomial distribution with parameter $(p, d, \tilde{\boldsymbol{\eta}}^{(t)})$, where $\tilde{\eta}_l^{(t)} = \eta_l^{(t)} \mathbb{1}(\eta_l^{(t)} > C_0/\log p) + \frac{C_0}{p} \mathbb{1}(\eta_l^{(t)} \le C_0/\log p)$ and d is sampled from the uniform distribution over $\{1, \ldots, D\}$
end
Set the threshold $\hat{\alpha}$ according to (2)
Construct the ensemble decision function $\nu_n(\boldsymbol{x}) = \frac{1}{B_1} \sum_{j=1}^{B_1} C_n^{S_{j*}^{(T)} - \mathcal{T}}(\boldsymbol{x})$
Output the predicted label $C_n^{RaSE}(\boldsymbol{x}) = \mathbb{1}(\nu_n(\boldsymbol{x}) > \hat{\alpha})$ and $\boldsymbol{\eta}^{(T)}$

3. Theoretical Studies

In this section, we investigate various theoretical properties of RaSE, including the impact of B_1 on the risk of RaSE classifier and expectation of misclassification rate. Furthermore, we will demonstrate that RIC achieves weak consistency. Throughout this section, we allow the dimension p grows with sample size n.

To streamline the presentation, we first introduce some additional notations. In this paper, we have three different sources of randomness: (1) the randomness of the training data, (2) the randomness of the subspaces, and (3) the randomness of the test data. We will use the following notations to differentiate them.

- Analogous to Cannings and Samworth (2017), we write $\mathbf{P}$ and $\mathbf{E}$ to represent the probability and expectation with respect to the collection of $B_1 B_2$ random subspaces $\{S_{jk} : 1 \le j \le B_1, 1 \le k \le B_2\}$;
- $\mathbb{P}$ and $\mathbb{E}$ are used when the randomness comes from the training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$;

- We use P and E when considering all three sources of randomness.

Recall that in Algorithm 1, the decision function is

$$\nu_n(\boldsymbol{x}) = \frac{1}{B_1}\sum_{j=1}^{B_1} C_n^{S_{j*}}(\boldsymbol{x}).$$

For a given threshold $\alpha \in (0,1)$, the RaSE classifier is

$$C_n^{RaSE}(\boldsymbol{x}) = \begin{cases} 1, & \nu_n(\boldsymbol{x}) > \alpha, \\ 0, & \nu_n(\boldsymbol{x}) \le \alpha. \end{cases}$$

By the weak law of large numbers, as $B_1 \to \infty$, ν_n will converge in probability to its expectation

$$\mu_n(\boldsymbol{x}) = \mathbf{P}\left(C_n^{S_{1*}}(\boldsymbol{x}) = 1\right).$$

It should be noted that here as both the training data and the criterion $\mathcal{C}$ in Algorithm 1 are fixed, S_{1*} is a deterministic function of $\{S_{1k}\}_{k=1}^{B_2}$. Nevertheless, $\mu_n(\boldsymbol{x})$ is still random due to randomness of the new $\boldsymbol{x}$. Then, it is helpful to define the conditional cumulative distribution function of μ_n for class $0, 1$ respectively as $G_n^{(0)}(\alpha') = \mathrm{P}(\mu_n(\boldsymbol{x}) \le \alpha' | y = 0) = \mathrm{P}^{(0)}(\mu_n(\boldsymbol{x}) \le \alpha')$ and $G_n^{(1)}(\alpha') = \mathrm{P}(\mu_n(\boldsymbol{x}) \le \alpha' | y = 1) = \mathrm{P}^{(1)}(\mu_n(\boldsymbol{x}) \le \alpha')$. Since the distribution $\mathcal{D}$ of subspaces is discrete, $\mu_n(\boldsymbol{x})$ takes finite unique values almost surely, implying $G_n^{(0)}, G_n^{(1)}$ to be step functions. We denote the corresponding probability mass functions of $G_n^{(0)}$ and $G_n^{(1)}$ as $g_n^{(0)}$ and $g_n^{(1)}$, respectively. Since for any $\boldsymbol{x}$, $\nu_n(\boldsymbol{x}) \to \mu_n(\boldsymbol{x})$ as $B_1 \to \infty$, we consider the following randomized RaSE classifier in population

$$C_n^{RaSE*}(\boldsymbol{x}) = \begin{cases} 1, & \mu_n(\boldsymbol{x}) > \alpha, \\ 0, & \mu_n(\boldsymbol{x}) < \alpha, \\ \text{Bernoulli}\left(\frac{1}{2}\right), & \mu_n(\boldsymbol{x}) = \alpha. \end{cases}$$

as the infinite simulation RaSE classifier with $B_1 \to \infty$.

In the following sections, we would like to study different properties of C_n^{RaSE}. In Section 3.1, we condition on the training data and study the impact of B_1 via the relationship between test error of C_n^{RaSE} and C_n^{RaSE*} and Monte-Carlo variance $\mathbf{Var}(R(C_n^{RaSE}))$, which can reflect the stability of RaSE classifier as B_1 increases. It will be demonstrated that conditioned on training data, both the difference between the test errors of C_n^{RaSE} and C_n^{RaSE*}, and the Monte-Carlo variance of C_n^{RaSE}, converge to zero as $B_1 \to \infty$ for almost every threshold $\alpha \in (0,1)$ at an exponential rate. In Section 3.3, we will prove an upper bound for the expected misclassification rate $R(C_n^{RaSE})$ with respect to all the randomness for fixed threshold α. Next, we introduce several scaling notations. For two sequences a_n and b_n, we use $a_n = o(b_n)$ or $a_n \ll b_n$ to denote $|a_n/b_n| \to 0$; $a_n = O(b_n)$ or $a_n \lesssim b_n$ to denote $|a_n/b_n| < \infty$. The corresponding stochastic scaling notations are o_p and O_p, where $a_n = o_p(b_n), a_n = O_p(b_n)$ imply that $|a_n/b_n| \xrightarrow{p} 0$ and for any $\epsilon > 0$, there exists $M > 0$ such that $\mathrm{P}(|a_n/b_n| > M) \le \epsilon$, $\forall n$. Also, we use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the smallest and largest eigenvalues of a square matrix A. For any vector $\boldsymbol{x} = (x_1, \ldots, x_p)^T$, the Euclidean norm $\|\boldsymbol{x}\| = \sqrt{\sum_i x_i^2}$. And for any matrix $A = (a_{ij})_{p\times p}$, we define the operator

norm $\|A\|_2 = \sup_{\|\boldsymbol{x}\|_2=1} \|A\boldsymbol{x}\|_2$, the infinity norm $\|A\|_\infty = \sup_i \sum_{j=1}^p |a_{ij}|$, the maximum norm $\|A\|_{\max} = \max_{i,j} |a_{ij}|$, and the Frobenius norm $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$.

3.1 Impact of B_1

In this section, we study the impact of B_1 by presenting upper bounds of the absolute difference between the test error of C_n^{RaSE} and C_n^{RaSE*}, and Monte-Carlo variance for RaSE when conditioned on the training data as B_1 grows. Note that the discrete distribution of random subspaces leads to both bounds vanishing at exponential rates, except for a finite set of thresholds α, which is very appealing.

Theorem 10 (Risk for the RaSE classifier conditioned on training data) *Denote $G_n(\alpha') = \pi_1 G_n^{(1)}(\alpha') - \pi_0 G_n^{(0)}(\alpha')$ and $\{\alpha_i\}_{i=1}^N$ represents the discontinuity points of G_n. Given training samples with size n, we have the following bound for expected misclassification rate of RaSE classifier with threshold α when $B_1 \to \infty$ as*

$$|\mathbf{E}[R(C_n^{RaSE})] - R(C_n^{RaSE*})| \le \begin{cases} O\left(\frac{1}{\sqrt{B_1}}\right), & \alpha \in \{\alpha_i\}_{i=1}^N, \\ \exp\{-C_\alpha B_1\}, & otherwise, \end{cases}$$

where $C_\alpha = 2 \min_{1\le i\le N} (|\alpha - \alpha_i|^2)$.

It shows that as $B_1 \to \infty$, the RaSE classifier C_n^{RaSE} and its infinite simulation version C_n^{RaSE*} achieve the same expected misclassification rate conditioned on the training data. Many similar results about the excess risk of randomized ensembles have been studied in literature (Cannings and Samworth, 2017; Lopes et al., 2019; Lopes, 2020). Next, we provide a similar upper bound for the MC variance of the RaSE classifier. Suppose the discontinuity points of $G_n^{(0)}$ and $G_n^{(1)}$ are $\{\alpha_i^0\}_{i=1}^{N_0}$ and $\{\alpha_j^1\}_{j=1}^{N_1}$, respectively.

Theorem 11 (MC variance for the RaSE classifier) *It holds that*

$$\mathbf{Var}[R(C_n^{RaSE})] \le \begin{cases} \frac{1}{2}\left[\pi_0 (g_n^{(0)}(\alpha))^2 + \pi_1 (g_n^{(1)}(\alpha))^2\right] + O\left(\frac{1}{\sqrt{B_1}}\right), & \alpha \in \{\alpha_i^{(0)}\}_{i=1}^{N_0} \cup \{\alpha_j^{(1)}\}_{j=1}^{N_1} \\ \exp\{-C_\alpha B_1\}, & otherwise, \end{cases}$$

where $C_\alpha = 2 \min_{\substack{1\le i\le N_0 \\ 1\le j\le N_1}} (|\alpha - \alpha_i^{(0)}|^2, |\alpha - \alpha_j^{(1)}|^2)$.

This theorem asserts that except for a finite set of threshold α, the MC variance of RaSE classifier shrinks to zero at an exponential rate.

3.2 Theoretical Properties of RIC

An important step of the RaSE classifier is the choice of an "optimal" subspace among B_2 subspaces for each of the B_1 weak classifiers. Before showing the screening consistency and weak consistency of RIC, we first present a proposition that explains the intuition of why RIC can succeed.

Proposition 12 *When Assumptions 1 and 2 hold, we have the following conclusions:*

(i) $\mathrm{KL}(f_S^{(0)}||f_S^{(1)}) = \mathrm{KL}(f_{S^*}^{(0)}||f_{S^*}^{(1)}), \mathrm{KL}(f_S^{(1)}||f_S^{(0)}) = \mathrm{KL}(f_{S^*}^{(1)}||f_{S^*}^{(0)})$ *hold for any* $S \supseteq S^*$*;*

(ii) $\pi_0\mathrm{KL}(f_S^{(0)}||f_S^{(1)}) + \pi_1\mathrm{KL}(f_S^{(1)}||f_S^{(0)}) < \pi_0\mathrm{KL}(f_{S^*}^{(0)}||f_{S^*}^{(1)}) + \pi_1\mathrm{KL}(f_{S^*}^{(1)}||f_{S^*}^{(0)})$ *if* $S \not\supseteq S^*$*;*

(iii) $\mathrm{KL}(f_{S^*}^{(0)}||f_{S^*}^{(1)}) = \sup_S \mathrm{KL}(f_S^{(0)}||f_S^{(1)}), \mathrm{KL}(f_{S^*}^{(1)}||f_{S^*}^{(0)}) = \sup_S \mathrm{KL}(f_S^{(1)}||f_S^{(0)})$.

From Proposition 12, if we define the population RIC without penalty as

$$\mathrm{RIC}(S) = -2\left[\pi_0\mathrm{KL}(f_S^{(0)}||f_S^{(1)}) + \pi_1\mathrm{KL}(f_S^{(1)}||f_S^{(0)})\right],$$

it can be easily seen that $\sup_{S:S\supseteq S^*} \mathrm{RIC}(S) = \mathrm{RIC}(S^*) < \inf_{S:S\not\supseteq S^*} \mathrm{RIC}(S)$. To successfully differentiate S^* from $\{S : S \not\supseteq S^*\}$ using RIC_n, we need to impose a condition on the minimum gap of RIC on S^* and that on S where $S \not\supseteq S^*$. Similar assumptions such as the "beta-min" condition appears in the high-dimensional variable selection literature (Bühlmann and Van De Geer, 2011). Denote

$$\psi(n,p,D) = \sqrt{\frac{D\log p + \kappa_1 \log D}{n}} \max\left\{D^{\kappa_1}(D^{\kappa_3} + D^{\kappa_4}), D^{\kappa_5}, D^{2\kappa_1+\kappa_2}\sqrt{\frac{D\log p + \kappa_1 \log D}{n}}\right\}.$$

The complete set of conditions is presented as follows.

Assumption 3 *Suppose densities* $f^{(0)}$ *and* $f^{(1)}$ *are in parametric forms with* $f^{(0)}(\boldsymbol{x}) = f^{(0)}(\boldsymbol{x};\boldsymbol{\theta})$*,* $f^{(1)}(\boldsymbol{x}) = f^{(1)}(\boldsymbol{x};\boldsymbol{\theta})$*, where* $\boldsymbol{\theta}$ *contains all parameters for both* $f^{(0)}$ *and* $f^{(1)}$*. Note that not all elements of* $\boldsymbol{\theta}$ *appear in* $f^{(0)}$ *or* $f^{(1)}$*. Denote the dimension of* $\boldsymbol{\theta}$ *as* p'*. As in Section 2.1,* D *represents the upper bound of the subspace size. Define* $L_S(\boldsymbol{x},\boldsymbol{\theta}) = \log\left(\frac{f_S^{(0)}(\boldsymbol{x}_S;\boldsymbol{\theta}_S)}{f_S^{(1)}(\boldsymbol{x}_S;\boldsymbol{\theta}_S)}\right)$*. Assume the following conditions hold, where* κ_1*,* κ_2*,* κ_3*,* κ_4*,* $\kappa_5 \geq 0$ *and* C_1*,* C_2*,* C_3*,* C_4*,* $C_5 > 0$ *are some universal constants which does not depend on* n*,* p *or* D*:*

(i) p' *is a function of* p *and* $p'(p) \lesssim p^{\kappa_1}$*;*

(ii) $\max\left[\mathrm{KL}(f^{(0)}||f^{(1)}), \mathrm{KL}(f^{(1)}||f^{(0)})\right] \lesssim (p^*)^{\kappa_1} \lesssim D^{\kappa_1}$*;*

(iii) There exists a family of functions $\{V_S(\{\boldsymbol{x}_{i,S}\}_{i=1}^n)\}_S$ *where* $V_S(\{\boldsymbol{x}_{i,S}\}_{i=1}^n) \in \mathbb{R}$*, such that for* $\forall\{\boldsymbol{x}_{i,S}\}_{i=1}^n \in \mathcal{X}$ *and subset* S *with* $|S| \leq D$*, there exists a constant* ζ *such that if* $\|\boldsymbol{\theta}'_S - \boldsymbol{\theta}_S\|_2 \leq \zeta$*, then*

$$\left\|\frac{1}{n}\sum_{i=1}^n \nabla^2_{\boldsymbol{\theta}_S} L_S(\boldsymbol{x}_i,\boldsymbol{\theta}')\right\|_{\max} \leq V_S(\{\boldsymbol{x}_{i,S}\}_{i=1}^n),$$

and the following tail probability bound holds for V_S*:*

$$\mathbb{P}\left(V_S(\{\boldsymbol{x}_{i,S}\}_{i=1}^n) > C_1 D^{\kappa_2}\right) \lesssim \exp\{-C_2 n\},$$

where $\boldsymbol{x}_{i,S} \sim f_S^{(0)}$ *or* $f_S^{(1)}$*;*

(iv) *Each component of* $\nabla_{\boldsymbol{\theta}_S} L_S(\boldsymbol{x}, \boldsymbol{\theta})$ *is* $\sqrt{2C_3}D^{\kappa_3}$*-subGaussian for any subset* S *with* $|S| \leq D$ *where* $\boldsymbol{x}_S \sim f_S^{(0)}$ *or* $\boldsymbol{x}_S \sim f_S^{(1)}$*, respectively;*

(v) $\sup_{S:|S|\leq D} \left\| \mathbb{E}_{\boldsymbol{x}_S \sim f_S^{(0)}} \nabla_{\boldsymbol{\theta}_S} L_S(\boldsymbol{x}, \boldsymbol{\theta}) \right\|_\infty$, $\sup_{S:|S|\leq D} \left\| \mathbb{E}_{\boldsymbol{x}_S \sim f_S^{(1)}} \nabla_{\boldsymbol{\theta}_S} L_S(\boldsymbol{x}, \boldsymbol{\theta}) \right\|_\infty \lesssim D^{\kappa_4}$*;*

(vi) $L_S(\boldsymbol{x}, \boldsymbol{\theta})$ *is a* $\sqrt{2C_4}D^{\kappa_5}$*-subGaussian variable, where* $\boldsymbol{x}_S \sim f_S^{(0)}$ *or* $\boldsymbol{x}_S \sim f_S^{(1)}$*;*

(vii) *Denoting the MLE of* $\boldsymbol{\theta}$ *based on subset* S *with* $|S| \leq D$ *as*

$$\hat{\boldsymbol{\theta}}_S = \arg\max_{\boldsymbol{\theta}_S} \sum_{i=1}^{n} \sum_{r=0}^{1} \mathbb{1}(y_i = r) \log f_S^{(r)}(\boldsymbol{x}_{i,S}; \boldsymbol{\theta}_S),$$

then when ϵ *is smaller than a positive constant, it holds that*

$$\mathbb{P}(\|\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S\|_\infty > \epsilon) \lesssim |S|^{\kappa_1} \exp\{-C_5 n\epsilon^2\} \Rightarrow \|\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S\|_\infty = O_p\left(\sqrt{\frac{\kappa_1 \log|S|}{n}}\right).$$

(viii) *The signal strength satisfies*

$$\Delta \coloneqq \inf_{\substack{S: S \not\supseteq S^* \\ |S| \leq D}} \mathrm{RIC}(S) - \mathrm{RIC}(S^*) \gg \psi(n, p, D),$$

where $\psi(n, p, D) = o(1)$.

Here condition (i) assumes the number of parameters in the model grows slower than a polynomial rate of the dimension. Condition (ii) assumes an upper bound for the two KL divergences. Conditions (iii)-(vi) are imposed to guarantee the accuracy of the second-order approximation for RIC. And condition (vii) is usually satisfied for common distribution families (Van der Vaart, 2000). The last condition is a requirement for the signal strength. A condition of this type is necessary to prove the consistency result for any information criterion.

For condition (viii), it imposes a constraint among n, p, and D. When $D \ll p$, it can be simplified as

$$D^\kappa \cdot \frac{\log p}{n} = o(1),$$

where κ is some positive constant. To cover all the signals in S^*, we require $D \geq p^*$. Therefore, an implied requirement for n, p, and p^* is

$$(p^*)^\kappa \cdot \frac{\log p}{n} = o(1),$$

which is similar to the conditions in the literature of variable selection. Here, we allow p to grow at an exponential rate of n.

To help readers understand these conditions better, we show that some commonly used conditions (presented in Assumption 4) for high-dimensional LDA model are sufficient for Assumption 3.(i)-(vii) to hold with the results presented in Proposition 14. Assumption 3.(viii) can be relaxed under the LDA model.

Assumption 4 (LDA model) *Suppose the following conditions are satisfied, where m, M, M' are constants:*

(i) $\lambda_{\min}(\Sigma) \geq m > 0, \|\Sigma\|_{\max} \leq M < \infty$;

(ii) $\|\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)}\|_\infty \leq M' < \infty$;

(iii) Denote $\boldsymbol{\delta}_S = \Sigma_{S,S}^{-1}(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}), \gamma = \inf_j |(\boldsymbol{\delta}_{S^*})_j| > 0$, *then*

$$\gamma^2 \gg D^2\sqrt{\frac{\log p}{n}} = o(1).$$

Remark 13 *Here, condition (i) constrains the eigenvalues of the common covariance matrix, which is similar to condition (C2) in Hao et al. (2018) and Wang (2009), condition (2) in Shao et al. (2011), condition (C4) in Li and Liu (2019). Condition (ii) imposes an upper bound for the maximal componentwise mean difference of the two classes, which is similar to condition (3) in Shao et al. (2011). Condition (iii) assumes a lower bound on the minimum signal strength γ, which is in a similar spirit to condition 2 in Mai et al. (2012).*

Proposition 14 *Suppose $f^{(0)} \sim N(\boldsymbol{\mu}^{(0)}, \Sigma), f^{(1)} \sim N(\boldsymbol{\mu}^{(1)}, \Sigma)$, where Σ is positive definite. If Assumption 4 holds, then Assumption 3.(i)-(vii) hold with $\boldsymbol{\theta}_S = ((\boldsymbol{\mu}_S^{(0)})^T, (\boldsymbol{\mu}_S^{(1)})^T, \mathrm{vec}(\Sigma_{S,S})^T)^T$ and $\kappa_1 = 2, \kappa_2 = \kappa_4 = 1, \kappa_3 = \kappa_5 = \frac{1}{2}$.*

The detailed proof for this proposition can be found in Appendix B. We are now ready to present the consistency result for RIC as defined in (5).

Theorem 15 (Consistency of RIC) *Under Assumptions 1-3, we have*

(i) If $\sup_{S:|S|\leq D} \deg(S) \cdot c_n/\Delta = o(1)$, *then the following screening consistency holds for RIC:*[2]

$$\mathbb{P}\left(\sup_{\substack{S:S\supseteq S^* \\ |S|\leq D}} \mathrm{RIC}_n(S) < \inf_{\substack{S:S\not\supseteq S^* \\ |S|\leq D}} \mathrm{RIC}_n(S)\right) \geq 1 - O\left(p^D D^{\kappa_1} \exp\left\{-Cn\left(\frac{\Delta}{D^{2\kappa_1+\kappa_2}}\right)\right\}\right)$$

$$- O\left(p^D D^{\kappa_1} \exp\left\{-Cn\left(\frac{\Delta}{D^{\kappa_1}(D^{\kappa_3}+D^{\kappa_4})}\right)^2\right\}\right) - O\left(p^D \exp\left\{-Cn\left(\frac{\Delta}{D^{\kappa_5}}\right)^2\right\}\right)$$

$$\to 1.$$

(ii) If in addition $c_n \gg \psi(n,p,D)$, then the weak consistency in the following sense holds for RIC:

$$\mathbb{P}\left(\mathrm{RIC}_n(S^*) = \inf_{S:|S|\leq D} \mathrm{RIC}_n(S)\right) \geq 1 - O\left(p^D D^{\kappa_1} \exp\left\{-Cn\left(\frac{c_n}{D^{2\kappa_1+\kappa_2}}\right)\right\}\right)$$

2. Note that here we assume that MLE of $\boldsymbol{\theta}$ is well-defined for all models under consideration.

$$- O\left(p^D D^{\kappa_1} \exp\left\{-Cn\left(\frac{c_n}{D^{\kappa_1}(D^{\kappa_3}+D^{\kappa_4})}\right)^2\right\}\right) - O\left(p^D \exp\left\{-Cn\left(\frac{c_n}{D^{\kappa_5}}\right)^2\right\}\right)$$
$$\to 1.$$

Corollary 16 *Denote* $p_{S^*} = \mathbf{P}(S_{11} \supseteq S^*) = \frac{1}{D}\sum_{p^* \le d \le D} \frac{\binom{p-p^*}{d-p^*}}{\binom{p}{d}}$, *where the subspaces are generated from the hierarchical uniform distribution. Assume the conditions stated in Assumption 1-2 hold, and in addition there holds*

$$B_2 p_{S^*} \gg 1.$$

If we select the optimal subspace by minimizing RIC as the criterion in RaSE where $D \ge p^*$, *then we have*

$$\mathrm{P}(S_{1*} \supseteq S^*) \ge \mathbb{P}\left(\sup_{\substack{S:S\supseteq S^* \\ |S|\le D}} \mathrm{RIC}_n(S) < \inf_{\substack{S:S\not\supseteq S^* \\ |S|\le D}} \mathrm{RIC}_n(S)\right) \cdot \mathbf{P}\left(\bigcup_{j=1}^{B_2}\{S_{1j} \supseteq S^*\}\right), \quad (7)$$

where

$$\mathbf{P}\left(\bigcup_{j=1}^{B_2}\{S_{1j} \supseteq S^*\}\right) = 1 - (1-p_{S^*})^{B_2} \ge 1 - O\left(\exp\{-B_2 p_{S^*}\}\right).$$

Remark 17 *Note that this corollary actually holds when we replace* RIC_n *with a general criterion* Cr_n *(smaller value leads to better subspace) with more discussions in Section 3.5. And for RIC, the bounds for the first probability on the right-hand side of* (7) *in Theorems 15, 18 and 20 can be plugged in to get the explicit bounds.*

Also, we want to point out that direct analyses of RIC for discriminant analysis models are also insightful and interesting. We can show similar consistency results as those in Theorem 15 from properties of discriminant analysis approach itself based on some common conditions used in literature about sparse discriminant analysis, instead of applying the general analysis of KL divergence.

Theorem 18 (LDA consistency) *For the LDA model, under Assumption 4, we have*

(i) If $Dc_n/\gamma^2 = o(1)$, *then the following screening consistency holds for RIC:*

$$\mathbb{P}\left(\sup_{\substack{S:S\supseteq S^* \\ |S|\le D}} \mathrm{RIC}_n(S) < \inf_{\substack{S:S\not\supseteq S^* \\ |S|\le D}} \mathrm{RIC}_n(S)\right) \ge 1 - O\left(p^2 \exp\left\{-Cn\left(\frac{\gamma^2}{D^2}\right)^2\right\}\right) \to 1.$$

(ii) If in addition $c_n \gg D^2\sqrt{\frac{\log p}{n}}$, *then RIC is weakly consistent:*

$$\mathbb{P}\left(\mathrm{RIC}_n(S^*) = \inf_{S:|S|\le D} \mathrm{RIC}_n(S)\right) \ge 1 - O\left(p^2 \exp\left\{-Cn\left(\frac{c_n}{D^2}\right)^2\right\}\right) \to 1.$$

Assumption 5 (QDA model) *Denote* $\Omega_{S,S}^{(r)} = (\Sigma_{S,S}^{(r)})^{-1}, r = 0, 1$. *Suppose the following conditions are satisfied, where* m, M, M' *are constants:*

(i) $\lambda_{\min}(\Sigma^{(r)}) \geq m > 0, \lambda_{\max}(\Sigma^{(r)}) \leq M < \infty, r = 0, 1$;

(ii) $\|\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)}\|_\infty \leq M' < \infty$;

(iii) Denote $\gamma_l = \inf_j \left|(\Omega_{S,S}^{(1)}\boldsymbol{\mu}_S^{(1)} - \Omega_{S,S}^{(0)}\boldsymbol{\mu}_S^{(0)})_j\right| > 0, \gamma_q = \inf_i \sup_j \left|(\Omega_{S_q^*,S_q^*}^{(1)} - \Omega_{S_q^*,S_q^*}^{(0)})_{ij}\right| > 0$, *then*

$$\min\{\gamma_l^2, \gamma_q^2, \gamma_q\} \gg D^2\sqrt{\frac{\log p}{n}} = o(1).$$

Remark 19 *The conditions here are similar to Assumption 4 for the LDA model. A set of analogous conditions were used in Jiang et al. (2018).*

Theorem 20 (QDA consistency) *For the QDA model, under Assumption 5,*

(i) If $D^2 c_n / \min\{\gamma_l^2, \gamma_q^2, \gamma_q\} = o(1)$, *then RIC is screening consistent:*

$$\mathbb{P}\left(\sup_{\substack{S:S\supseteq S^* \\ |S|\leq D}} \mathrm{RIC}_n(S) < \inf_{\substack{S:S\not\supseteq S^* \\ |S|\leq D}} \mathrm{RIC}_n(S)\right) \geq 1 - O\left(p^2 \exp\left\{-Cn\left(\frac{\min\{\gamma_l^2, \gamma_q^2, \gamma_q\}}{D^2}\right)^2\right\}\right)$$

$$\to 1.$$

(ii) Further, if $c_n \gg D^2\sqrt{\frac{\log p}{n}}$, *then RIC is weakly consistent:*

$$\mathbb{P}\left(\mathrm{RIC}_n(S^*) = \inf_{S:|S|\leq D} \mathrm{RIC}_n(S)\right) \geq 1 - O\left(p^2 \exp\left\{-Cn\left(\frac{c_n}{D^2}\right)^2\right\}\right) \to 1.$$

The proof is available in Appendix B. Note that the bound here is tighter than the results from Proposition 14 and Theorem 15.

Based on the consistency of RIC, in the next section, we will construct an upper bound for the expectation of the misclassification rate $R(C_n^{RaSE})$.

3.3 Misclassification Rate of the RaSE Classifier

In the following theorem, we present an upper bound on the misclassification rate for the RaSE classifier, which holds for any criterion to choose optimal subspaces.

Theorem 21 (General misclassification rate) *For the RaSE classifier with threshold* α *and any criterion to choose optimal subspaces, it holds that*

$$\mathbb{E}\{\mathbf{E}[R(C_n^{RaSE}) - R(C_{Bayes})]\} \leq \frac{\mathbb{E}\sup_{\substack{S:S\supseteq S^* \\ |S|\leq D}} [R(C_n^S) - R(C_{Bayes})] + \mathrm{P}(S_{1*} \not\supseteq S^*)}{\min(\alpha, 1-\alpha)}. \quad (8)$$

Here, the upper bound consists of two terms. The first term involving $\mathbb{E}\sup_{\substack{S:S\supseteq S^* \\ |S|\le D}}[R(C_n^{S^*})-R(C_{Bayes})]$ can be seen as the maximum discrepancy between the risk of models trained in any subspace covering S^* based on finite samples with the Bayes risk, which will be investigated in detail in the next subsection. This term shrinks to zero under certain conditions (see details in Section 3.4). The second term corresponds to the event that at least one signal is missed in the selected subspace. Corollary 16 in Section 3.2 shows that B_2 needs to be sufficiently large to ensure this term goes to 0. We will show in Section 3.5 that the iterative RaSE could relax the requirement on B_2 under certain scenarios.

Specifically, if we use the criterion of minimizing training error (misclassification rate on the training set) or leave-one-out cross-validation error, a similar guarantee of performance can be arrived as follows.

Theorem 22 (Misclassification rate when minimizing training error or leave-one-out cross-validation error) *If the criterion of minimal training error or leave-one-out cross-validation error is applied for the RaSE classifier with threshold α, it holds that*

$$\begin{aligned}&\mathbb{E}\{\mathbf{E}[R(C_n^{RaSE})-R(C_{Bayes})]\}\\ &\le \frac{\mathbb{E}\sup_{\substack{S:S\supseteq S^* \\ |S|\le D}}[R(C_n^S)-R(C_{Bayes})]+\left[\mathbb{E}(\epsilon_n)+\mathbb{E}\sup_{\substack{S:S\supseteq S^* \\ |S|\le D}}|\epsilon_n^S|\right]+(1-p_{S^*})^{B_2}}{\min(\alpha,1-\alpha)},\end{aligned} \tag{9}$$

where $\epsilon_n^S = R(C_n^S)-R_n(C_n^S), \epsilon_n = \mathbf{E}[R(C_n^{S_{1*}})-R_n(C_n^{S_{1*}})]$. Here $R_n(C)$ is the training error or leave-one-out cross-validation error of classifier C.

This theorem is closely related to Theorem 3 in Cannings and Samworth (2017) and derived along similar lines. The merit of Theorem 22 compared with Theorem 21 is that we don't have the term $\mathrm{P}(S_{1*} \not\supseteq S^*)$ in the bound, which can be difficult to quantify when minimizing training error or leave-one-out cross-validation error. Regarding the upper bound in (9), the first term is the same as the first term of the bound in Theorem 21. The second term involving $\mathbb{E}(\epsilon_n)+\mathbb{E}\sup_{\substack{S:S\supseteq S^* \\ |S|\le D}}|\epsilon_n^S|$ is relative to the distance between the training error and test error, which usually shrinks to zero for some specific classifiers under certain conditions (see details in Section 3.4). The third term involving $(1-p_{S^*})^{B_2}$ reflects the possibility that S^* is not selected in any of the B_2 subspaces we generate, which is similar to the second term in the bound given by Theorem 21. This term shrinks to zero under the condition of Corollary 16.

3.4 Detailed Analysis for Several Base Classifiers

In this section, we work out the technical details for the RaSE classifier when the base classifier is chosen to be LDA, QDA, and kNN.

3.4.1 Linear Discriminant Analysis (LDA)

LDA was proposed by Fisher (1936) and corresponds to model (1) where $f^{(r)} \sim N(\boldsymbol{\mu}^{(r)}, \Sigma), r = 0, 1$. For given training data $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ in subspace S, using the MLEs given in Proposition 8, the classifier can be constructed as

$$C_n^{S-LDA}(\boldsymbol{x}) = \begin{cases} 1, & L_S(\boldsymbol{x}_S|\hat{\pi}_0, \hat{\pi}_1, \hat{\boldsymbol{\mu}}_S^{(0)}, \hat{\boldsymbol{\mu}}_S^{(1)}, \hat{\Sigma}_{S,S}) > 0, \\ 0, & \text{otherwise}, \end{cases}$$

where the decision function

$$L_S(\boldsymbol{x}_S|\hat{\pi}_0, \hat{\pi}_1, \hat{\boldsymbol{\mu}}_S^{(0)}, \hat{\boldsymbol{\mu}}_S^{(1)}, \hat{\Sigma}_{S,S}) = \log(\hat{\pi}_1/\hat{\pi}_0) + (\boldsymbol{x}_S - (\hat{\boldsymbol{\mu}}_S^{(0)} + \hat{\boldsymbol{\mu}}_S^{(1)})/2)^T (\hat{\Sigma}_{S,S})^{-1} (\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)}).$$

And the degree of freedom of the LDA model with feature subspace S is $\deg(S) = |S| + 1$. Efron (1975) derived that

$$\begin{aligned} R(C_n^{S-LDA}) - R(C_{Bayes}) = \pi_1 &\left[\Phi\left(-\frac{\hat{\Delta}_S}{2} + \hat{\tau}_S \right) - \Phi\left(-\frac{\Delta_S}{2} + \tau_S \right) \right] \\ &+ \pi_0 \left[\Phi\left(-\frac{\hat{\Delta}_S}{2} - \hat{\tau}_S \right) - \Phi\left(-\frac{\Delta_S}{2} - \tau_S \right) \right], \end{aligned} \tag{10}$$

where $\Delta_S = \sqrt{\left(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}\right)^T \Sigma_{S,S}^{-1} \left(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}\right)}$, $\hat{\Delta}_S = \sqrt{\left(\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)}\right)^T \hat{\Sigma}_{S,S}^{-1} \left(\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)}\right)}$, $\tau_S = \log(\pi_1/\pi_0)/\Delta_S, \hat{\tau}_S = \log(\hat{\pi}_1/\hat{\pi}_0)/\hat{\Delta}_S$.

Proposition 23 *If Assumption 4 holds, then we have*

$$\mathbb{E} \sup_{\substack{S: S \supseteq S^* \\ |S| \le D}} [R(C_n^{S-LDA}) - R(C_{Bayes})] \lesssim D^2 \sqrt{\frac{\log p}{n}} \cdot \max\{(p^*)^{-\frac{1}{2}} \gamma^{-1}, (p^*)^{-\frac{3}{2}} \gamma^{-3}\}.$$

Regarding the second term in the upper bound in Theorem 22, due to Theorem 23.1 in Devroye et al. (2013), for any subset S, we have

$$\mathbb{P}\left(|\epsilon_n^S| > \epsilon\right) \le 8n^D \exp\left\{-\frac{1}{32} n\epsilon^2\right\},$$

which yields

$$\mathbb{P}\left(\sup_{\substack{S: S \supseteq S^* \\ |S| \le D}} |\epsilon_n^S| > \epsilon \right) \le \sum_{\substack{S: S \supseteq S^* \\ |S| \le D}} \mathbb{P}\left(|\epsilon_n^S| > \epsilon\right) \le 8n^D p^{D-p^*} \exp\left\{-\frac{1}{32} n\epsilon^2\right\}.$$

By Lemma 39 in Appendix B, it follows

$$\mathbb{E} \sup_{\substack{S: S \supseteq S^* \\ |S| \le D}} |\epsilon_n^S| \le \sqrt{\frac{32[(D - p^*)\log p + D \log n + 3\log 2 + 1]}{n}}. \tag{11}$$

Also since

$$\mathbb{E}|\epsilon_n| \leq \mathbf{E}\left[\mathbb{E}\left(\sup_{k=1,\ldots,B_2}|\epsilon_n^{S_{1k}}|\right)\right],$$

and

$$\mathbb{P}\left(\sup_{k=1,\ldots,B_2}|\epsilon_n^{S}|>\epsilon\right) \leq \sum_{k=1}^{B_2}\mathbb{P}\left(|\epsilon_n^{S_{1k}}|>\epsilon\right) \leq 8B_2n^D\exp\left\{-\frac{1}{32}n\epsilon^2\right\},$$

again by Lemma 39, we have

$$\mathbb{E}|\epsilon_n| \leq \sqrt{\frac{32[\log B_2 + D\log n + 3\log 2 + 1]}{n}}. \tag{12}$$

By plugging these bounds in the right-hand side of (8) and (9), we can get the explicit upper bound of the misclassification rate for the LDA model.

3.4.2 Quadratic Discriminant Analysis (QDA)

QDA considers the model (1) analogous to LDA while $\boldsymbol{x}|y=r \sim N(\boldsymbol{\mu}^{(r)}, \Sigma^{(r)}), r=0,1$, where $\Sigma^{(0)}$ can be different from $\Sigma^{(1)}$. On the basis of training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ in subspace S, it admits the following form of classifier based on the MLEs given in Proposition 9:

$$C_n^{S-QDA}(\boldsymbol{x}) = \begin{cases} 1, & Q_S(\boldsymbol{x}_S|\hat{\pi}_0, \hat{\pi}_1, \hat{\boldsymbol{\mu}}_S^{(0)}, \hat{\boldsymbol{\mu}}_S^{(1)}, \hat{\Sigma}_{S,S}^{(0)}, \hat{\Sigma}_{S,S}^{(1)}) > 0, \\ 0, & \text{otherwise}, \end{cases}$$

where the decision function

$$\begin{aligned} Q_S(\boldsymbol{x}_S|\hat{\pi}_0, \hat{\pi}_1, \hat{\boldsymbol{\mu}}_S^{(0)}, \hat{\boldsymbol{\mu}}_S^{(1)}, \hat{\Sigma}_{S,S}^{(0)}, \hat{\Sigma}_{S,S}^{(1)}) = \log(\hat{\pi}_1/\hat{\pi}_0) &- \frac{1}{2}\boldsymbol{x}_S^T[(\hat{\Sigma}_{S,S}^{(1)})^{-1} - (\hat{\Sigma}_{S,S}^{(0)})^{-1}]\boldsymbol{x}_S \\ &+ \boldsymbol{x}_S^T[(\hat{\Sigma}_{S,S}^{(1)})^{-1}\hat{\boldsymbol{\mu}}_S^{(1)} - (\hat{\Sigma}_{S,S}^{(0)})^{-1}\hat{\boldsymbol{\mu}}_S^{(0)}] \\ &- \frac{1}{2}(\hat{\boldsymbol{\mu}}_S^{(1)})^T(\hat{\Sigma}_{S,S}^{(1)})^{-1}\hat{\boldsymbol{\mu}}_S^{(1)} + \frac{1}{2}(\hat{\boldsymbol{\mu}}_S^{(0)})^T(\hat{\Sigma}_{S,S}^{(0)})^{-1}\hat{\boldsymbol{\mu}}_S^{(0)}. \end{aligned}$$

And the degree of freedom of QDA model with feature subspace S is $\deg(S) = |S|(|S|+1)/2 + |S| + 1$.

To analyze the first term in (8) and (9), as in Jiang et al. (2018), for any constant c, we define

$$u_c = \max\left\{\operatorname*{ess\,sup}_{z\in[-c,c]} h^{(r)}(z), r=0,1\right\},$$

where ess sup represents the essential supremum defined as the supremum except a set with measure zero and $h^{(r)}(z)$ is the density of $Q_{S^*}(\boldsymbol{x}_{S^*}|\pi_1, \pi_0, \boldsymbol{\mu}_{S^*}^{(0)}, \boldsymbol{\mu}_{S^*}^{(1)}, \Sigma_{S^*,S^*}^{(0)}, \Sigma_{S^*,S^*}^{(1)})$ given that $y=r$.

Proposition 24 *If Assumption 5 and the following conditions hold:*

(i) There exist positive constants c, U_c such that $u_c \leq U_c < \infty$;

(ii) There exists a positive number $\varpi_0 \in (0,1)$, $p \lesssim \exp\{n^{\varpi_0}\}$;

then we have

$$\mathbb{E} \sup_{\substack{S:S\supseteq S^* \\ |S|\leq D}} \left[R(C_n^{S-QDA}) - R(C_{Bayes})\right] \lesssim D^2 \left(\frac{\log p}{n}\right)^{\frac{1-2\varpi}{2}}.$$

for any $\varpi \in (0, 1/2)$.

Also, by applying Theorem 23.1 in Devroye et al. (2013) and Lemma 39, we have similar conclusions as (11) and (12) in the following

$$\mathbb{E} \sup_{\substack{S:S\supseteq S^* \\ |S|\leq D}} |\epsilon_n^S| \leq \sqrt{\frac{32[(D-p^*)\log p + D(D+3)/2 \cdot \log n + 3\log 2 + 1]}{n}},$$

$$\mathbb{E}|\epsilon_n| \leq \sqrt{\frac{32[\log B_2 + D(D+3)/2 \cdot \log n + 3\log 2 + 1]}{n}}.$$

The explicit upper bound of misclassification rate for the QDA model follows when we plug these inequalities into the right-hand sides of (8) and (9).

3.4.3 k-nearest Neighbor (kNN)

kNN method was firstly proposed by Fix (1951). Given $\boldsymbol{x}$, it tries to mimic the regression function $E[y|\boldsymbol{x}]$ in the local region around $\boldsymbol{x}$ by using the average of k nearest neighbors. kNN is a non-parametric method and its success has been witnessed in a wide range of applications.

Given training data $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, for the new observation $\boldsymbol{x}$ and subspace S, rank the training data by the increasing ℓ^2 distance in Euclidean space to $\boldsymbol{x}_S$ as $\{\boldsymbol{x}_{m_i,S}\}_{i=1}^n$ such that

$$\|\boldsymbol{x}_S - \boldsymbol{x}_{m_1,S}\|_2 \leq \|\boldsymbol{x}_S - \boldsymbol{x}_{m_2,S}\|_2 \leq \ldots \leq \|\boldsymbol{x}_S - \boldsymbol{x}_{m_n,S}\|_2,$$

where $\|\cdot\|_2$ represents the ℓ^2 norm in the corresponding Euclidean space. Then the kNN classifier admits the following form:

$$C_n^{S-k\text{NN}}(\boldsymbol{x}) = \begin{cases} 1, & \frac{1}{k}\sum_{i=1}^k y_{m_i} > 0.5, \\ 0, & \text{otherwise.} \end{cases}$$

By Devroye and Wagner (1979) and Cannings and Samworth (2017), it holds the following tail bound:

$$\mathbb{P}\left(\sup_{\substack{S:S\supseteq S^* \\ |S|\leq D}} |\epsilon_n^S| > \epsilon\right) \leq \sum_{\substack{S:S\supseteq S^* \\ |S|\leq D}} \mathbb{P}\left(|\epsilon_n^S| > \epsilon\right) \leq 8p^{D-p^*} \exp\left\{-\frac{n\epsilon^3}{108k(3^D+1)}\right\}.$$

Then by Lemma 39 and similar to the analysis for deriving (12), it follows

$$\mathbb{E} \sup_{\substack{S:S\supseteq S^* \\ |S|\leq D}} |\epsilon_n^S| \leq [108k(3^D+1)]^{\frac{1}{3}} \cdot \left(\frac{3\log 2 + (D-p^*)\log p + 1}{n}\right)^{\frac{1}{3}},$$

$$\mathbb{E}|\epsilon_n| \leq [108k(3^D+1)]^{\frac{1}{3}} \cdot \left(\frac{3\log 2 + \log B_2 + 1}{n}\right)^{\frac{1}{3}}.$$

However, for kNN, due to its lack of parametric form, it is much more involved to derive a similar upper bound as those presented in Propositions 23 and 24. We decide to leave this analysis as future work.

3.5 Theoretical Analysis of Iterative RaSE

Recall that to control the misclassification rate when minimizing RIC, we showed in Section 3.3 that to control $\mathrm{P}(S_{1*} \not\supseteq S^*)$, B_2 needs to be sufficiently large. In particular, a sufficient condition regarding B_2 was presented in Corollary 16. Since

$$\frac{\binom{p-p^*}{d-p^*}}{\binom{p}{d}} \leq \frac{\binom{p-p^*}{D-p^*}}{\binom{p}{D}} \leq \left(\frac{D}{p-p^*+1}\right)^{p^*},$$

condition (ii) in Corollary 16 implies $B_2 \gg \left(\frac{p-p^*+1}{D}\right)^{p^*}$, which could be very large for high-dimensional settings. Next, we show the iterative RaSE in Algorithm 2 can sometimes relax the requirement on B_2 substantially.

Different from the hierarchical uniform distribution over the subspaces in RaSE (Algorithm 1), the iterative RaSE in Algorithm 2 uses a non-uniform distribution from the second iteration. The non-uniform distribution works by assigning higher probabilities to subspaces that include the more frequently appeared variables among the B_1 subspaces chosen in the previous step. We will show that the subspaces generated from such non-uniform distributions require a smaller B_2 to cover S^*.

In the following analysis, we study the iterative RaSE algorithm that minimizes a general criterion Cr, which is a real-valued function on any subspace with its sample version denoted as Cr_n.

To guarantee the success of Algorithm 2, we need the following conditions.

Assumption 6 *Suppose the following conditions are satisfied:*

(i) There exists a positive function of n, p, D called ν satisfying $\nu(n,p,D) = o(1)$, such that

$$\sup_{S:|S|\leq D} |\mathrm{Cr}_n(S) - \mathrm{Cr}(S)| = O_p(\nu(n,p,D)).$$

(ii) (Stepwise detectable condition) There exists a series $\{M_n\}_{n=1}^{\infty} \to \infty$ and a specific positive integer $\bar{p}^* \leq p^*$ such that

(a) for any feature subset $\tilde{S}_*^{(1)} \subseteq S^*$ where $|\tilde{S}_*^{(1)}| \leq p^* - \bar{p}^*$, there exists $S_*^{(2)} \subseteq S^* \backslash \tilde{S}_*^{(1)}$ and $|S_*^{(2)}| = \bar{p}^*$ such that

$$\mathrm{Cr}(S') - \mathrm{Cr}(S) > M_n \nu(n,p,D)$$

holds for any n, any S and S' satisfying $S \cap S^* \neq S' \cap S^*$, $S \cap S^* = \tilde{S}_*^{(1)} \cup S_*^{(2)}$, $S' \cap S^* = S_1' \cup S_2'$ where $S_1' \subseteq \tilde{S}_*^{(1)}, S_2' \subseteq S^* \backslash \tilde{S}_*^{(1)}, |S_2'| \leq \bar{p}^*$;

(b) the criterion satisfies

$$\inf_{\substack{S:|S|\le D\\ S\not\supseteq S^*}} \mathrm{Cr}(S) - \sup_{\substack{S:|S|\le D\\ S\supseteq S^*}} \mathrm{Cr}(S) > M_n\nu(n,p,D).$$

(iii) We have

$$D\log\log p \ll \log p.$$

In Assumption 6, condition (i) provides a uniform bound for the specific criterion we use. Condition (ii)(a) is introduced to make Algorithm 2 detect additional signals in each iteration until all signals are covered. Condition (ii)(b) is imposed to help us find discriminative sets among B_2 subspaces, and it holds for RIC by previous analysis in Section 3.2. Condition (iii) characterizes the requirement on the dimension p and the maximal subspace size D.

Theorem 25 *For Algorithm 2, B_2 in the first step is set as*

$$Dp^{\bar{p}^*} \lesssim B_2 \ll \left(\frac{p}{\bar{p}^* D}\right)^{\bar{p}^*+1},$$

and B_2 in the following steps is set as

$$\left(1+\frac{D}{C_0}\right)^D p^{\bar{p}^*}(\log p)^{p^*} \lesssim B_2 \ll \left(\frac{p}{\bar{p}^* D}\right)^{\bar{p}^*+1},$$

where C_0 is the same as in Algorithm 2. Also we set B_1 such that $B_1 \gg \log p^*$. If Assumption 6 holds, then after T iterations where $\frac{e^{B_1}}{p^*} \gg T \ge \lceil \frac{p^*}{\bar{p}^*} \rceil$, as $n, B_1, B_2 \to \infty$ there holds

$$\mathrm{P}(S_{1*}^{(T)} \not\supseteq S^*) \to 0.$$

Next, we compare the requirements on B_2 for iterative RaSE with that for the vanilla RaSE. For simplicity, we assume D, p^*, and $\bar{p}^*$ are constants. When $\bar{p}^* < p^*$, iterative RaSE requires $B_2 \gtrsim p^{\bar{p}^*}(\log p)^{p^*}$, which is much weaker than the requirement $B_2 \gg p^{p^*}$ for vanilla RaSE implied by Corollary 16.

Using the results in Theorem 21, an upper bound for the error rate could be obtained. Also note that the rate of p is constrained by Assumption 6.(i), where we assume $\nu(n,p,D) = o(1)$. For example, with LDA and QDA model, by Lemmas 32 and 38 in Appendix B, under Assumptions 4 and 5 respectively, we have

$$\nu(n,p,D) = D^2\sqrt{\frac{\log p}{n}}.$$

Therefore the constraint is $D^2\sqrt{\frac{\log p}{n}} = o(1)$.

4. Computational and Practical Issues

4.1 Tuning Parameter Selection

In Algorithm 1, there are five tuning parameters, including the number of weak learners B_1, the number of candidate subspaces B_2 to explore for each weak learner, the distribution $\mathcal{D}$ of subspaces, the criterion $\mathcal{C}$ for selecting the optimal subspace for each weak learner, and the threshold $\hat{\alpha}$.

If we set $\mathcal{C}$ as minimizing the RIC, the difference between the risk of RaSE and Bayes risk, as well as the MC variance of RaSE, vanish at an exponential rate when $B_1 \to \infty$, except for a finite set of thresholds α. This implies the RaSE classifier becomes more accurate and stable as B_1 increases. Regarding the impact of B_2, by Corollary 16, Theorem 15 and Theorem 25, under the minimal RIC criterion with some conditions, as $B_2, n \to \infty$, the subspace chosen for each weaker learner recovers the minimal determinative set with high probability. By Theorem 21, the expectation of the misclassification rate becomes closer to the Bayes error as the sample size n and B_2 increase, which motivates us to use a large B_2 if we have sufficient computational power. However, when we choose "minimizing training error" as the criterion $\mathcal{C}$ to select the optimal subspace, Theorem 22 shows that the influence of B_2 becomes more subtle. In our implementation, we set $B_1 = 200$ and $B_2 = 500$ as default. For LDA and QDA classifier, $\mathcal{C}$ is set to choose the optimal subspace by minimizing the RIC, while for kNN, the default setting is minimizing the leave-one-out cross-validation error.

Without prior information about the features, as we mentioned in Section 2.1, $\mathcal{D}$ is set as the hierarchical uniform distribution over the subspaces. To generate the size d of subspaces from the uniform distribution over $\{1, \ldots, D\}$, another parameter D has to be determined. In practice, for QDA base classifier we set $D = \min(p, [\sqrt{n_0}], [\sqrt{n_1}])$ and for LDA, kNN and all the other base classifiers, we set $D = \min(p, [\sqrt{n}])$, where $[a]$ denotes the largest integer not larger than a. The threshold $\hat{\alpha}$ is chosen by (2) to minimize the training error. When using non-parametric estimate of RIC corresponding to (6), following Wang et al. (2009) and Ganguly et al. (2018), we set $k_0 = [\sqrt{n_0}]$ and $k_1 = [\sqrt{n_1}]$ to satisfy the conditions they presented for proving the consistency.

4.2 Computational Cost

RaSE is an ensemble framework, generating $B_1 B_2$ subspaces in total following distribution $\mathcal{D}$. If we use the uniform distribution introduced in the last section to generate one subspace, the time required equals to the time for sampling at most D features from p ones, which is $O(pD)$. And the time for training the base model is denoted as T_{train}, which equals to $O(nD^2)$ for LDA and QDA base classifiers. Similarly, the time for predicting test data is denoted as T_{test}, which equals to $O(n_{\text{test}}D)$ for LDA base classifier, $O(n_{\text{test}}D^2)$ for QDA base classifier, and $O(n \cdot n_{\text{test}}D)$ for k-NN base classifier. In total, the computation cost for the training process is $O(B_1 B_2 T_{\text{train}} + B_1 B_2 \log B_2)$ time. Here, $O(B_2 \log B_2)$ is the time needed to find the optimal subspace among B_2 ones based on the sorting of their scores calculated under some criterion. The computation cost for prediction process is $O(B_1 T_{\text{test}})$ and RaSE algorithm takes approximately $O(B_1 B_2 T_{\text{train}} + B_1 B_2 \log B_2 + B_1 T_{\text{test}})$ for both model fitting and prediction.

In practice, this type of framework is very convenient to apply the parallel computing paradigm, making the computing quite fast. And for specific classifiers like LDA and QDA, we have simplified the RIC expression, which can be directly used to speed up calculation. Compared to the projection generation process in Cannings and Samworth (2017), RaSE is more efficient since we only need to select features based on certain distribution without doing any complicated matrix operations.

4.3 Feature Ranking

There are many powerful tools in statistics and machine learning for variable selection and feature ranking. For the sparse classification approaches like sparse-LDA and sparse-QDA (Mai et al., 2012; Jiang et al., 2018; Fan et al., 2012; Shao et al., 2011; Hao et al., 2018; Fan et al., 2015; Mai et al., 2015; Fan et al., 2016), or independent regularization approach like nearest shrunken centroids (Tibshirani et al., 2003), this is usually directly implied by the methodology. For model-free classification methods, however, it's not straightforward to rank features. For random forest, Breiman (2001) proposed a feature ranking method by randomly permuting the value of each feature and calculating the misclassification rate on the out-of-bag data for the new random forest.

For RaSE, as an ensemble framework, it's quite natural to rank variables by their frequencies of appearing in B_1 subspaces corresponding to B_1 weak learners. Following Corollary 16, as $n, B_2 \to \infty$, under some conditions for signal strength and the increasing rate of B_2, by applying the criterion of minimizing RIC, the chosen subspace tends to cover the minimal discriminative set S^* with high probability, which intuitively illustrates why this idea works to rank variables. When we do not have sufficient computational resources to set a very large B_2, as Theorem 25 indicates, under some conditions, the iterative RaSE (Algorithm 2) can cover S^* with high probability after a few steps with a smaller B_2. In practice, the frequencies of signals in S^* tend to increase after iterations, which can improve the performance of the RaSE classifier and provide a better ranking. We will demonstrate this via extensive simulation studies in the next section.

5. Simulations and Real-data Experiments

We use six simulation settings and four real data sets to demonstrate the effectiveness of the RaSE method, coupled with RIC and leave-one-out cross-validation as the minimizing criterion to choose the optimal subspace. The performance of RaSE classifiers with LDA, QDA, Gamma and kNN as base classifiers with different iteration numbers are compared with that of other competitors, including the standard LDA, QDA, and kNN classifiers, sparse LDA (sLDA) (Mai et al., 2012), regularization algorithm under marginality principle (RAMP) (Hao et al., 2018), nearest shrunken centroids (NSC) (Tibshirani et al., 2003), random forests (RF) (Breiman, 2001), and random projection ensemble classifier (RPEnsemble) (Cannings and Samworth, 2017). For the LDA model, we also implemented the non-parametric estimate for RIC to show its effectiveness and robustness.

The standard LDA and QDA methods are implemented by using R package `MASS`. And the kNN classifier is implemented by `knn`, `knn.cv` in R package `class` and function `knn3` in R package `caret`. We utilize package `dsda` to fit sLDA model. RAMP is implemented through package `RAMP`. For the RF, we use R package `RandomForest`; the number of trees

are set as 500 (as default) and $[\sqrt{p}]$ variables are randomly selected when training each tree (as default). And the NSC model is fitted by calling function `pamr.train` in package `pamr`. RPEnsemble is implemented by R package `RPEnsemble`. To obtain the MLE of parameters in the Gamma distribution, the Newton's method is applied via function `nlm` in package `stats` and the initial point is chosen to be the moment estimator. To get the non-parametric estimate of KL divergence and RIC, we call function `KL.divergence` in package `FNN`.

When fitting the standard kNN classifier, and the kNN base classifier in RPEnsemble and RaSE method, the number of neighbors k is chosen from $\{3, 5, 7, 9, 11\}$ via leave-one-out cross-validation, following Cannings and Samworth (2017). For RAMP, the response type is set as "binomial", for which the logistic regression with interaction effects is considered. In sLDA, the penalty parameter λ is chosen to minimize cross-validation error. In RPEnsemble method, LDA, QDA, kNN, are set as base classifiers with default parameters, and the number of weak learner $B_1 = 500$ and the number of projection candidates for each weak learner $B_2 = 50$ and the dimension of projected space $d = 5$. The projection generation method is set to be "Haar". The criterion of choosing optimal projection is set to minimize training error for LDA and QDA and minimize leave-one-out cross-validation error for kNN. For the RaSE method, for LDA, QDA, and independent Gamma classifier (to be illustrated later in Section 5.1.2), the criterion is set to be minimizing RIC, and for k-NN, the strategy of minimizing leave-one-out cross-validation error is applied. Other parameter settings in RaSE are the same as in the last section. For simulations, the number of iterations T in Algorithm 2 is set to be 0, 1, 2, while for real-data experiments, we only consider RaSE methods with 0 or 1 iteration. We write the iteration number on the subscript, and if it is zero, the subscript will be omitted. For example, RaSE_1-LDA represents the RaSE classifier with $T = 1$ iteration and LDA base classifier. In addition, we use LDAn to denote the LDA base classifier with the non-parametric estimate of RIC.

For all experiments, 200 replicates are considered, and the average test errors (percentage) are calculated based on them. The standard deviation of the test errors over 200 replicates is also calculated for each approach and written on the subscript. The approach with minimal test error for each setting is highlighted in bold, and methods that achieve test error within one standard deviation of the best one are highlighted in italics. Also, for all simulations and the madelon data set, the average selected percentage of features in B_1 subspaces for the largest sample size setting in the RaSE method in 200 replicates are presented, which provides a natural way for feature ranking. For the average selected percentage in the case of the smallest sample size, refer to Appendix A. To highlight the different behaviors of signals and noises, we present the average selected percentages of all noise features as a box-plot marked with "N".

The RaSE classifier competes favorably with existing classification methods. Its misclassification rate is the lowest in 27 out of 30 (simulation and real-data) settings and within one standard deviation of the lowest in the remaining four settings.

5.1 Simulations

For the simulated data, model 1 follows model 1 in Mai et al. (2012), which is a sparse LDA-adapted setting. In model 2, for each class, a Gamma distribution with independent components is used. Model 3 follows from the setting of model 3 in Fan et al. (2015), which is

a QDA-adapted model. The marginal distribution for two classes is set to be $\pi_0 = \pi_1 = 0.5$ for the first three simulation models. Model 4 is motivated by the kNN algorithm, and the data generation process will be introduced below. To test the robustness of RaSE, two non-sparse settings, model 1' and 4' are investigated as well, where model 1' has decaying signals in the LDA model while model 4' inherits from model 4 by increasing the number of signals to 30 with the signal strength decreased.

For simulations, we consider the "signal" model as a benchmark. These models use the correct model on the minimal discriminative set S^*, mimicking the behavior of the Bayes classifier when S^* is sparse.

5.1.1 Models 1 and 1' (LDA)

First we consider a sparse LDA setting (model 1). Let $\boldsymbol{x}|y = r \sim N(\boldsymbol{\mu}^{(r)}, \Sigma), r = 0, 1$, where $\Sigma = (\Sigma_{ij})_{p\times p} = (0.5^{|i-j|})_{p\times p}, \boldsymbol{\mu}^{(0)} = \mathbf{0}_{p\times 1}, \boldsymbol{\mu}^{(1)} = \Sigma \times 0.556(3, 1.5, 0, 0, 2, \mathbf{0}_{1\times(p-5)})^T$. Here $p = 400$, and the training sample size $n \in \{200, 400, 1000\}$. Test data of size 1000 is independently generated from the same model.

As analyzed in Example 1, feature subset $\{1, 2, 5\}$ is the minimal discriminative set S^*. On the left panel of Table 1, the performance of various methods on model 1 for different sample sizes are presented. As we could see, RaSE$_1$-LDAn performs the best when the sample size $n = 200$ and 400. sLDA achieves similar performances to the best classifiers for each setting, and RaSE$_2$-QDA ranks the top when $n = 1000$. Also, since this model is very sparse, the default value of B_2 cannot guarantee that the minimal discriminative set can be selected. Therefore the iterative version of RaSE improves the performance of RaSE a lot. And NSC also achieves a comparably small misclassification rate when $n = 200$.

In Figure 1, the average selected percentages of 400 features in 200 replicates when $n = 1000$ are presented. Note that after two iterations, the three signals can be captured by almost all $B_1 = 200$ subspaces for all three base classifiers, and all noises are rarely selected across B_1 subspaces except when the non-parametric estimate of RIC is applied.

Next, we consider a non-sparse LDA model (model 1'). Let $\boldsymbol{\mu}^{(1)} = \Sigma\cdot(0.9, 0.9^2, \ldots, 0.9^{50}, \mathbf{0}_{1\times(p-50)})^T$ and keep other parameters the same as above. Now S^* contains the first 50 features. Under this non-sparse setting, as the right panel of Table 1 shows, although most methods obtain similar error rates, RaSE$_1$-LDAn achieves the best performance when $n = 200$ and 400. RaSE$_1$-kNN performs the best when $n = 1000$. From the table, it can be seen that despite the non-sparse design, the iterations can still improve the performance of RaSE.

An interesting phenomenon is observed from Figure 2, which exhibits the average selected percentages for model 1'. Note that the selected percentages are decaying as the signal strength decreases except for the first feature. One possible reason is that the marginal discriminative powers of feature 2 and 3 are the strongest among all features due to the specific correlation structure in our setting.

5.1.2 Model 2 (Gamma distribution)

In this model we investigate the Gamma distribution with independent features, which is rarely studied in the literature. $\boldsymbol{x}|y = r$ at j-th coordinate follows Gamma distribution

Method	Results for model 1			Results for model 1'		
	$n = 200$	$n = 400$	$n = 1000$	$n = 200$	$n = 400$	$n = 1000$
RaSE-LDA	$12.99_{1.42}$	$12.38_{1.20}$	$\mathit{11.16}_{1.10}$	$21.43_{4.64}$	$\mathit{19.56}_{3.67}$	$\mathit{17.98}_{2.74}$
RaSE-LDAn	$13.40_{1.31}$	$13.23_{1.20}$	$12.64_{1.11}$	$21.64_{4.22}$	$20.25_{3.01}$	$19.42_{2.59}$
RaSE-QDA	$13.78_{1.26}$	$13.69_{1.19}$	$13.23_{1.04}$	$22.36_{3.96}$	$20.46_{2.83}$	$19.83_{2.51}$
RaSE-kNN	$13.21_{1.48}$	$12.57_{1.23}$	$11.19_{1.10}$	$\mathit{19.98}_{3.74}$	$\mathit{18.21}_{2.89}$	$\mathit{16.97}_{2.10}$
RaSE$_1$-LDA	$\mathit{11.48}_{1.26}$	$\mathit{10.42}_{1.07}$	$\mathit{10.25}_{1.08}$	$\mathit{20.15}_{3.65}$	$\mathit{18.39}_{2.69}$	$\mathit{17.20}_{2.14}$
RaSE$_1$-LDAn	$\mathbf{10.58}_{1.15}$	$\mathbf{10.28}_{1.01}$	$\mathit{10.16}_{1.04}$	$\mathbf{18.21}_{2.92}$	$\mathbf{17.55}_{2.26}$	$\mathit{16.86}_{1.74}$
RaSE$_1$-QDA	$\mathit{11.28}_{1.47}$	$\mathit{10.76}_{1.25}$	$\mathit{10.40}_{1.22}$	$21.24_{4.75}$	$\mathit{19.65}_{3.56}$	$\mathit{18.23}_{2.51}$
RaSE$_1$-kNN	$\mathit{11.11}_{1.25}$	$\mathit{10.58}_{1.09}$	$\mathit{10.33}_{1.05}$	$\mathit{18.90}_{3.16}$	$\mathit{17.75}_{2.51}$	$\mathbf{16.80}_{1.98}$
RaSE$_2$-LDA	$12.62_{1.45}$	$11.41_{1.16}$	$\mathit{10.13}_{1.03}$	$21.78_{3.44}$	$\mathit{19.22}_{2.54}$	$\mathit{17.53}_{1.98}$
RaSE$_2$-LDAn	$\mathit{10.90}_{1.16}$	$\mathit{10.42}_{0.96}$	$\mathit{10.17}_{1.06}$	$\mathit{18.75}_{2.87}$	$\mathit{17.95}_{2.19}$	$\mathit{17.30}_{1.70}$
RaSE$_2$-QDA	$11.74_{1.45}$	$\mathit{10.39}_{1.01}$	$\mathbf{10.09}_{1.07}$	$21.84_{5.03}$	$20.00_{3.62}$	$19.04_{2.57}$
RaSE$_2$-kNN	$\mathit{11.17}_{1.31}$	$\mathit{10.60}_{0.99}$	$\mathit{10.34}_{1.02}$	$\mathit{18.96}_{3.10}$	$\mathit{17.77}_{2.29}$	$\mathit{17.19}_{1.78}$
RP-LDA	$17.26_{1.53}$	$15.03_{1.24}$	$13.37_{1.09}$	$25.66_{1.98}$	$23.85_{1.79}$	$21.99_{1.50}$
RP-QDA	$17.96_{1.60}$	$15.26_{1.34}$	$13.50_{1.12}$	$26.41_{2.04}$	$24.03_{1.88}$	$22.06_{1.51}$
RP-kNN	$18.54_{1.60}$	$16.15_{1.21}$	$14.23_{1.21}$	$27.03_{2.27}$	$25.06_{1.76}$	$23.01_{1.55}$
LDA	—†	$46.39_{2.62}$	$18.62_{1.53}$	—†	$47.81_{2.87}$	$27.44_{1.95}$
QDA	—†	—†	$47.74_{1.73}$	—†	—†	$48.90_{1.65}$
kNN	$29.08_{2.77}$	$26.73_{1.97}$	$24.53_{1.60}$	$35.67_{2.59}$	$34.07_{2.33}$	$32.36_{1.72}$
sLDA	$\mathit{10.80}_{1.26}$	$\mathit{10.48}_{1.16}$	$\mathit{10.23}_{1.07}$	$\mathit{18.80}_{3.19}$	$\mathit{17.62}_{2.28}$	$\mathit{17.24}_{1.81}$
RAMP	$14.09_{2.67}$	$\mathit{10.56}_{1.33}$	$\mathit{10.10}_{1.04}$	$21.91_{5.49}$	$\mathit{19.22}_{3.25}$	$\mathit{17.55}_{2.04}$
NSC	$\mathit{11.50}_{1.13}$	$11.50_{0.97}$	$11.67_{1.09}$	$\mathit{18.49}_{1.99}$	$\mathit{19.02}_{1.78}$	$19.41_{1.39}$
RF	$12.66_{1.43}$	$11.77_{1.04}$	$11.25_{1.10}$	$21.33_{3.01}$	$20.00_{2.10}$	$19.25_{1.51}$
Sig-LDA	$10.07_{0.94}$	$10.09_{0.95}$	$10.07_{1.03}$	$23.70_{3.14}$	$20.90_{2.27}$	$18.95_{1.64}$

† Not applicable.

Table 1: Error rates for models 1 and 1'

$Gamma(\alpha_j^{(r)}, \beta_j^{(r)})$, which has the density function

$$f_j^{(r)}(x; \alpha_j^{(r)}, \beta_j^{(r)}) = \frac{1}{(\beta_j^{(r)})^{\alpha_j^{(r)}} \Gamma(\alpha_j^{(r)})} x^{\alpha_j^{(r)}-1} \exp\{-x/\beta_j^{(r)}\} \mathbb{1}(x \geq 0), j = 1, \ldots, p, r = 0, 1. \tag{13}$$

Denote $\boldsymbol{\alpha}^{(r)} = (\alpha_1^{(r)}, \ldots, \alpha_p^{(r)})^T, \boldsymbol{\beta}^{(r)} = (\beta_1^{(r)}, \ldots, \beta_p^{(r)})^T$. Here, we let $\boldsymbol{\alpha}^{(0)} = (2, 1.5, 1.5, 2, 2, \mathbf{1}_{1\times(p-5)})^T$, $\boldsymbol{\alpha}^{(1)} = (2.5, 1.5, 1.5, 1, 1, \mathbf{1}_{1\times(p-5)})^T, \boldsymbol{\beta}^{(0)} = (1.5, 3, 1, 1, 1, \mathbf{1}_{1\times(p-5)})^T, \boldsymbol{\beta}^{(1)} = (2, 1, 3, 1, 1, \mathbf{1}_{1\times(p-5)})^T, p = 400, n \in \{100, 200, 400\}$. Hence, the minimal discriminative set S^* is $\{1, 2, 3, 4, 5\}$, due to Proposition 3.

MLEs of $\boldsymbol{\alpha}^{(0)}, \boldsymbol{\alpha}^{(1)}, \boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}$ can be obtained by numerical approaches like the gradient descent or Newton's method. And the marginal probabilities are estimated by the proportion of two classes in training samples. Then the Bayes classifier is estimated by these MLEs and applied to classify new observations, which is denoted as an independent Gamma classifier in Table 2. For this example, we also apply RaSE with the independent Gamma classifier as one of the base classifiers. According to (3) and (13), the decision function of independent Gamma classifier estimated in subspace S is

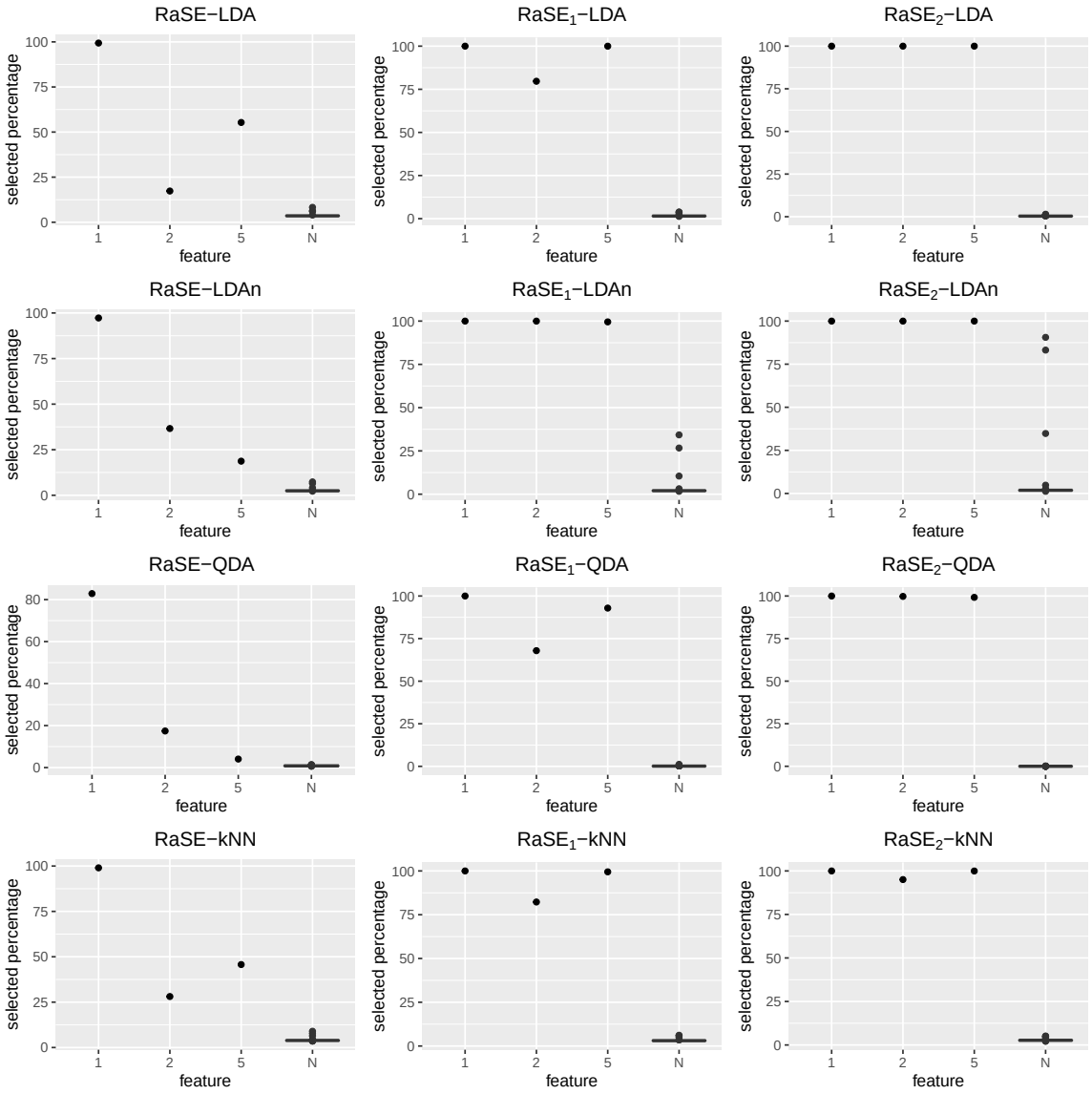

Figure 1: Average selected percentages of features for model 1 in 200 replicates when $n = 1000$

$$C_n^{S-Gamma}(\boldsymbol{x}) = \mathbb{1}\left(\frac{\hat{\pi}_1}{\hat{\pi}_0} \cdot \prod_{j \in S} \frac{f_j^{(1)}(x_j; \hat{\alpha}_j^{(1)}, \hat{\beta}_j^{(1)})}{f_j^{(0)}(x_j; \hat{\alpha}_j^{(0)}, \hat{\beta}_j^{(0)})} > 0.5\right),$$

where $\hat{\pi}_1, \hat{\pi}_0, \hat{\alpha}_j^{(1)}, \hat{\alpha}_j^{(0)}, \hat{\beta}_j^{(1)}, \hat{\beta}_j^{(0)}$ are corresponding MLEs.

This is also a very sparse model. Therefore the iteration process can improve the RaSE method with a lower misclassification rate. The left panel of Table 2 shows us the performance of various methods on this model. It demonstrates that RaSE$_1$-Gamma performs

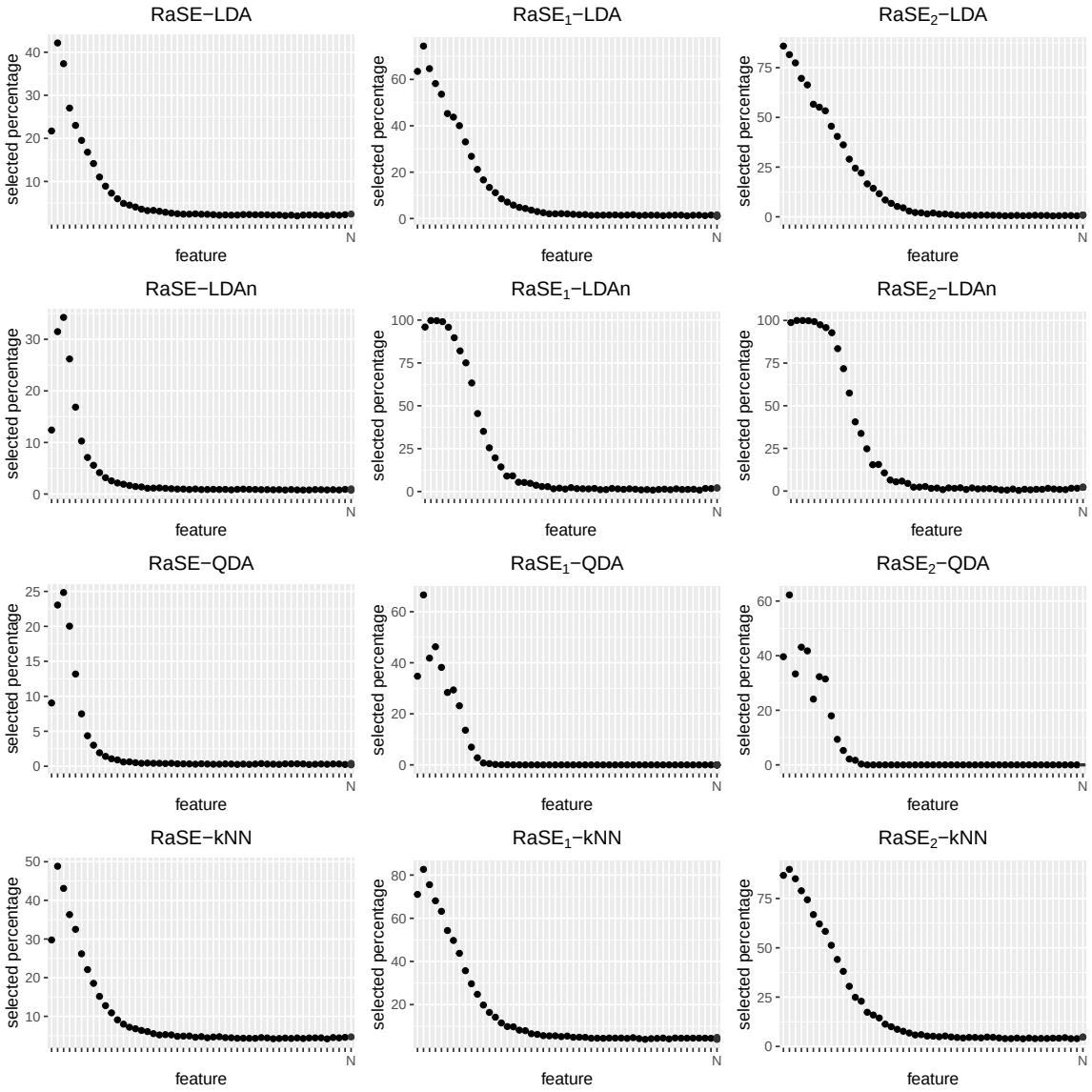

Figure 2: Average selected percentages of features for model 1' in 200 replicates when $n = 1000$

the best when $n = 100$. RaSE$_2$-Gamma incurs the lowest misclassification rate and low standard deviation for the other two settings.

Figure 3 shows us the average selected percentage of features when $n = 400$, from which we can see that due to the high sparsity, the default B_2 is not sufficient and makes it hard for the vanilla RaSE classifier to capture all the features in S^*. After iterations, the frequencies of the minimal discriminative set increase significantly, and S^* can be easily identified for all three base classifiers after two iterations.

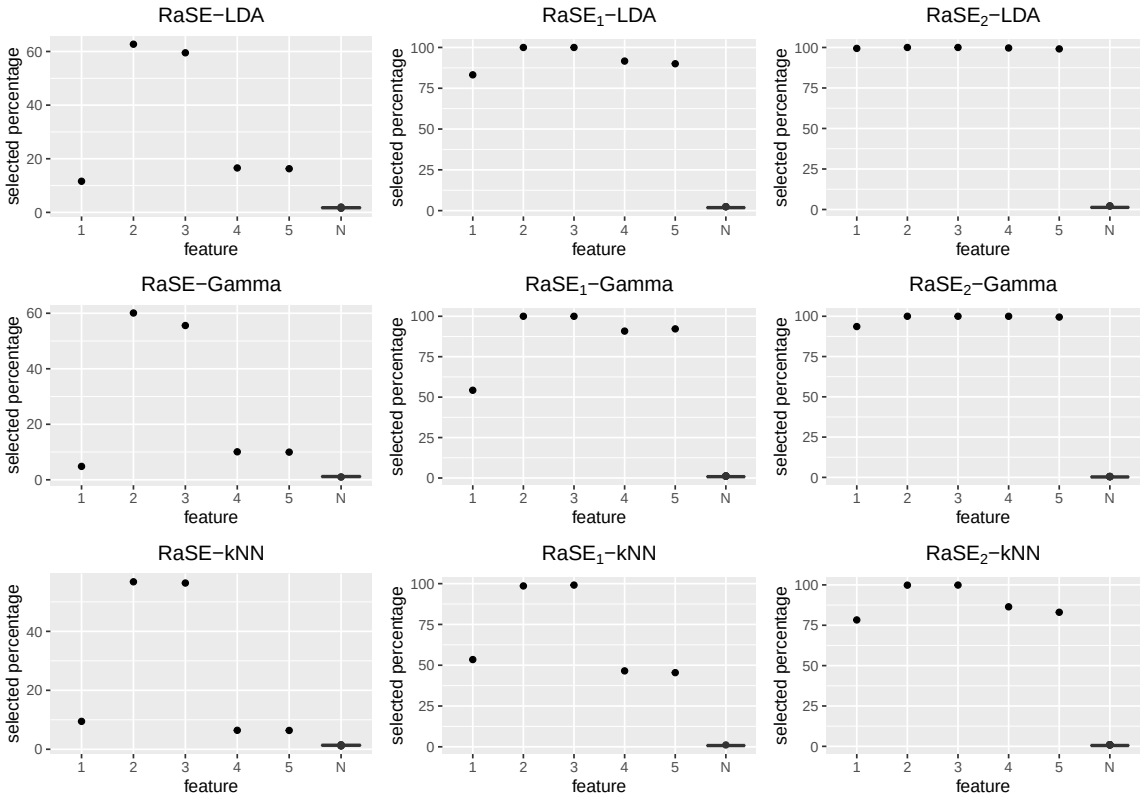

Figure 3: Average selected percentages of features for model 2 in 200 replicates when $n = 400$

5.1.3 Model 3 (QDA)

$\boldsymbol{x}|y = r \sim N(\boldsymbol{\mu}^{(r)}, \Sigma^{(r)}), r = 0, 1$, where $\Omega^{(0)} = (\Sigma^{(0)})^{-1}$ is a $p \times p$ band matrix with $(\Omega^{(0)})_{ii} = 1$ and $(\Omega^{(0)})_{ik} = 0.3$ for $|i-k| = 1$. All the other entries are zero. $\Omega^{(1)} = \Omega^{(0)} + \Omega$, where Ω is a $p \times p$ sparse symmetric matrix with $\Omega_{10,10} = -0.3758, \Omega_{10,30} = 0.0616, \Omega_{10,50} = 0.2037, \Omega_{30,30} = -0.5482, \Omega_{30,50} = 0.0286, \Omega_{50,50} = -0.4614$ and all the other entries are zero. Here $p = 200, n \in \{200, 400, 1000\}$.

As analyzed in Example 2, the minimal discriminative set $S^* = \{1, 2, 10, 30, 50\}$, where features 1 and 2 represent linear signals or main effects and features 10, 30, and 50 are quadratic signals. The right panel of Table 2 shows us the results. From it we can see that RaSE$_1$-QDA achieves the best performance when $n = 200$, and RaSE$_2$-QDA has the lowest test misclassification rate when $n = 400, 1000$, which is reasonable since data is generated from a QDA-adapted model. RaSE$_1$-kNN and RaSE$_2$-kNN also do a good job when n is large.

Figure 4 represents the average selected percentage of features when $n = 1000$, which exhibits that the frequencies of the elements in the minimal discriminative set are increasing after iterations. When the base classifier is QDA or kNN, all the five features stand out. On

the other hand, RaSE, with the LDA base classifier, only captures the two linear signals, which is expected since LDA does not consider the quadratic terms.

Method[3]	Results for model 2			Results for model 3		
	$n = 100$	$n = 200$	$n = 400$	$n = 200$	$n = 400$	$n = 1000$
RaSE-LDA	$21.51_{3.42}$	$18.52_{2.75}$	$17.43_{2.00}$	$37.30_{3.17}$	$36.11_{1.97}$	$35.67_{1.73}$
RaSE-Gamma/QDA	$23.57_{3.65}$	$21.41_{3.17}$	$20.29_{2.43}$	$32.52_{2.90}$	$30.44_{2.60}$	$29.00_{1.97}$
RaSE-kNN	$22.83_{3.09}$	$21.07_{3.07}$	$20.47_{2.57}$	$31.10_{3.23}$	$27.83_{2.41}$	$25.22_{1.56}$
RaSE$_1$-LDA	$19.01_{2.83}$	$15.14_{1.75}$	$13.55_{1.20}$	$36.09_{2.87}$	$32.82_{1.74}$	$32.68_{1.49}$
RaSE$_1$-Gamma/QDA	$\mathbf{15.05}_{2.31}$	$\textit{13.02}_{1.51}$	$\textit{12.50}_{1.18}$	$\mathbf{26.83}_{2.47}$	$\textit{25.07}_{1.89}$	$\textit{23.53}_{1.50}$
RaSE$_1$-kNN	$19.84_{2.90}$	$16.64_{1.86}$	$15.33_{1.39}$	$\textit{28.76}_{2.60}$	$\textit{25.88}_{1.98}$	$\textit{24.18}_{1.47}$
RaSE$_2$-LDA	$20.88_{3.03}$	$16.92_{2.05}$	$13.66_{1.14}$	$38.09_{2.48}$	$33.69_{1.83}$	$32.71_{1.55}$
RaSE$_2$-Gamma/QDA	$\textit{16.27}_{2.22}$	$\mathbf{12.64}_{1.31}$	$\mathbf{11.83}_{1.02}$	$\textit{26.99}_{2.68}$	$\mathbf{24.87}_{1.99}$	$\mathbf{23.11}_{1.60}$
RaSE$_2$-kNN	$22.39_{3.13}$	$17.40_{2.27}$	$14.59_{1.44}$	$\textit{28.73}_{2.56}$	$\textit{25.46}_{1.82}$	$\textit{23.76}_{1.54}$
RP-LDA	$38.89_{1.96}$	$35.00_{1.97}$	$30.89_{1.76}$	$44.90_{1.86}$	$42.82_{1.76}$	$40.38_{1.74}$
RP-QDA	$43.62_{3.62}$	$37.62_{2.35}$	$33.02_{2.14}$	$43.02_{2.07}$	$39.87_{1.88}$	$36.38_{1.78}$
RP-kNN	$41.48_{2.26}$	$38.90_{2.06}$	$36.69_{1.85}$	$44.32_{1.81}$	$42.46_{1.58}$	$40.80_{2.12}$
LDA	—†	—†	$47.50_{2.35}$	$49.03_{1.94}$	$42.88_{1.82}$	$38.68_{1.70}$
QDA	$32.06_{2.40}$	$26.22_{1.80}$	$21.56_{1.44}$	—†	—†	$45.13_{1.58}$
kNN	$45.48_{2.24}$	$44.68_{2.14}$	$44.07_{2.00}$	$45.67_{1.78}$	$44.63_{2.02}$	$43.43_{1.63}$
sLDA	$22.26_{3.52}$	$18.64_{2.12}$	$15.34_{1.55}$	$36.41_{3.15}$	$33.87_{2.01}$	$32.99_{1.52}$
RAMP	$20.64_{3.81}$	$16.72_{2.25}$	$13.31_{1.21}$	$36.94_{5.87}$	$32.65_{1.89}$	$32.42_{1.78}$
NSC	$26.00_{6.49}$	$19.92_{4.31}$	$16.87_{2.69}$	$41.14_{4.49}$	$38.24_{3.85}$	$35.13_{2.20}$
RF	$24.97_{5.74}$	$18.02_{2.77}$	$15.26_{1.47}$	$37.34_{2.91}$	$31.61_{2.19}$	$27.42_{1.60}$
Sig-Gamma/QDA	$12.65_{1.12}$	$12.12_{1.12}$	$11.76_{0.97}$	$23.62_{1.47}$	$22.72_{1.40}$	$22.16_{1.31}$

† Not applicable.

Table 2: Error rates for models 2 and 3

5.1.4 Models 4 and 4' (kNN)

As in the LDA models, we first study a sparse setting (model 4) with the data generating process motivated by the kNN classifier. First, 10 initial points $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{10}$ are generated i.i.d. from $N(\mathbf{0}_{p\times 1}, I_p)$, five of which are labeled as 0 and the other five are labeled as 1. Then each time one of $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{10}\}$ is randomly selected (suppose $\boldsymbol{z}_{k_i}$) and we then generate $\boldsymbol{x}_i \sim N((\boldsymbol{z}_{k_i,S^*}^T, \mathbf{0}_{1\times(p-5)})^T, 0.5^2 I_p)$. Here the minimal discriminative set is $S^* = \{1, 2, 3, 4, 5\}$, $p = 200$, and $n \in \{200, 400, 1000\}$. The results are presented in Table 3 and the average selected percentages of features in $B_1 = 200$ subspaces are presented in Figure 5.

From the left panel of Table 3, it can be seen that the performance of RaSE$_2$-kNN is surprising. It outperforms all the other methods, and the difference between its misclassification rate and others is very prominent. Note that in this case, the Sig-kNN classifier is calculated by applying kNN on only the first five features and also uses leave-one-out

3. To save space, for the rows involving "Gamma/QDA", it represents the independent Gamma classifier in model 2 and the QDA in model 3.

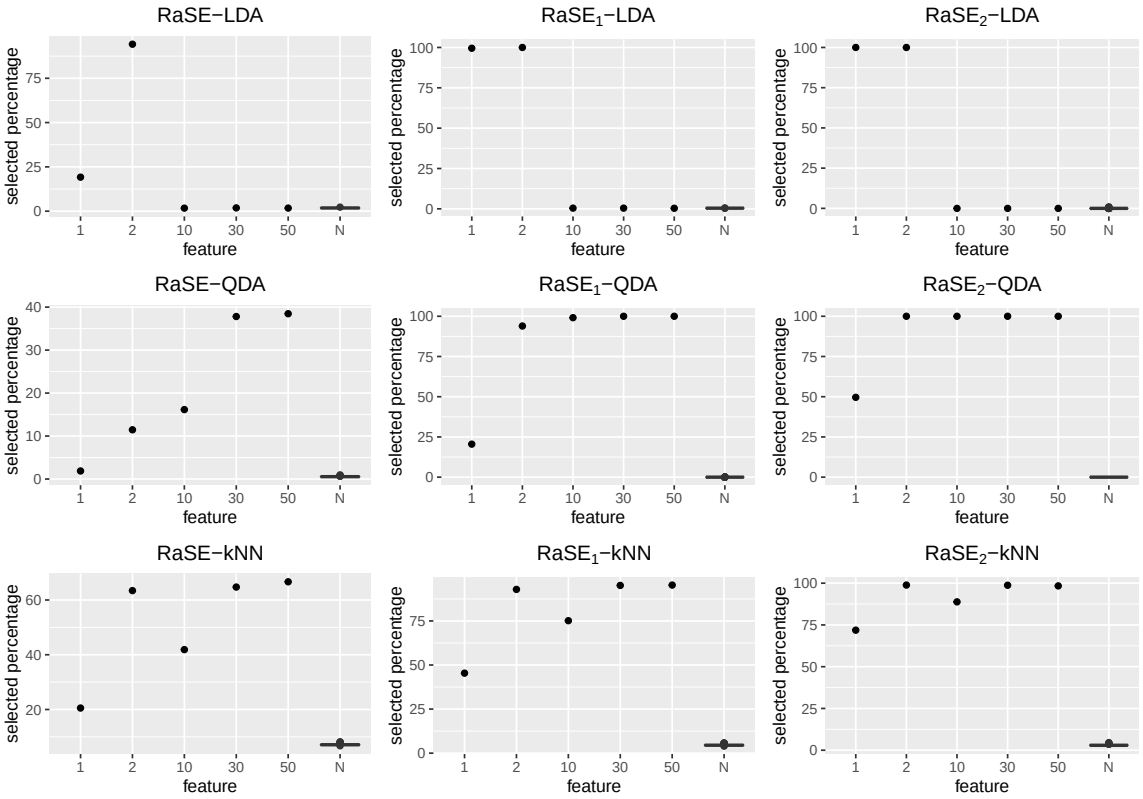

Figure 4: Average selected percentages of features for model 3 in 200 replicates when $n = 1000$

cross-validation to choose k from $\{3, 5, 7, 9, 11\}$. Note that it is not the optimal classifier due to the lack of a true model, which explains why RaSE$_2$-kNN can even achieve a better performance when $n = 400, 1000$.

Figure 5 shows that RaSE, based on all the three base classifiers, can capture features in the minimal discriminative set.

Now, we study a non-sparse setting (model 4'), where the number of signals are increased to 30 and each $\boldsymbol{x}_i \sim N((\boldsymbol{z}^T_{k_i,S^*}, \mathbf{0}_{1\times(p-30)})^T, 2I_p)$. The other parameters and the data generation mechanism are the same as model 4. From the right panel of Table 3, we observe that RaSE$_2$-kNN still achieves the best performance, and the iterations improve the performance of RaSE classifiers under this non-sparse setting. In addition, Figure 6 shows that RaSE can capture the signals while the noises keep a low selected percentage, which again verifies the robustness of RaSE.

Method	Results for model 4			Results for model 4'		
	$n = 200$	$n = 400$	$n = 1000$	$n = 200$	$n = 400$	$n = 1000$
RaSE-LDA	$27.38_{9.53}$	$25.08_{8.40}$	$24.92_{9.05}$	$26.50_{5.78}$	$23.84_{4.98}$	$20.92_{5.20}$
RaSE-QDA	$24.24_{7.20}$	$22.62_{6.77}$	$22.04_{6.73}$	$29.28_{4.34}$	$26.73_{3.63}$	$24.77_{3.35}$
RaSE-kNN	$13.26_{5.03}$	$10.67_{4.44}$	$8.85_{4.05}$	$20.83_{4.50}$	$15.70_{3.74}$	$10.33_{2.90}$
RaSE$_1$-LDA	$25.89_{10.13}$	$23.05_{8.49}$	$23.42_{8.95}$	$21.28_{4.76}$	$18.59_{3.97}$	$16.87_{3.92}$
RaSE$_1$-QDA	$13.83_{5.88}$	$12.54_{5.79}$	$12.70_{5.51}$	$21.97_{5.37}$	$19.32_{4.65}$	$15.84_{4.36}$
RaSE$_1$-kNN	$\mathit{7.51}_{3.80}$	$\mathit{6.16}_{3.48}$	$\mathit{5.90}_{3.12}$	$\mathit{16.24}_{3.52}$	$11.09_{2.84}$	$7.08_{2.04}$
RaSE$_2$-LDA	$27.49_{10.13}$	$23.39_{8.51}$	$23.39_{9.05}$	$20.73_{4.99}$	$17.34_{3.99}$	$15.72_{3.93}$
RaSE$_2$-QDA	$13.15_{5.00}$	$11.90_{5.38}$	$12.15_{5.22}$	$20.94_{5.06}$	$18.61_{4.61}$	$14.96_{4.11}$
RaSE$_2$-kNN	$\mathbf{7.06}_{3.62}$	$\mathbf{5.89}_{3.32}$	$\mathbf{5.74}_{3.02}$	$\mathbf{13.62}_{3.31}$	$\mathbf{8.64}_{2.40}$	$\mathbf{5.36}_{1.58}$
RP-LDA	$28.03_{8.91}$	$25.48_{7.80}$	$24.83_{8.09}$	$22.16_{4.49}$	$18.84_{4.42}$	$16.84_{4.10}$
RP-QDA	$26.22_{7.66}$	$23.93_{6.97}$	$22.75_{7.00}$	$21.37_{4.29}$	$17.58_{3.84}$	$15.53_{3.73}$
RP-kNN	$26.63_{8.09}$	$24.49_{7.15}$	$23.32_{7.35}$	$22.37_{4.69}$	$18.69_{4.19}$	$16.50_{3.95}$
LDA	$47.51_{2.66}$	$33.27_{7.65}$	$27.89_{8.97}$	$46.06_{3.05}$	$25.16_{4.12}$	$17.78_{4.05}$
QDA	—†	—†	$36.70_{4.83}$	—†	—†	$30.45_{2.89}$
kNN	$24.49_{6.64}$	$21.04_{6.60}$	$19.07_{6.50}$	$24.73_{4.35}$	$20.06_{3.95}$	$15.91_{3.59}$
sLDA	$24.90_{9.41}$	$22.80_{8.19}$	$23.22_{8.79}$	$19.78_{4.92}$	$16.41_{3.98}$	$14.59_{3.69}$
RAMP	$22.14_{11.50}$	$15.01_{7.83}$	$12.82_{7.13}$	$24.79_{6.49}$	$18.59_{4.56}$	$13.13_{3.13}$
NSC	$27.70_{9.44}$	$25.71_{8.35}$	$25.82_{8.72}$	$22.09_{6.20}$	$18.17_{4.51}$	$16.17_{4.09}$
RF	$23.64_{8.05}$	$17.39_{6.19}$	$14.84_{5.73}$	$21.73_{5.44}$	$15.70_{3.65}$	$11.54_{2.80}$
Sig-kNN	$6.89_{3.40}$	$6.03_{3.37}$	$6.01_{3.21}$	$7.60_{2.19}$	$5.40_{1.76}$	$4.06_{1.43}$

† Not applicable.

Table 3: Error rates for models 4 and 4'

5.2 Real-data Experiments

5.2.1 Madelon

The madelon data set (`http://archive.ics.uci.edu/ml/datasets/madelon`) is an artificial data set containing data points grouped in 32 clusters placed on the vertices of a five-dimensional hypercube and randomly labeled as 0 or 1 (Guyon et al., 2005). It consists of 2000 observations, 1000 of which are class 0, and the other 1000 are class 1. There are 500 features, among which only 20 are informative, and the others have no predictive power. The training sample size is set as $n \in \{200, 500, 1000\}$ for three different settings, and each time the remained data is used as test data. 200 replicates are applied, and the average misclassification rate with a standard deviation of all methods is reported in Table 4. Figure 7 represents the average selected percentage of features in different RaSE models when $n = 1000$.

From the left panel of Table 4, we can see that the misclassification rate of RaSE$_1$-kNN outperforms all the other methods in all the three settings. Figure 7 shows us that the RaSE model leads to sparse solutions since most of the features have frequencies that are close to zero.

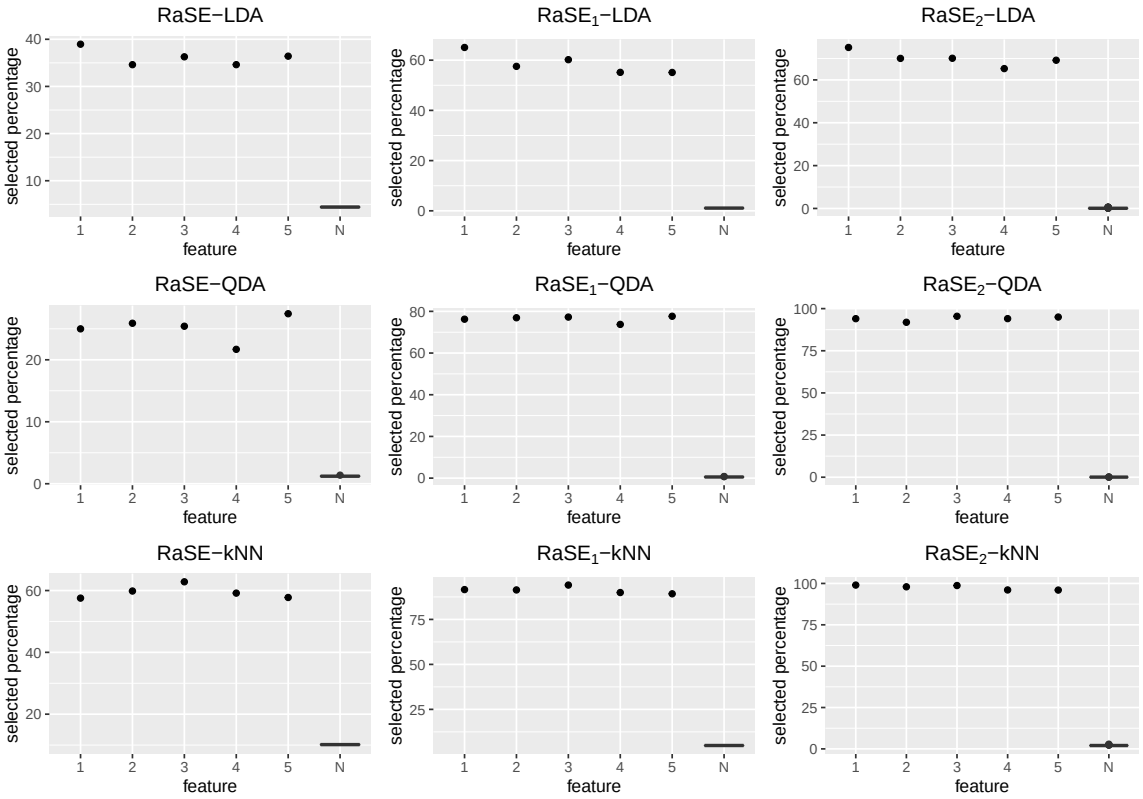

Figure 5: Average selected percentages of features for model 4 in 200 replicates when $n = 1000$

5.2.2 Musk

The musk data set (`https://archive.ics.uci.edu/ml/datasets/Musk+(Version+2)`) contains 6598 observations with 5581 non-musk (class 0) and 1017 musk (class 1) molecules. The molecule needs to be classified based on $p = 166$ shape measurements (Dua and Graff, 2019). The training sample size is set to be 200, 500, 1000, for each setting, and the remaining observations are used as the test data. 200 replicates are considered, and the average misclassification rates and standard deviations are reported in Table 4.

When the training sample size is 200, RP-kNN achieves the lowest misclassification rate, and RaSE-LDA, RaSE-kNN, RaSE$_1$-LDA, RaSE$_1$-kNN, sLDA, RP-QDA, and RF also have a good performance. As the sample size increases, RaSE-kNN turns to be the best one when $n = 500$ and RF yields a comparable performance. When $n = 1000$, RaSE$_1$-kNN and RF outperform the other methods.

5.2.3 Mice Protein Expression

The mice protein expression data set (`https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression`) contains 1080 instances with 570 healthy mice (class 0) and 510 mice

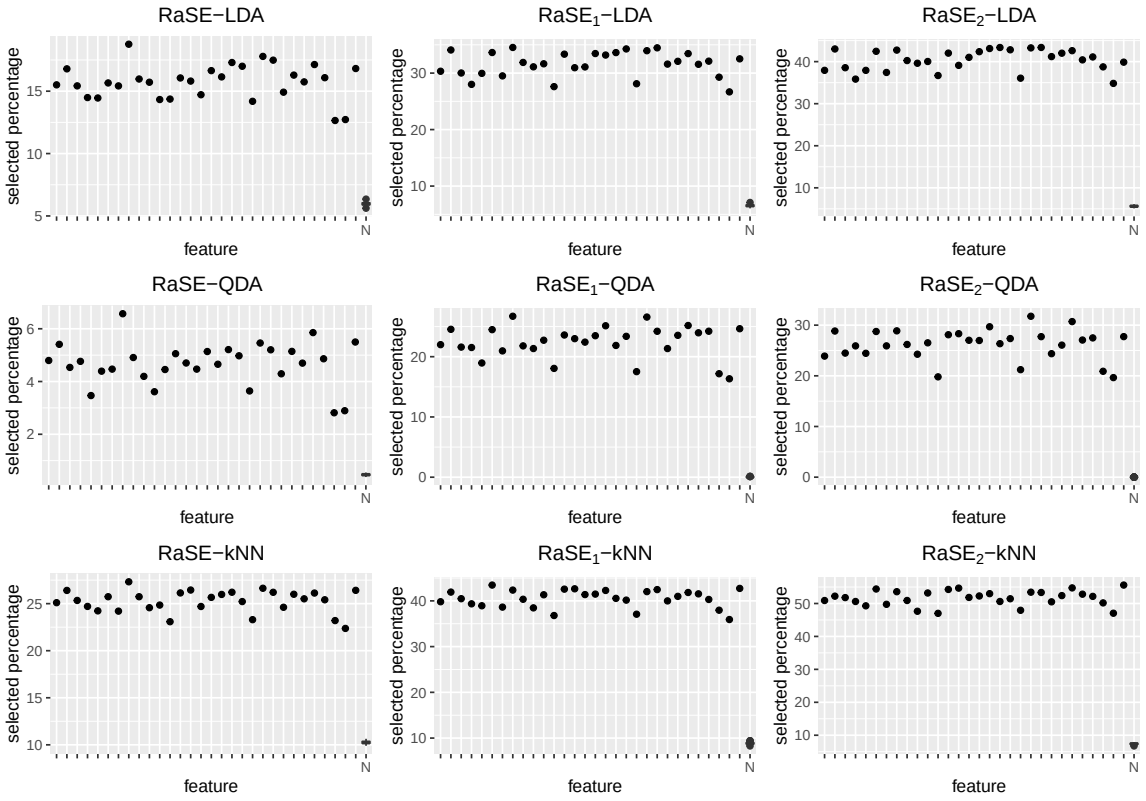

Figure 6: Average selected percentages of features for model 4 in 200 replicates when $n = 1000$

with Down's syndrome (class 1). There are 77 features representing the expression of 77 different proteins (Higuera et al., 2015). Training samples of size 200, 500, 800 are considered, and the remaining observations are set as the test data.

The average of test misclassification rates and the standard deviations of 200 replicates are calculated with results reported in Table 5. When $n = 200$, sLDA achieves the lowest error among all approaches, and RaSE-kNN achieves a similar performance. As the sample size increases to 500 and 800, the average misclassification rate of RaSE-kNN and RaSE$_1$-kNN decrease significantly, and they become the best classifier when $n = 500$ and 800, respectively. When $n = 800$, RP-kNN and RF also have a similar performance.

5.2.4 Hand-written Digits Recognition

The hand-written digits recognition data set (`https://archive.ics.uci.edu/ml/datasets/Multiple+Features`) consists of features of hand-written numerals (0-9) extracted from a collection of Dutch utility maps (Dua and Graff, 2019). Here we use the mfeat-fou data set, which records 76 Fourier coefficients of the character shapes. We extract the observations corresponding to number 7 (class 0) and 9 (class 1) from the original data. There are 400 observations, 200 of which belong to class 0, and the remaining 200 belong to class 1. The

Method	Madelon			Musk		
	$n = 200$	$n = 500$	$n = 1000$	$n = 200$	$n = 500$	$n = 1000$
RaSE-LDA	$43.39_{3.53}$	$39.87_{1.63}$	$39.16_{1.29}$	$\mathit{10.55}_{1.22}$	$8.86_{0.77}$	$7.87_{0.46}$
RaSE-QDA	$43.83_{3.28}$	$40.58_{1.73}$	$40.01_{1.57}$	$12.55_{7.02}$	$9.16_{0.97}$	$7.77_{0.73}$
RaSE-kNN	$35.02_{2.52}$	$26.78_{2.71}$	$21.45_{1.91}$	$\mathit{10.26}_{1.81}$	$\mathbf{7.33}_{0.96}$	$\mathit{5.76}_{0.74}$
RaSE$_1$-LDA	$45.98_{2.49}$	$39.76_{1.81}$	$38.57_{1.11}$	$\mathit{10.56}_{1.19}$	$8.88_{0.81}$	$7.82_{0.46}$
RaSE$_1$-QDA	$44.46_{5.36}$	$37.21_{2.85}$	$34.63_{2.15}$	$16.71_{9.58}$	$11.02_{2.99}$	$8.60_{0.85}$
RaSE$_1$-kNN	$\mathbf{26.05}_{3.33}$	$\mathbf{16.66}_{1.33}$	$\mathbf{13.57}_{1.00}$	$\mathit{10.52}_{1.95}$	$\mathit{7.49}_{0.97}$	$\mathbf{5.71}_{0.78}$
RP-LDA	$41.18_{1.77}$	$39.86_{1.14}$	$39.41_{1.17}$	$12.58_{1.86}$	$10.19_{0.76}$	$9.50_{0.46}$
RP-QDA	$39.97_{1.53}$	$39.29_{1.57}$	$38.79_{1.49}$	$\mathit{10.70}_{1.95}$	$8.75_{0.83}$	$8.18_{0.45}$
RP-kNN	$40.10_{1.66}$	$39.07_{1.54}$	$38.40_{1.43}$	$\mathbf{10.01}_{1.66}$	$\mathit{8.05}_{1.06}$	$6.89_{0.84}$
LDA	—†	$49.82_{1.24}$	$47.60_{1.49}$	$25.60_{3.76}$	$9.06_{0.80}$	$6.97_{0.40}$
QDA	—†	—†	—†	—†	—†	—†
kNN	$35.89_{1.42}$	$31.84_{1.43}$	$28.56_{1.56}$	$\mathit{11.35}_{1.55}$	$\mathit{8.19}_{0.86}$	$6.53_{0.55}$
sLDA	$43.01_{2.98}$	$40.40_{1.94}$	$39.57_{1.52}$	$\mathit{10.68}_{1.54}$	$\mathit{7.81}_{0.68}$	$6.85_{0.40}$
RAMP	$49.09_{3.46}$	$41.88_{5.16}$	$38.58_{1.36}$	$14.35_{1.93}$	$11.50_{1.40}$	$9.72_{1.10}$
NSC	$41.49_{2.32}$	$40.10_{1.09}$	$40.05_{1.24}$	$20.72_{5.11}$	$24.12_{3.62}$	$25.58_{1.01}$
RF	$43.91_{2.49}$	$38.59_{1.56}$	$34.17_{1.46}$	$\mathit{10.83}_{1.44}$	$\mathit{7.60}_{0.66}$	$\mathbf{5.71}_{0.48}$

† Not applicable.

Table 4: Error rates for madelon and musk data sets

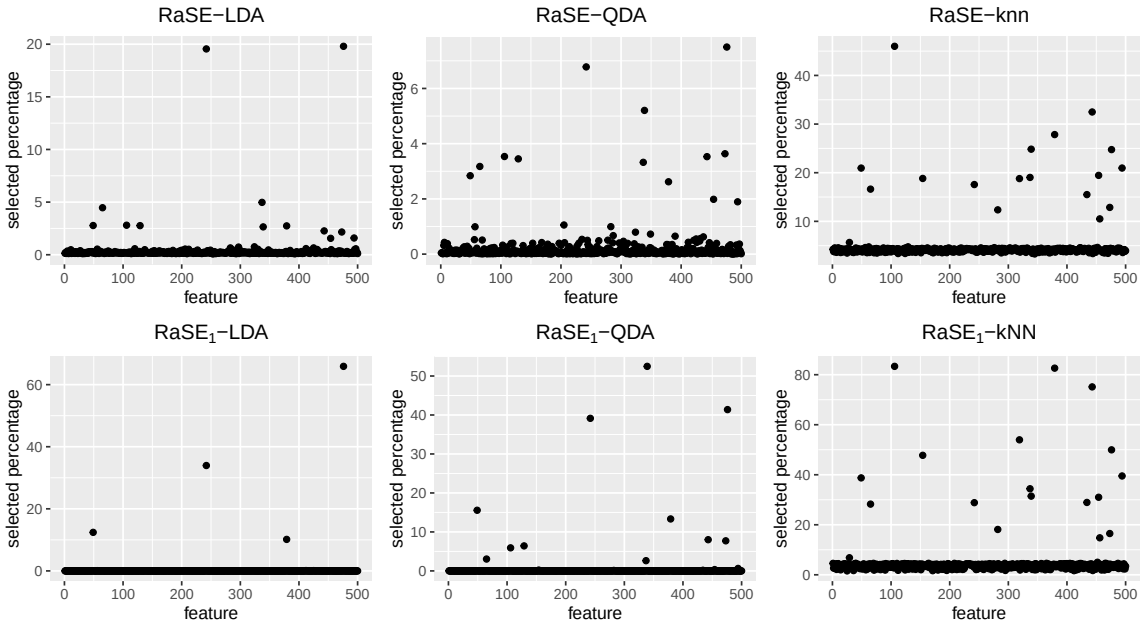

Figure 7: Average selected percentages of features for madelon data set in 200 replicates when $n = 1000$

Method	Mice protein expression			Hand-written digits recognition		
	$n = 200$	$n = 500$	$n = 800$	$n = 50$	$n = 100$	$n = 200$
RaSE-LDA	$7.41_{1.14}$	$5.70_{0.93}$	$4.65_{1.24}$	$\mathit{1.56}_{0.85}$	$1.13_{0.59}$	$\mathit{0.80}_{0.54}$
RaSE-QDA	$9.14_{2.58}$	$4.81_{1.17}$	$3.44_{1.23}$	$2.50_{1.47}$	$1.89_{0.91}$	$1.47_{0.96}$
RaSE-kNN	$6.80_{1.88}$	$\mathbf{1.55}_{0.88}$	$\mathit{0.62}_{0.55}$	$1.86_{0.96}$	$1.12_{0.66}$	$\mathit{0.75}_{0.45}$
RaSE$_1$-LDA	$7.24_{1.10}$	$5.53_{1.02}$	$4.49_{1.23}$	$\mathbf{1.06}_{0.63}$	$\mathit{0.70}_{0.35}$	$\mathbf{0.53}_{0.40}$
RaSE$_1$-QDA	$9.38_{2.21}$	$5.16_{1.16}$	$3.40_{1.16}$	$2.18_{1.66}$	$1.18_{0.71}$	$\mathit{0.85}_{0.61}$
RaSE$_1$-kNN	$7.43_{2.00}$	$\mathit{1.70}_{0.87}$	$\mathbf{0.60}_{0.56}$	$1.72_{0.95}$	$\mathit{1.02}_{0.62}$	$\mathit{0.60}_{0.44}$
RP-LDA	$24.84_{2.91}$	$22.79_{2.50}$	$22.34_{2.55}$	$1.75_{1.29}$	$1.22_{0.67}$	$1.04_{0.61}$
RP-QDA	$18.31_{2.57}$	$16.19_{2.08}$	$15.66_{2.38}$	$2.12_{1.67}$	$1.28_{0.94}$	$\mathit{0.92}_{0.62}$
RP-kNN	$11.77_{2.54}$	$2.57_{0.89}$	$\mathit{0.92}_{0.68}$	$\mathit{1.68}_{1.34}$	$\mathit{1.03}_{0.60}$	$\mathit{0.84}_{0.59}$
LDA	$7.07_{1.37}$	$3.88_{0.85}$	$3.13_{1.08}$	—†	$1.82_{0.96}$	$1.01_{0.56}$
QDA	—†	—†	—†	—†	—†	$3.25_{2.32}$
kNN	$20.53_{2.47}$	$7.75_{1.44}$	$2.80_{1.21}$	$\mathit{1.42}_{1.32}$	$\mathbf{0.67}_{0.41}$	$\mathit{0.60}_{0.47}$
sLDA	$\mathbf{5.70}_{1.10}$	$3.95_{0.91}$	$3.13_{1.05}$	$2.30_{1.36}$	$1.71_{1.27}$	$1.15_{0.95}$
RAMP	$11.76_{2.42}$	$8.52_{1.69}$	$7.02_{1.89}$	$3.31_{1.75}$	$2.26_{1.19}$	$1.70_{0.87}$
NSC	$30.49_{3.31}$	$29.70_{2.76}$	$29.88_{3.02}$	$3.22_{1.44}$	$3.53_{1.23}$	$3.59_{1.50}$
RF	$8.32_{1.71}$	$2.62_{0.94}$	$\mathit{1.04}_{0.73}$	$2.34_{1.24}$	$1.63_{0.73}$	$1.37_{0.74}$

† Not applicable.

Table 5: Error rates for mice protein expression and hand-written digits recognition data sets

training samples of size 50, 100, 200 are used, and the remained data is used as the test data.

Average of test misclassification rates and standard deviations are reported in Table 5, from which it can be seen that when $n = 50, 200$, RaSE$_1$-LDA enjoys the minimal test misclassification rate while standard kNN method is the best when $n = 100$. And we also note that all the RaSE classifiers get improved after 1 iteration for all three settings, implying that the underlying classification problem may be a sparse one.

6. Discussion

6.1 Summary

In this work, we introduce a flexible ensemble classification framework named RaSE, which is designed to solve the sparse classification problem. To select the optimal subspace for each weak learner, we define a new information criterion, ratio information criterion (RIC), based on Kullback-Leibler divergence, and it is shown to achieve screening consistency and weak consistency under some general model conditions. This guarantees that each weak learner is trained using features in the minimal discriminative set with a high probability for a sufficiently large sample size and the number of random subspaces. We also investigate the consistency of RIC for LDA and QDA models under some specific conditions. The theoretical analysis of RaSE classifiers with specific base classifiers is conducted. In addition,

we present two versions of RaSE algorithms, that is, the vanilla version (RaSE) and the iterative version (RaSE$_T$). We also apply RaSE for feature ranking based on the average selected percentage of features in subspaces. Theoretical analysis shows that when the stepwise detectable condition holds and the signal is sufficiently sparse, the iterative RaSE can cover the minimal discriminative set with high probability after a few iterations, with the required number of random subspaces smaller than that for the vanilla RaSE.

Multiple numerical experiments, including simulations and real data, verify that RaSE is a favorable classification method for sparse classification problems.

The RaSE algorithms are available in R package `RaSEn` (`https://cran.r-project.org/web/packages/RaSEn/index.html`).

6.2 Future Work

There are many interesting directions along which RaSE can be extended and explored. An interesting question is how to extend RaSE and RIC into multi-class problems. For example, the pair-wise KL divergences can be used to define the multi-class RIC. Moreover, we can also apply RaSE for variable selection. We can conduct thresholding to the average selected percentage of features in B_1 subspaces to select variables. When the sample size is small, a bootstrap-type idea can be used, and each time we apply RaSE on a bootstrap sample and at the end, take the average for the selected percentage to do variable selection or feature ranking. Finally, we aggregate the classifiers by taking a simple average over all weak learners. However, the boosting-type idea can also be applied here to assign different weights for different weak learners according to the training error, which may further improve the performance of RaSE. In addition, the distribution for random subspaces can also be chosen to be different for each weak learner and can also be updated using a similar idea to boosting.

Acknowledgments

The authors would like to thank the Action Editor and anonymous referees for many constructive comments which have greatly improved the paper. This work was partially supported by National Science Foundation CAREER grant DMS-2013789.

Appendix A. Additional Figures of Simulations

We present figures of the selected percentage for each feature when n equals to the smallest value among three settings.

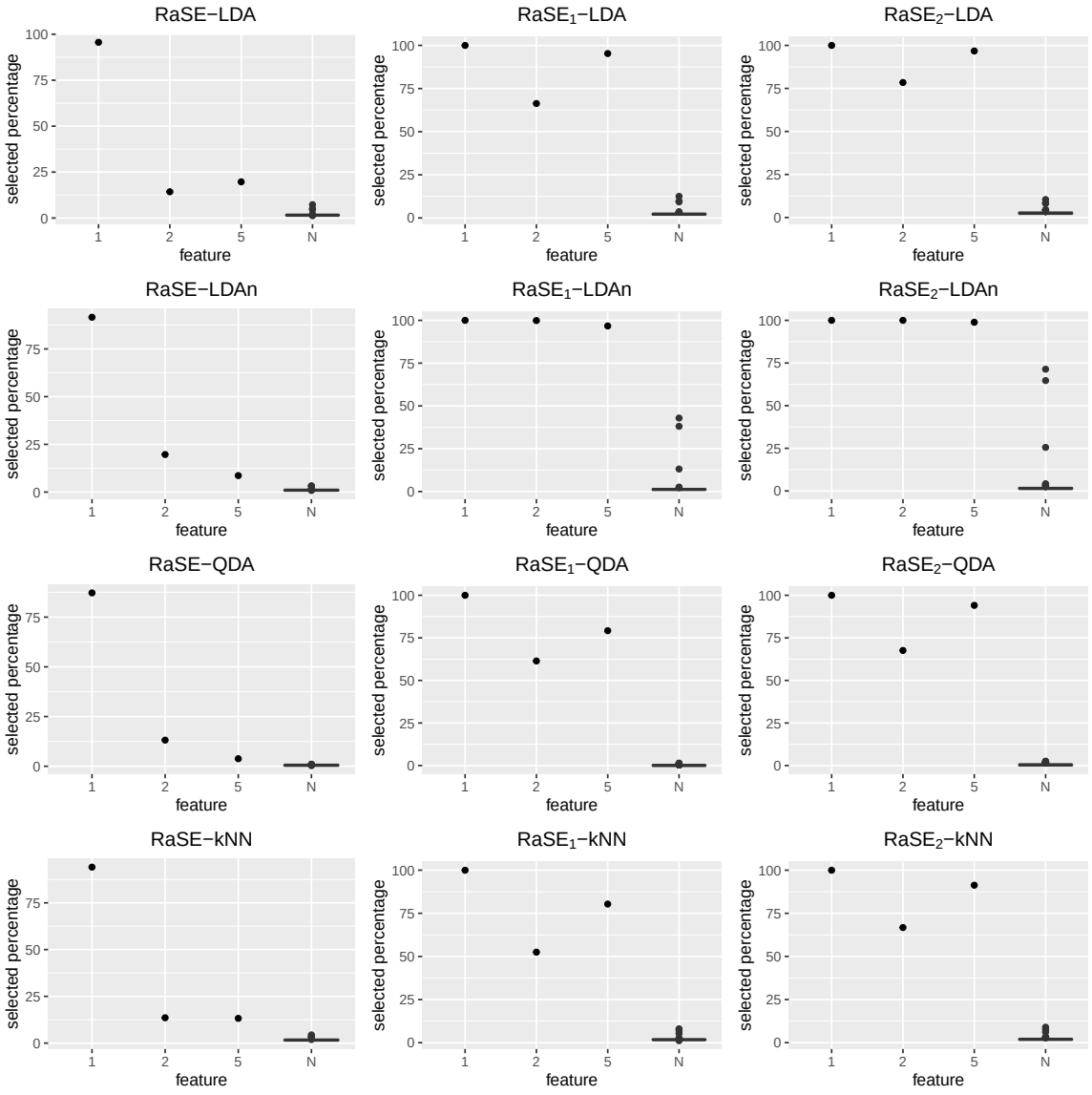

Figure 8: Average selected percentages of features for model 1 in 200 replicates when $n = 200$

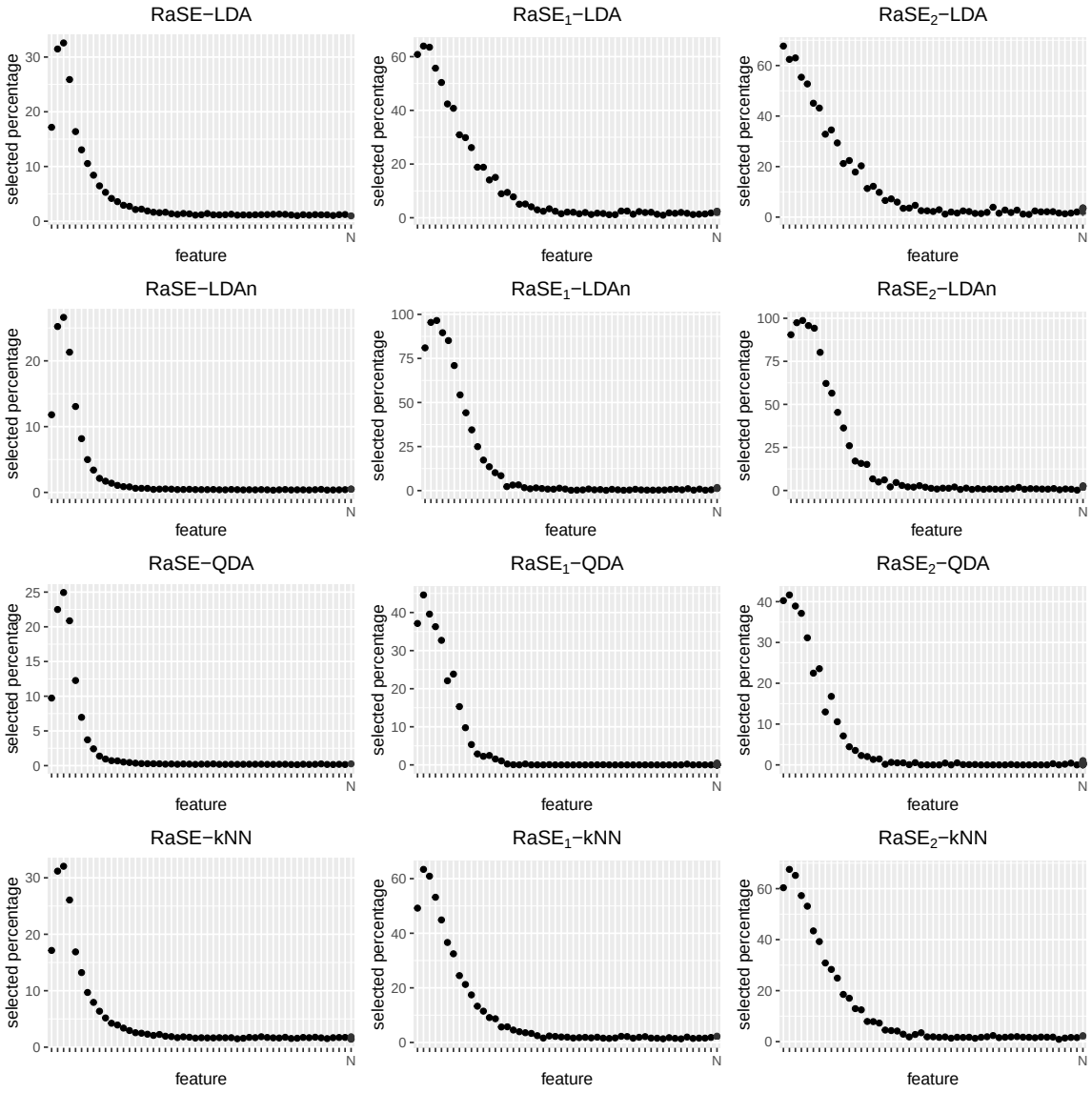

Figure 9: Average selected percentages of features for model 1' in 200 replicates when $n = 200$

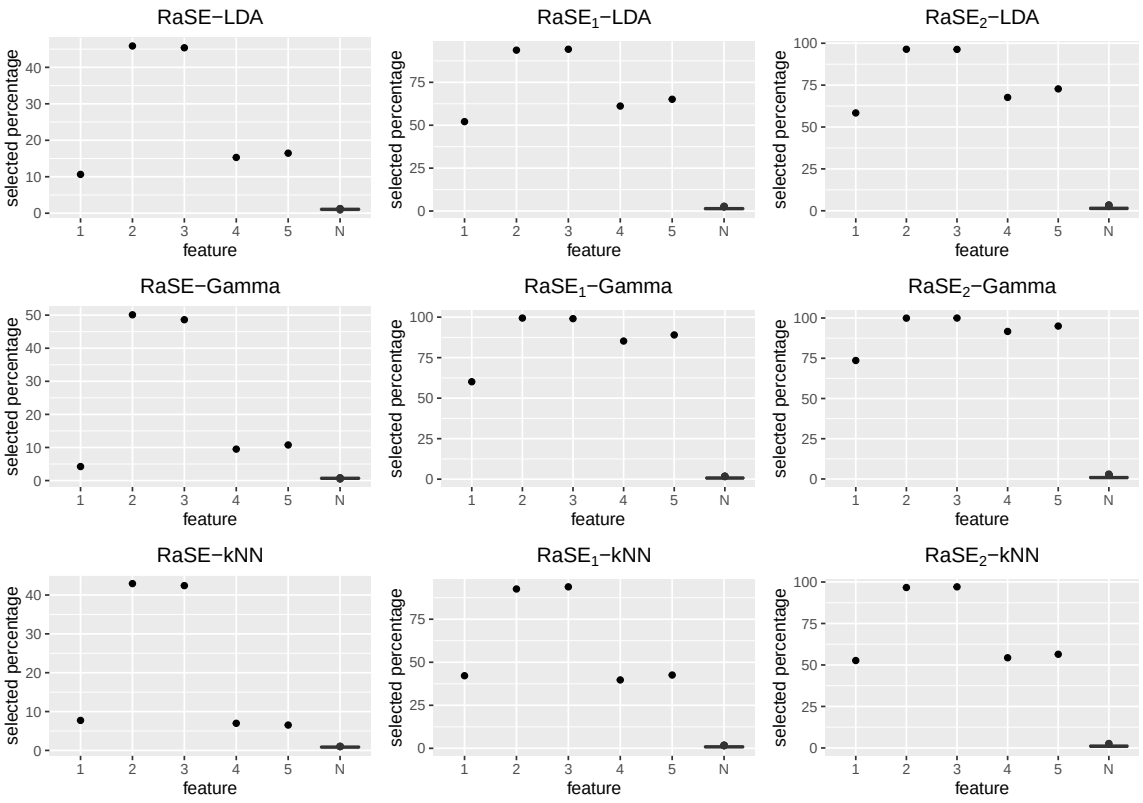

Figure 10: Average selected percentages of features for model 2 in 200 replicates when $n = 100$

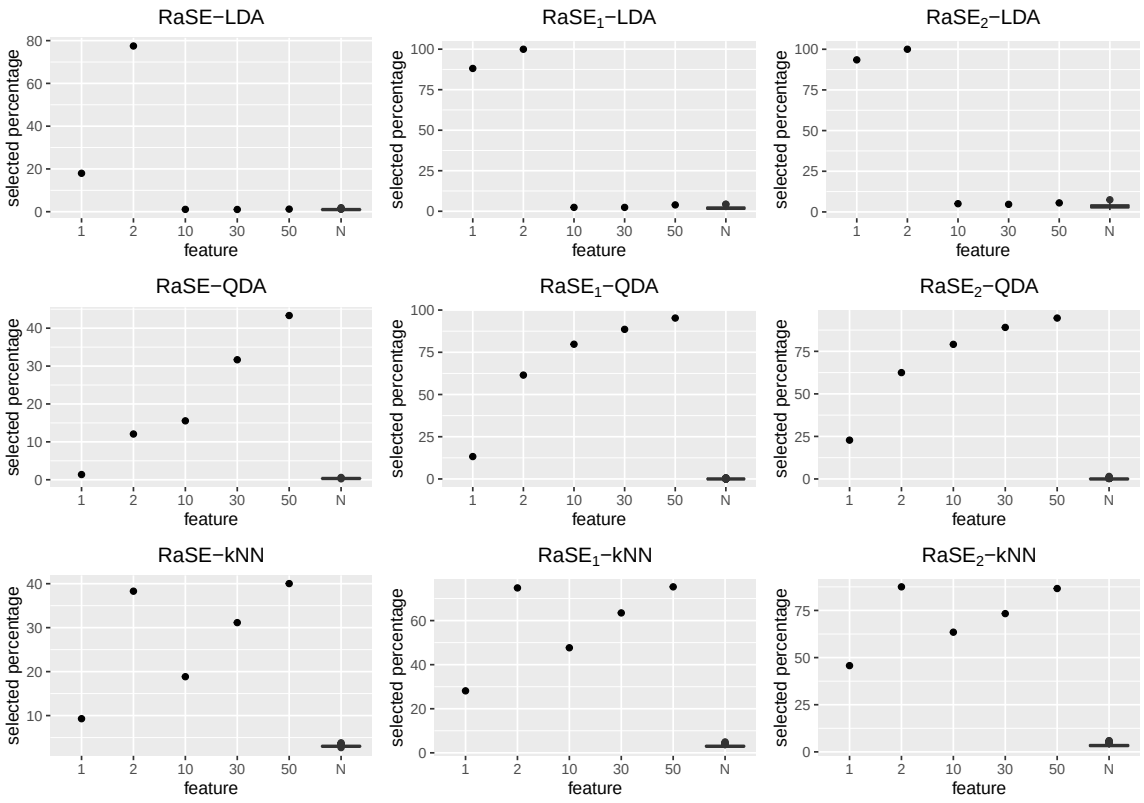

Figure 11: Average selected percentages of features for model 3 in 200 replicates when $n = 200$

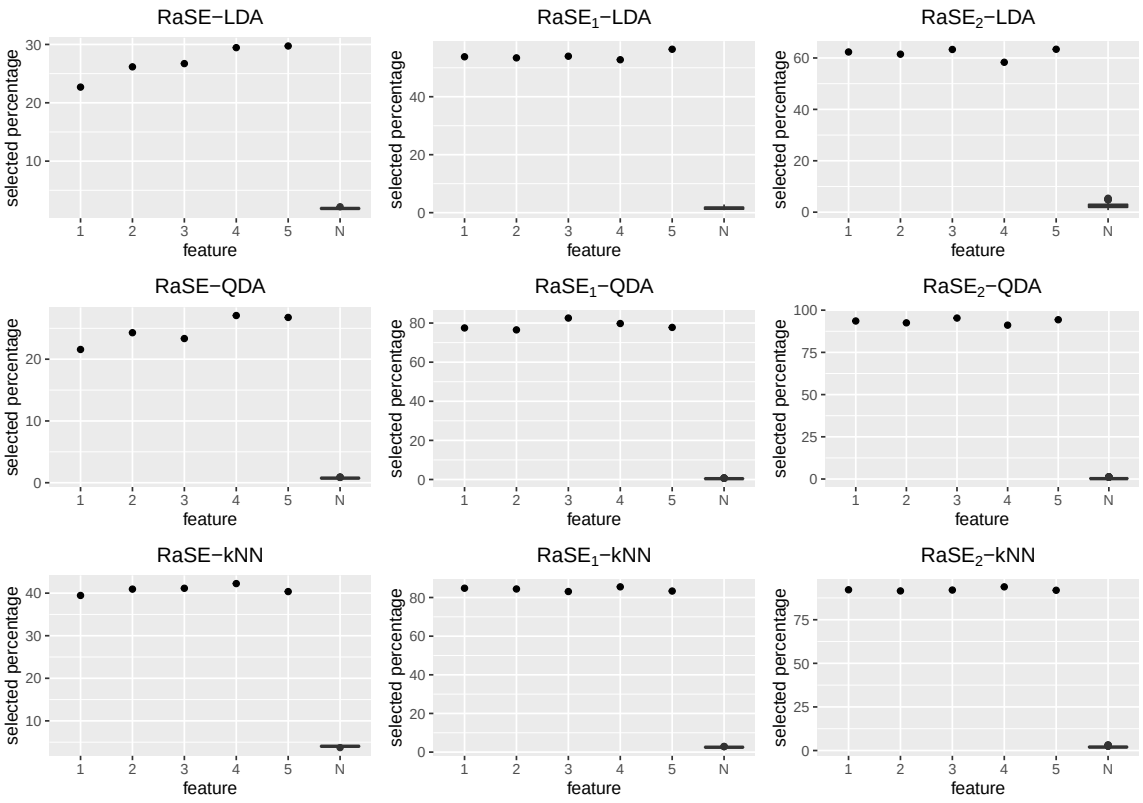

Figure 12: Average selected percentages of features for model 4 in 200 replicates when $n = 200$

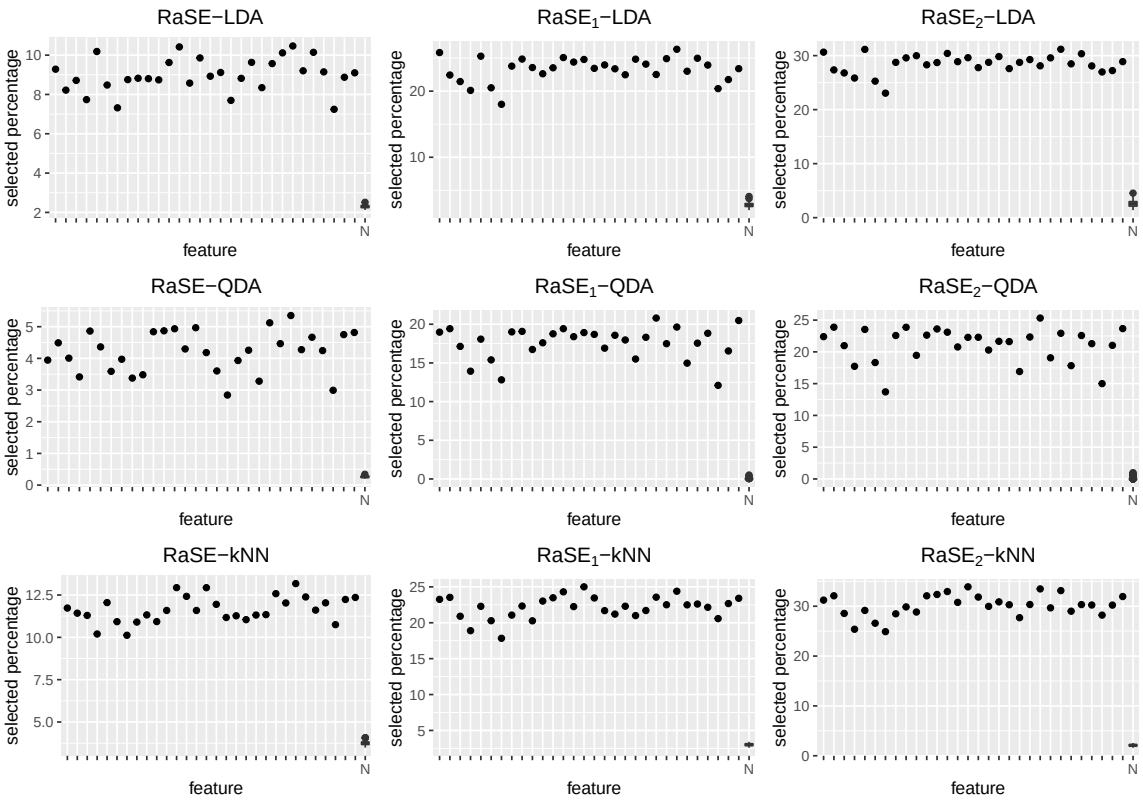

Figure 13: Average selected percentages of features for model 4' in 200 replicates when $n = 200$

Appendix B. Main Proofs

In the following proofs, for convenience, we use C to represent a positive constant, which could be different at different occurences.

B.1 Proof of Proposition 3

(i) The conditional probability is

$$\mathrm{P}(y=1|\boldsymbol{x}) = \frac{\pi_1 f^{(1)}(\boldsymbol{x})}{\pi_1 f^{(1)}(\boldsymbol{x}) + \pi_0 f^{(0)}(\boldsymbol{x})}.$$

The definition of discriminative set is equivalent to

$$\mathrm{P}(y=1|\boldsymbol{x}) = \mathrm{P}(y=1|\boldsymbol{x}_S), \tag{14}$$

or

$$\mathrm{P}(y=0|\boldsymbol{x}) = \mathrm{P}(y=0|\boldsymbol{x}_S)$$

almost surely, which is also equivalent to

$$\frac{f^{(1)}(\boldsymbol{x})}{f^{(0)}(\boldsymbol{x})} = \frac{f_S^{(1)}(\boldsymbol{x}_S)}{f_S^{(0)}(\boldsymbol{x}_S)}. \tag{15}$$

(ii) First, we observe the condition is equivalent to $f^{(1)}(\boldsymbol{x}) = h(\boldsymbol{x}_S) f^{(0)}(\boldsymbol{x})$. Taking integration on both sides with respect to $\boldsymbol{x}_{S^c}$, we get $f_S^{(1)}(\boldsymbol{x}_S) = h(\boldsymbol{x}_S) f_S^{(0)}(\boldsymbol{x}_S)$. Due to the equivalence of (14) and (15), S is a discriminative set.

B.2 Proof of Proposition 5

To facilitate our analysis, we first state the following lemma.

Lemma 26 *A discriminative set S is unique, if it will not be a discriminative set anymore after removing any features from it.*

Proof [Proof of Lemma 26] Suppose the conclusion is not correct, then there exist two different discriminative sets S_1 and S_2 satisfying such a property, that is, deleting any features from them leads to non-discriminative sets. Then according to Proposition 3 (i), we have

$$\frac{f_{S_1}^{(1)}(\boldsymbol{x}_{S_1})}{f_{S_1}^{(0)}(\boldsymbol{x}_{S_1})} = \frac{f_{S_2}^{(1)}(\boldsymbol{x}_{S_2})}{f_{S_2}^{(0)}(\boldsymbol{x}_{S_2})}, \tag{16}$$

almost surely w.r.t $\mathrm{P}^{\boldsymbol{x}}$, where $\mathrm{P}^{\boldsymbol{x}} = \pi_0 \mathrm{P}^{(0)} + \pi_1 \mathrm{P}^{(1)}$. Since $f^{(0)}, f^{(1)}$ are supported on the whole $\mathbb{R}^p$, $\mathrm{P}^{\boldsymbol{x}}$ dominates the Lebesgue measure. Combined with the explicit form of density functions of Gaussian distribution and denoting $\Omega_{S_1,S_1} = (\Sigma^{(1)}_{S_1,S_1})^{-1} - (\Sigma^{(0)}_{S_1,S_1})^{-1}, \Omega_{S_2,S_2} = (\Sigma^{(1)}_{S_2,S_2})^{-1} - (\Sigma^{(0)}_{S_2,S_2})^{-1}, \boldsymbol{\delta}_{S_1} = (\Sigma^{(0)}_{S_1,S_1})^{-1}\boldsymbol{\mu}^{(0)}_{S_1} - (\Sigma^{(1)}_{S_1,S_1})^{-1}\boldsymbol{\mu}^{(1)}_{S_1}, \boldsymbol{\delta}_{S_2} = (\Sigma^{(0)}_{S_2,S_2})^{-1}\boldsymbol{\mu}^{(0)}_{S_2} - (\Sigma^{(1)}_{S_2,S_2})^{-1}\boldsymbol{\mu}^{(1)}_{S_2}$, it can be obtained from (4) and (16) that

$$c + \frac{1}{2}\boldsymbol{x}_{S_1}^T \Omega_{S_1,S_1} \boldsymbol{x}_{S_1} + \boldsymbol{\delta}_{S_1}^T \boldsymbol{x}_{S_1} = c' + \frac{1}{2}\boldsymbol{x}_{S_2}^T \Omega_{S_2,S_2} \boldsymbol{x}_{S_2} + \boldsymbol{\delta}_{S_2}^T \boldsymbol{x}_{S_2},$$

where c, c' are constants irrelative to S_1, S_2. Considering the combined vector $\boldsymbol{x}_{S_1 \cup S_2}$, there exists some matrix Ω and vector $\boldsymbol{\delta}$ such that the equation above can be simplified as

$$\boldsymbol{x}_{S_1 \cup S_2}^T \Omega \boldsymbol{x}_{S_1 \cup S_2} + \boldsymbol{\delta}^T \boldsymbol{x}_{S_1 \cup S_2} + c - c' = 0, \tag{17}$$

for almost every $\boldsymbol{x}_{S_1 \cup S_2}$ in the Euclidean space. Since S_1 is a discriminative set, by Example 2, for every feature $j \in S_1$, the corresponding row of Ω_{S_1, S_1} is not zero vector or the corresponding component of $\boldsymbol{\delta}_{S_1}$ is not zero. The same argument holds for S_2 as well. Since $S_1 \backslash S_2$ and $S_2 \backslash S_1$ are not empty, at least one of Ω and $\boldsymbol{\delta}$ contain non-zero components. However, it's obvious that (17) cannot hold for almost every $\boldsymbol{x}_{S_1 \cup S_2}$ in Euclidean space, which leads to contradiction. Thus S^* is unique. ■

Now let's proceed on the proof of Proposition 5.

By Definition 1, the minimal discriminative set S^* satisfies the property described in Lemma 26, therefore by this lemma S^* is obviously unique, which implies (i).

For (ii), if there exists a discriminative set $S \not\supseteq S^*$, we can remove elements from S until we arrive at a discriminative set S' that satisfies the property stated in Lemma 26. It's apparent $S' \neq S^*$, which contradicts with Lemma 26.

(iii) is trivial from Definition 1.

B.3 Proof of Proposition 8

By Definition 6 and the fact that $\deg(S) = |S| + 1$, it's easy to obtain that

$$\begin{aligned}
\mathrm{RIC}_n(S) = \frac{2}{n} \sum_{i=1}^{n} \mathbb{1}(y_i = 0) \Bigg[& ((\hat{\boldsymbol{\mu}}_S^{(1)})^T \hat{\Sigma}_{S,S}^{-1} - (\hat{\boldsymbol{\mu}}_S^{(0)})^T \hat{\Sigma}_{S,S}^{-1}) \boldsymbol{x}_{i,S} + \frac{1}{2} (\hat{\boldsymbol{\mu}}_S^{(0)})^T \hat{\Sigma}_{S,S}^{-1} \hat{\boldsymbol{\mu}}_S^{(0)} \\
& - \frac{1}{2} (\hat{\boldsymbol{\mu}}_S^{(1)})^T \hat{\Sigma}_{S,S}^{-1} \hat{\boldsymbol{\mu}}_S^{(1)} \Bigg] + \frac{2}{n} \sum_{i=1}^{n} \mathbb{1}(y_i = 1) \Bigg[((\hat{\boldsymbol{\mu}}_S^{(0)})^T \hat{\Sigma}_{S,S}^{-1} - (\hat{\boldsymbol{\mu}}_S^{(1)})^T \hat{\Sigma}_{S,S}^{-1}) \boldsymbol{x}_{i,S} \\
& + \frac{1}{2} (\hat{\boldsymbol{\mu}}_S^{(1)})^T \hat{\Sigma}_{S,S}^{-1} \hat{\boldsymbol{\mu}}_S^{(1)} - \frac{1}{2} (\hat{\boldsymbol{\mu}}_S^{(0)})^T \hat{\Sigma}_{S,S}^{-1} \hat{\boldsymbol{\mu}}_S^{(0)} \Bigg] + c_n \cdot (|S| + 1) \\
= & \, 2\hat{\pi}_0 \left[((\hat{\boldsymbol{\mu}}_S^{(1)})^T \hat{\Sigma}_{S,S}^{-1} - (\hat{\boldsymbol{\mu}}_S^{(0)})^T \hat{\Sigma}_{S,S}^{-1}) \hat{\boldsymbol{\mu}}_S^{(0)} + \frac{1}{2} (\hat{\boldsymbol{\mu}}_S^{(0)})^T \hat{\Sigma}_{S,S}^{-1} \hat{\boldsymbol{\mu}}_S^{(0)} - \frac{1}{2} (\hat{\boldsymbol{\mu}}_S^{(1)})^T \hat{\Sigma}_{S,S}^{-1} \hat{\boldsymbol{\mu}}_S^{(1)} \right] \\
& + 2\hat{\pi}_1 \left[((\hat{\boldsymbol{\mu}}_S^{(0)})^T \hat{\Sigma}_{S,S}^{-1} - (\hat{\boldsymbol{\mu}}_S^{(1)})^T \hat{\Sigma}_{S,S}^{-1}) \hat{\boldsymbol{\mu}}_S^{(1)} + \frac{1}{2} (\hat{\boldsymbol{\mu}}_S^{(1)})^T \hat{\Sigma}_{S,S}^{-1} \hat{\boldsymbol{\mu}}_S^{(1)} - \frac{1}{2} (\hat{\boldsymbol{\mu}}_S^{(0)})^T \hat{\Sigma}_{S,S}^{-1} \hat{\boldsymbol{\mu}}_S^{(0)} \right] \\
& + c_n \cdot (|S| + 1) \\
= & -(\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)})^T \hat{\Sigma}_{S,S}^{-1} (\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)}) + c_n \cdot (|S| + 1),
\end{aligned}$$

which completes the proof.

B.4 Proof of Proposition 9

By Definition 6 and the fact $\deg(S) = |S|(|S| + 3)/2 + 1$, it's easy to obtain that

$$\mathrm{RIC}_n(S) = \frac{2}{n} \sum_{i=1}^{n} \mathbb{1}(y_i = 0) \left[\frac{1}{2} \boldsymbol{x}_{i,S}^T ((\hat{\Sigma}_{S,S}^{(0)})^{-1} - (\hat{\Sigma}_{S,S}^{(1)})^{-1}) \boldsymbol{x}_{i,S} + ((\hat{\boldsymbol{\mu}}_S^{(1)})^T (\hat{\Sigma}_{S,S}^{(1)})^{-1} \right.$$

$$
\begin{aligned}
&- (\hat{\boldsymbol{\mu}}_S^{(0)})^T(\hat{\Sigma}_{S,S}^{(0)})^{-1})\boldsymbol{x}_{i,S} + \frac{1}{2}(\hat{\boldsymbol{\mu}}_S^{(0)})^T(\hat{\Sigma}_{S,S}^{(0)})^{-1}\hat{\boldsymbol{\mu}}_S^{(0)} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_S^{(1)})^T(\hat{\Sigma}_{S,S}^{(1)})^{-1}\hat{\boldsymbol{\mu}}_S^{(1)} \\
&+ \log(|\hat{\Sigma}_{S,S}^{(0)}|) - \log(|\hat{\Sigma}_{S,S}^{(1)}|)\Big] + \frac{2}{n}\sum_{i=1}^{n}\mathbb{1}(y_i = 1)\Big[\frac{1}{2}\boldsymbol{x}_{i,S}^T((\hat{\Sigma}_{S,S}^{(1)})^{-1} - (\hat{\Sigma}_{S,S}^{(0)})^{-1})\boldsymbol{x}_{i,S} \\
&+ ((\hat{\boldsymbol{\mu}}_S^{(0)})^T(\hat{\Sigma}_{S,S}^{(0)})^{-1} - (\hat{\boldsymbol{\mu}}_S^{(1)})^T(\hat{\Sigma}_{S,S}^{(1)})^{-1})\boldsymbol{x}_{i,S} + \frac{1}{2}(\hat{\boldsymbol{\mu}}_S^{(1)})^T(\hat{\Sigma}_{S,S}^{(1)})^{-1}\hat{\boldsymbol{\mu}}_S^{(1)} \\
&- \frac{1}{2}(\hat{\boldsymbol{\mu}}_S^{(0)})^T(\hat{\Sigma}_{S,S}^{(0)})^{-1}\hat{\boldsymbol{\mu}}_S^{(0)} + \log(|\hat{\Sigma}_{S,S}^{(1)}|) - \log(|\hat{\Sigma}_{S,S}^{(0)}|)\Big] + c_n \cdot (|S|(|S|+3)/2 + 1).
\end{aligned}
$$

Denote the observations with class r as $\{\boldsymbol{x}_i^{(r)}\}_{i=1}^{n_r}, r = 0, 1$. And it holds that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i = 0)\left[\boldsymbol{x}_{i,S}^T((\hat{\Sigma}_{S,S}^{(0)})^{-1} - (\hat{\Sigma}_{S,S}^{(1)})^{-1})\boldsymbol{x}_{i,S}\right] \\
&= \frac{n_0}{n}\cdot\frac{1}{n_0}\sum_{i=1}^{n_0}\left[(\boldsymbol{x}_{i,S}^{(0)})^T((\hat{\Sigma}_{S,S}^{(0)})^{-1} - (\hat{\Sigma}_{S,S}^{(1)})^{-1})\boldsymbol{x}_{i,S}^{(0)}\right] \\
&= \hat{\pi}_0\cdot\frac{1}{n_0}\sum_{i=1}^{n_0}\mathrm{Tr}\left[((\hat{\Sigma}_{S,S}^{(0)})^{-1} - (\hat{\Sigma}_{S,S}^{(1)})^{-1})\boldsymbol{x}_{i,S}^{(0)}(\boldsymbol{x}_{i,S}^{(0)})^T\right] \\
&= \hat{\pi}_0\mathrm{Tr}\left[((\hat{\Sigma}_{S,S}^{(0)})^{-1} - (\hat{\Sigma}_{S,S}^{(1)})^{-1})\frac{1}{n_0}\sum_{i=1}^{n_0}\boldsymbol{x}_{i,S}^{(0)}(\boldsymbol{x}_{i,S}^{(0)})^T\right] \\
&= \hat{\pi}_0\mathrm{Tr}\left[((\hat{\Sigma}_{S,S}^{(0)})^{-1} - (\hat{\Sigma}_{S,S}^{(1)})^{-1})(\hat{\boldsymbol{\mu}}_S^{(0)}(\hat{\boldsymbol{\mu}}_S^{(0)})^T + \hat{\Sigma}_{S,S}^{(0)})\right] \\
&= \hat{\pi}_0(\hat{\boldsymbol{\mu}}_S^{(0)})^T((\hat{\Sigma}_{S,S}^{(0)})^{-1} - (\hat{\Sigma}_{S,S}^{(1)})^{-1})\hat{\boldsymbol{\mu}}_S^{(0)} + \hat{\pi}_0\mathrm{Tr}\left[((\hat{\Sigma}_{S,S}^{(0)})^{-1} - (\hat{\Sigma}_{S,S}^{(1)})^{-1})\hat{\Sigma}_{S,S}^{(0)}\right].
\end{aligned}
$$

Similarly we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i = 1)\left[\boldsymbol{x}_{i,S}^T((\hat{\Sigma}_{S,S}^{(1)})^{-1} - (\hat{\Sigma}_{S,S}^{(0)})^{-1})\boldsymbol{x}_{i,S}\right] \\
&= \hat{\pi}_1(\hat{\boldsymbol{\mu}}_S^{(1)})^T((\hat{\Sigma}_{S,S}^{(1)})^{-1} - (\hat{\Sigma}_{S,S}^{(0)})^{-1})\hat{\boldsymbol{\mu}}_S^{(1)} + \hat{\pi}_1\mathrm{Tr}\left[((\hat{\Sigma}_{S,S}^{(1)})^{-1} - (\hat{\Sigma}_{S,S}^{(0)})^{-1})\hat{\Sigma}_{S,S}^{(1)}\right].
\end{aligned}
$$

Combining with the fact

$$
\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i = r)\boldsymbol{x}_{i,S} = \hat{\pi}_r\boldsymbol{\mu}_S^{(r)}, r = 0, 1,
$$

we obtain that

$$
\begin{aligned}
\mathrm{RIC}_n(S) = &-(\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)})^T[\hat{\pi}_1(\hat{\Sigma}_{S,S}^{(0)})^{-1} + \hat{\pi}_0(\hat{\Sigma}_{S,S}^{(1)})^{-1}](\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)}) \\
&+ \mathrm{Tr}[((\hat{\Sigma}_{S,S}^{(1)})^{-1} - (\hat{\Sigma}_{S,S}^{(0)})^{-1})(\hat{\pi}_1\hat{\Sigma}_{S,S}^{(1)} - \hat{\pi}_0\hat{\Sigma}_{S,S}^{(0)})] + (\hat{\pi}_1 - \hat{\pi}_0)(\log|\hat{\Sigma}_{S,S}^{(1)}| - \log|\hat{\Sigma}_{S,S}^{(0)}|) \\
&+ c_n \cdot (|S|(|S|+3)/2 + 1),
\end{aligned}
$$

which completes the proof.

B.5 Proof of Theorem 10

Denote $g_n(\alpha') = \pi_1 g_n^{(1)}(\alpha') - \pi_0 g_n^{(0)}(\alpha')$. By the definition of $\{\alpha_i\}_{i=1}^N$, it holds that

$$g_n(\alpha) = 0, \text{when } \alpha \notin \{\alpha_i\}_{i=1}^N.$$

Recall that $\boldsymbol{x}|y=0 \sim \mathrm{P}^{(0)}, \boldsymbol{x}|y=1 \sim \mathrm{P}^{(1)}$, where $\mathrm{P}^{(0)}, \mathrm{P}^{(1)}$ are the corresponding cumulative distribution functions. We have

$$\mathbf{E}[R(C_n^{RaSE})] = \pi_0 \int_{\mathcal{X}} \mathbf{P}(\nu_n(\boldsymbol{x}) > \alpha) d\mathrm{P}^{(0)} + \pi_1 \int_{\mathcal{X}} \mathbf{P}(\nu_n(\boldsymbol{x}) \leq \alpha) d\mathrm{P}^{(1)}.$$

For given $\boldsymbol{x}$ and the corresponding $\alpha' = \mu_n(\boldsymbol{x})$, we can construct a random variable $T = \sum_{j=1}^{B_1} \mathbb{1}\left\{C_n^{S_{j*}}(\boldsymbol{x}) = 1\right\} \sim Bin(B_1, \alpha')$. Then

$$\int_{\mathcal{X}} \mathbf{P}(\nu_n(\boldsymbol{x}) \leq \alpha) d\mathrm{P}^{(1)} = \int_{[0,1]} \mathrm{P}(T \leq B_1\alpha) dG_n^{(1)}(\alpha').$$

Similarly,

$$\int_{\mathcal{X}} \mathbf{P}(\nu_n(\boldsymbol{x}) > \alpha) d\mathrm{P}^{(0)} = 1 - \int_{[0,1]} \mathrm{P}(T \leq B_1\alpha) dG_n^{(0)}(\alpha').$$

This leads to

$$\mathbf{E}[R(C_n^{RaSE})] = \pi_0 + \int_{[0,1]} \mathrm{P}(T \leq B_1\alpha) dG_n(\alpha'),$$

where $G_n(\alpha') = \pi_1 G_n^{(1)}(\alpha') - \pi_0 G_n^{(0)}(\alpha')$. And there also holds that

$$\begin{aligned}
R(C_n^{RaSE*}) &= \pi_0 \mathrm{P}^{(0)}(\mu_n(\boldsymbol{x}) > \alpha) + \pi_1 \mathrm{P}^{(1)}(\mu_n(\boldsymbol{x}) < \alpha) \\
&\quad + \frac{1}{2}[\pi_0 \mathrm{P}^{(0)}(\mu_n(\boldsymbol{x}) = \alpha)) + \pi_1 \mathrm{P}^{(1)}(\mu_n(\boldsymbol{x}) = \alpha)] \\
&= \pi_0(1 - G_n^{(0)}(\alpha)) + \pi_1 G_n^{(1)}(\alpha) + \frac{1}{2}[\pi_0 g_n^{(0)}(\alpha) - \pi_1 g_n^{(1)}(\alpha)] \\
&= \pi_0 + G_n(\alpha) - \frac{1}{2} g_n(\alpha),
\end{aligned}$$

where $g_n(\alpha') = \pi_1 g_n^{(1)}(\alpha') - \pi_0 g_n^{(0)}(\alpha')$. This implies

$$\mathbf{E}[R(C_n^{RaSE})] - R(C_n^{RaSE*}) = \int_{[0,1]} [\mathrm{P}(T \leq B_1\alpha) - \mathbb{1}_{\{\alpha' \leq \alpha\}}] dG_n(\alpha') + \frac{1}{2} g_n(\alpha). \tag{18}$$

(i) When $\alpha \notin \{\alpha_i\}_{i=1}^N$, $g_n(\alpha) = 0$ holds. For $\alpha' \in \{\alpha_i\}_{i=1}^N$, by Hoeffding's inequality (Petrov, 2012), we obtain that

$$\begin{aligned}
&|\mathrm{P}(T \leq B_1\alpha) - \mathbb{1}_{\{\alpha' \leq \alpha\}}| \\
&\leq \mathrm{P}(T - B_1\alpha' > B_1(\alpha - \alpha'))\mathbb{1}_{\{\alpha' \leq \alpha\}} + \mathrm{P}(B_1\alpha' - T \geq B_1(\alpha' - \alpha))\mathbb{1}_{\{\alpha' > \alpha\}} \\
&\leq \exp\{-2B_1(\alpha' - \alpha)^2\} \\
&\leq \exp\{-C_\alpha B_1\},
\end{aligned}$$

where $C_\alpha = \min_{1 \leq i \leq N} |\alpha - \alpha_i|^2$. This leads to the final bound $|\mathbf{E}[R(C_n^{RaSE})] - R(C_n^{RaSE*})| \leq \exp\{-C_\alpha B_1\}$.

(ii) When $\alpha = \alpha_{i_0}, i_0 \in \{1, 2, \ldots, N\}$, for $\alpha' \neq \alpha_{i_0}$, we have

$$|\mathrm{P}(T \leq B_1\alpha) - \mathbb{1}_{\{\alpha' \leq \alpha\}}| \leq \exp\{-2B_1(\alpha' - \alpha)^2\},$$

leading to

$$\sum_{i \neq i_0} |\mathrm{P}(T \leq B_1\alpha) - \mathbb{1}_{\{\alpha' \leq \alpha\}}| g_n(\alpha_i) = \exp\{-C_\alpha B_1\} \sum_{i \neq i_0} g_n(\alpha_i) \leq \exp\{-C_\alpha B_1\},$$

where $C_\alpha = 2 \min_{1 \leq i \leq N} |\alpha - \alpha_i|^2 = 2 \min_{i \neq i_0} |\alpha_{i_0} - \alpha_i|^2$. Therefore, again by (18), we have

$$|\mathbf{E}[R(C^{RaSE})] - R(C_n^{RaSE*})| \leq \left| [\mathrm{P}(T \leq B_1\alpha_{i_0}) - 1] \cdot g_n(\alpha_{i_0}) + \frac{1}{2} g_n(\alpha_{i_0}) \right| + \exp\{-C_\alpha B_1\}.$$

By Berry-Esseen theorem (Esseen, 1956),

$$\mathrm{P}(T \leq B_1\alpha_{i_0}) = \mathrm{P}\left(\frac{T - B_1\alpha_{i_0}}{\sqrt{B_1(\alpha_{i_0}(1 - \alpha_{i_0}))}} \leq 0 \right) = \frac{1}{2} + O\left(\frac{1}{\sqrt{B_1}} \right), \tag{19}$$

as $B_1 \to \infty$. Eventually there holds that

$$|\mathbf{E}[R(C^{RaSE})] - R(C_n^{RaSE*})| \leq O\left(\frac{1}{\sqrt{B_1}} \right).$$

This completes the proof.

B.6 Proof of Theorem 11

Referring to the proof of the bound on MC-variance in Cannings and Samworth (2017), it can be obtained that

$$\mathbf{Var}\left(\int_{\mathbb{R}^p} \mathbb{1}_{\{\nu_n(\boldsymbol{x}) > \alpha\}} d\mathrm{P}^{(0)} \right) \leq 2 \int_{[0,1]} \int_{[0,\alpha'']} \mathrm{P}(T' \leq B_1\alpha)\mathrm{P}(T'' > B_1\alpha) dG_n^{(0)}(\alpha') dG_n^{(0)}(\alpha''),$$

where given $\boldsymbol{x}$, $T' \sim Bin(B_1, \alpha'), T'' \sim Bin(B_1, \alpha'')$ and T', T'' are independent.

(i) For $\alpha \notin \{\alpha_i^{(0)}\}_{i=1}^{N_0} \cup \{\alpha_i^{(1)}\}_{i=1}^{N_1}$: constant $C_\alpha^{(0)} = 2 \min_{1 \leq i \leq N_0} (|\alpha - \alpha_i^{(0)}|^2) > 0, C_\alpha^{(1)} = 2 \min_{1 \leq i \leq N_1} (|\alpha - \alpha_i^{(1)}|^2) > 0$. If $\alpha \leq \alpha' \leq \alpha''$: Then by Hoeffding's inequality, we have

$$\mathrm{P}(T' \leq B_1\alpha) = \mathrm{P}(T' - B_1\alpha' \leq B_1(\alpha - \alpha')) \leq \exp\{-B_1 C_\alpha^{(0)}\}. \tag{20}$$

If $\alpha' \leq \alpha'' \leq \alpha$: Similarly we have

$$\mathrm{P}(T'' > B_1\alpha) \leq \exp\{-B_1 C_\alpha^{(0)}\}. \tag{21}$$

If $\alpha' \leq \alpha \leq \alpha''$: We have both (20) and (21).

Thus we always have

$$\mathrm{P}(T' \leq B_1\alpha)\mathrm{P}(T'' > B_1\alpha) \leq \exp\{-B_1 C_\alpha^{(0)}\}.$$

Since here the integration actually is the finite summation, it implies that

$$\mathbf{Var}\left(\int_{\mathcal{X}} \mathbb{1}_{\{\nu_n(\boldsymbol{x})>\alpha\}} d\mathrm{P}^{(0)}\right) \leq \exp\{-B_1 C_\alpha^{(0)}\}.$$

Since $\alpha \notin \{\alpha_i^{(1)}\}_{i=1}^{N_1}$: We can obtain the similar conclusion that

$$\mathbf{Var}\left(\int_{\mathcal{X}} \mathbb{1}_{\{\nu_n(\boldsymbol{x})\leq\alpha\}} d\mathrm{P}^{(1)}\right) \leq \exp\{-B_1 C_\alpha^{(1)}\}.$$

Thus for any $\alpha \notin \{\alpha_i^{(0)}\}_{i=1}^{N_0} \cup \{\alpha_i^{(1)}\}_{i=1}^{N_1}$, setting $C_\alpha = 2 \min\limits_{\substack{1\leq i\leq N_0 \\ 1\leq j\leq N_1}} (|\alpha-\alpha_i^{(0)}|^2, |\alpha-\alpha_j^{(1)}|^2) = \min(C_\alpha^{(0)}, C_\alpha^{(1)}) > 0$, then by the convexity, we have

$$\begin{aligned}
\mathbf{Var}\left(R(C_n^{RaSE})\right) &= \mathbf{Var}\left(\pi_0 \int_{\mathcal{X}} \mathbb{1}_{\{\nu_n(\boldsymbol{x})>\alpha\}} d\mathrm{P}^{(0)} + \pi_1 \int_{\mathcal{X}} \mathbb{1}_{\{\nu_n(\boldsymbol{x})\leq\alpha\}} d\mathrm{P}^{(1)}\right) \\
&\leq \pi_0 \mathbf{Var}\left(\int_{\mathcal{X}} \mathbb{1}_{\{\nu_n(\boldsymbol{x})>\alpha\}} d\mathrm{P}^{(0)}\right) + \pi_1 \mathbf{Var}\left(\int_{\mathcal{X}} \mathbb{1}_{\{\nu_n(\boldsymbol{x})\leq\alpha\}} d\mathrm{P}^{(1)}\right) \\
&\leq \exp\{-B_1 C_\alpha\}.
\end{aligned}$$

(ii) For $\alpha = \alpha_{i_0}^{(0)}$ or $\alpha_{i_1}^{(1)}$: Without loss of generality, suppose $\alpha = \alpha_{i_0}^{(0)}$. When $\alpha' < \alpha''$, similar to (i), there exists positive number C'_α such that

$$\mathrm{P}(T' \leq B_1\alpha)\mathrm{P}(T'' > B_1\alpha) \leq \exp\{-B_1 C'_\alpha\}.$$

When $\alpha' = \alpha'' = \alpha_{i_0}^{(0)}$, similar to (19), there holds

$$\mathrm{P}(T' \leq B_1\alpha)\mathrm{P}(T'' > B_1\alpha) = \frac{1}{4} + O\left(\frac{1}{\sqrt{B_1}}\right).$$

Thus it holds

$$\mathbf{Var}\left(\int_{\mathcal{X}} \mathbb{1}_{\{\nu_n(\boldsymbol{x})>\alpha\}} dP^{(0)}\right) \leq \frac{1}{2}(g_n^{(0)}(\alpha_{i_0}^{(0)}))^2 + O\left(\frac{1}{\sqrt{B_1}}\right).$$

And similar results hold for $\mathbf{Var}\left(\int_{\mathcal{X}} \mathbb{1}_{\{\nu_n(\boldsymbol{x})\leq\alpha\}} dP^{(1)}\right)$. Eventually we would have

$$\begin{aligned}
\mathbf{Var}\left(R(C_n^{RaSE})\right) &\leq \pi_0 \mathbf{Var}\left(\int_{\mathcal{X}} \mathbb{1}_{\{\nu_n(\boldsymbol{x})>\alpha\}} d\mathrm{P}^{(0)}\right) + \pi_1 \mathbf{Var}\left(\int_{\mathcal{X}} \mathbb{1}_{\{\nu_n(\boldsymbol{x})\leq\alpha\}} d\mathrm{P}^{(1)}\right) \\
&\leq \frac{1}{2}\left[\pi_0 (g_n^{(0)}(\alpha))^2 + \pi_1 (g_n^{(1)}(\alpha))^2\right] + O\left(\frac{1}{\sqrt{B_1}}\right),
\end{aligned}$$

which completes the proof.

B.7 Proof of Proposition 12

(i) Due to Proposition 3, this is trivial to be seen to hold here.

(ii) First let's consider subspace $S = S' \cup \{j\}$, where $|S'| \geq 1, j \notin S'$. The conditional densities of $j|S'$ are denoted as $f^{(0)}_{j|S'}, f^{(1)}_{j|S'}$ and the components of $\boldsymbol{x}_S$ corresponding to S' and $\{j\}$ are $\boldsymbol{x}_{S'}, \boldsymbol{x}_j$, respectively. The definition of KL divergence and Fubini theorem incur

$$\begin{aligned}
\mathrm{KL}(f^{(0)}_S||f^{(1)}_S) &= \mathrm{E}_{\boldsymbol{x}_S\sim f^{(0)}_S}\left[\log\left(\frac{f^{(0)}_{S'}(\boldsymbol{x}_{S'})}{f^{(1)}_{S'}(\boldsymbol{x}_{S'})}\right) + \log\left(\frac{f^{(0)}_{j|S'}(\boldsymbol{x}_j)}{f^{(1)}_{j|S'}(\boldsymbol{x}_j)}\right)\right] \\
&= \mathrm{E}_{\boldsymbol{x}_{S'}\sim f^{(0)}_{S'}}\left[\log\left(\frac{f^{(0)}_{S'}(\boldsymbol{x}_{S'})}{f^{(1)}_{S'}(\boldsymbol{x}_{S'})}\right)\right] + \mathrm{E}_{\boldsymbol{x}_{S'}\sim f^{(0)}_{S'}}\left[\mathrm{E}_{\boldsymbol{x}_j\sim f^{(0)}_{j|S'}}\log\left(\frac{f^{(0)}_{j|S'}(\boldsymbol{x}_j)}{f^{(1)}_{j|S'}(\boldsymbol{x}_j)}\right)\right] \\
&= \mathrm{KL}(f^{(0)}_{S'}||f^{(1)}_{S'}) + \mathrm{E}_{\boldsymbol{x}_{S'}\sim f^{(0)}_{S'}}\left[\mathrm{KL}(f^{(0)}_{j|S'}||f^{(1)}_{j|S'})\right] \\
&\geq \mathrm{KL}(f^{(0)}_{S'}||f^{(1)}_{S'}).
\end{aligned}$$

Here "=" holds if and only if $f^{(0)}_{j|S'}(\boldsymbol{x}_j|\boldsymbol{x}_{S'}) = f^{(1)}_{j|S'}(\boldsymbol{x}_j|\boldsymbol{x}_{S'})$ *a.s.* with respect to $\mathrm{P}^{(0)}$. By induction, this indicates that for any $S \supseteq S'$ and $S' \neq \emptyset$, there holds

$$\mathrm{KL}(f^{(0)}_S||f^{(1)}_S) \geq \mathrm{KL}(f^{(0)}_{S'}||f^{(1)}_{S'}). \tag{22}$$

Note that in (22), by Proposition 3, if $f^{(0)}_{j|S'}(\boldsymbol{x}_j|\boldsymbol{x}_{S'}) = f^{(1)}_{j|S'}(\boldsymbol{x}_j|\boldsymbol{x}_{S'})$ *a.s.* with respect to $\mathrm{P}^{(0)}$, we have

$$\frac{f^{(0)}_S(\boldsymbol{x}_S)}{f^{(1)}_S(\boldsymbol{x}_S)} = \frac{f^{(0)}_{S'}(\boldsymbol{x}_{S'})}{f^{(1)}_{S'}(\boldsymbol{x}_{S'})},\ a.s.\ , w.r.t.\ \mathrm{P}^{(0)}. \tag{23}$$

Similarly, we have

$$\mathrm{KL}(f^{(1)}_S||f^{(0)}_S) \geq \mathrm{KL}(f^{(1)}_{S'}||f^{(0)}_{S'}), \tag{24}$$

where the " = " holds if and only if

$$\frac{f^{(1)}_S(\boldsymbol{x}_S)}{f^{(0)}_S(\boldsymbol{x}_S)} = \frac{f^{(1)}_{S'}(\boldsymbol{x}_{S'})}{f^{(0)}_{S'}(\boldsymbol{x}_{S'})},\ a.s.\ , w.r.t.\ \mathrm{P}^{(1)}. \tag{25}$$

If $S \not\supseteq S^*$, consider $\bar{S} = S \cup (S^* \backslash S) \supseteq S^*$, then by (i) and (22), there holds

$$\begin{aligned}
\pi_0\mathrm{KL}(f^{(0)}_S||f^{(1)}_S) + \pi_1\mathrm{KL}(f^{(1)}_S||f^{(0)}_S) &\leq \pi_0\mathrm{KL}(f^{(0)}_{\bar{S}}||f^{(1)}_{\bar{S}}) + \pi_1\mathrm{KL}(f^{(1)}_{\bar{S}}||f^{(0)}_{\bar{S}}) \\
&= \pi_0\mathrm{KL}(f^{(0)}_{S^*}||f^{(1)}_{S^*}) + \pi_1\mathrm{KL}(f^{(1)}_{S^*}||f^{(0)}_{S^*}).
\end{aligned} \tag{26}$$

Then by Proposition 3, if the "=" holds in (26), due to (23) and (25), we have

$$\frac{f^{(0)}_S(\boldsymbol{x}_S)}{f^{(1)}_S(\boldsymbol{x}_S)} = \frac{f^{(0)}_{\bar{S}}(\boldsymbol{x}_{\bar{S}})}{f^{(1)}_{\bar{S}}(\boldsymbol{x}_{\bar{S}})} = \frac{f^{(0)}_{S^*}(\boldsymbol{x}_{S^*})}{f^{(1)}_{S^*}(\boldsymbol{x}_{S^*})},\ a.s.\ , w.r.t.\ \mathrm{P}^{\boldsymbol{x}} = \pi_0\mathrm{P}^{(0)} + \pi_1\mathrm{P}^{(1)},$$

implying that S is a discriminative set but $S \not\supseteq S^*$, which yields a contradiction. Therefore (ii) holds here.

(iii) holds since the full model $S_{\text{Full}} \supseteq S^*$ and the KL divergence is monotone in the sense of (22) and (24).

B.8 Proof of Proposition 14

Denote $R_S^{0|1}(\boldsymbol{\mu}_S^{(0)}, \boldsymbol{\mu}_S^{(1)}, \Sigma_{S,S}, \boldsymbol{x}_S) = \log\left(\frac{f_S^{(0)}(\boldsymbol{x}_S)}{f_S^{(1)}(\boldsymbol{x}_S)}\right) = \left(\boldsymbol{\mu}_S^{(0)} - \boldsymbol{\mu}_S^{(1)}\right)^T \Sigma_{S,S}^{-1}\left[\boldsymbol{x}_S - \frac{1}{2}(\boldsymbol{\mu}_S^{(0)} + \boldsymbol{\mu}_S^{(1)})\right]$. To eliminate any confusion and better illustrate the meaning of the gradient and second-order derivative matrix, we will use $\frac{\partial R_S^{0|1}}{\partial \boldsymbol{\mu}_S^{(0)}}, \frac{\partial R_S^{0|1}}{\partial \boldsymbol{\mu}_S^{(1)}}, \frac{\partial R_S^{0|1}}{\partial \Sigma_{S,S}}$ to represent the gradient and use $\frac{\partial^2 R_S^{0|1}}{\partial \boldsymbol{\mu}_S^{(0)} \partial \Sigma_{S,S}}, \frac{\partial^2 R_S^{0|1}}{\partial \boldsymbol{\mu}_S^{(1)} \partial \Sigma_{S,S}}, \frac{\partial^2 R_S^{0|1}}{\partial \Sigma_{S,S} \partial \Sigma_{S,S}}$ to represent the second-order derivative matrix.

According to Brookes (2005) and Petersen and Pedersen (2012), with some calculation we can obtain that

$$
\begin{aligned}
\frac{\partial R_S^{0|1}}{\partial \boldsymbol{\mu}_S^{(0)}} &= \Sigma_{S,S}^{-1}\left(\boldsymbol{x}_S - \boldsymbol{\mu}_S^{(0)}\right), \\
\frac{\partial R_S^{0|1}}{\partial \boldsymbol{\mu}_S^{(1)}} &= \Sigma_{S,S}^{-1}\left(\boldsymbol{x}_S - \boldsymbol{\mu}_S^{(1)}\right), \\
\frac{\partial R_S^{0|1}}{\partial \Sigma_{S,S}} &= \Sigma_{S,S}^{-1}\left(\boldsymbol{\mu}_S^{(0)} - \boldsymbol{\mu}_S^{(1)}\right)\left(\boldsymbol{x}_S - \frac{1}{2}(\boldsymbol{\mu}_S^{(0)} + \boldsymbol{\mu}_S^{(1)})\right)^T \Sigma_{S,S}^{-1}, \\
\frac{\partial^2 R_S^{0|1}}{\partial \boldsymbol{\mu}_S^{(0)} \partial \boldsymbol{\mu}_S^{(0)}} &= \frac{\partial^2 R_S^{0|1}}{\partial \boldsymbol{\mu}_S^{(1)} \partial \boldsymbol{\mu}_S^{(1)}} = -\Sigma_{S,S}^{-1}, \\
\frac{\partial^2 R_S^{0|1}}{\partial \Sigma_{S,S} \partial \boldsymbol{\mu}_S^{(0)}} &= -\left(\Sigma_{S,S}^{-1} \otimes \Sigma_{S,S}^{-1}\right)\left(I_{|S|} \otimes \left(\boldsymbol{x}_S - \boldsymbol{\mu}_S^{(0)}\right)\right), \\
\frac{\partial^2 R_S^{0|1}}{\partial \Sigma_{S,S} \partial \boldsymbol{\mu}_S^{(1)}} &= -\left(\Sigma_{S,S}^{-1} \otimes \Sigma_{S,S}^{-1}\right)\left(I_{|S|} \otimes \left(\boldsymbol{x}_S - \boldsymbol{\mu}_S^{(1)}\right)\right), \\
\frac{\partial^2 R_S^{0|1}}{\partial \Sigma_{S,S} \partial \Sigma_{S,S}} &= -\left[I_{|S|} \otimes \Sigma_{S,S}^{-1}\left(\boldsymbol{\mu}_S^{(0)} - \boldsymbol{\mu}_S^{(1)}\right)\left(\boldsymbol{x}_S - \frac{1}{2}(\boldsymbol{\mu}_S^{(0)} + \boldsymbol{\mu}_S^{(1)})\right)^T \right. \\
&\quad \left. + \left(\boldsymbol{x}_S - \frac{1}{2}(\boldsymbol{\mu}_S^{(0)} + \boldsymbol{\mu}_S^{(1)})\right)\left(\boldsymbol{\mu}_S^{(0)} - \boldsymbol{\mu}_S^{(1)}\right)^T \Sigma_{S,S}^{-1} \otimes I_{|S|}\right]\left(\Sigma_{S,S}^{-1} \otimes \Sigma_{S,S}^{-1}\right),
\end{aligned}
$$

where $\otimes$ is the Kronecker product.

Then let's check conditions in Assumption 3 one by one. Without loss of generality, for (iii), we only check the case that $\boldsymbol{x}_{i,S} \overset{i.i.d.}{\sim} f_S^{(0)}$. And for $(iv), (v)$, we only check the case that $\boldsymbol{x}_S \sim f_S^{(0)}$.

(i) It's easy to see that the number of parameters in LDA model in p-dimensional space is $2 + 2p + \frac{p(p+1)}{2}$, therefore $\kappa_1 = 2$.

(ii) According to Assumption 4, we have

$$
\mathrm{KL}(f^{(0)} || f^{(1)}) = \left(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)}\right)^T \Sigma^{-1}\left(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)}\right)
$$

$$\begin{aligned}
&\leq \|\boldsymbol{\mu}_{S^*}^{(1)} - \boldsymbol{\mu}_{S^*}^{(0)}\|_2^2 \cdot \|\Sigma_{S^*,S^*}^{-1}\|_2 \\
&\leq p^*(M')^2 m^{-1} \\
&\leq D(M')^2 m^{-1}.
\end{aligned}$$

The similar conclusion holds for $\mathrm{KL}(f^{(1)}||f^{(0)})$ as well.

(iii) For any $(\tilde{\boldsymbol{\mu}}_S^{(0)}, \tilde{\boldsymbol{\mu}}_S^{(1)}, \tilde{\Sigma}_{S,S})$:

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 R_S^{0|1}}{\partial \Sigma_{S,S} \partial \Sigma_{S,S}} (\tilde{\boldsymbol{\mu}}_S^{(0)}, \tilde{\boldsymbol{\mu}}_S^{(1)}, \tilde{\Sigma}_{S,S}, \boldsymbol{x}_i) \right\|_{\max} \\
&\leq \left\| \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 R_S^{0|1}}{\partial \Sigma_{S,S} \partial \Sigma_{S,S}} (\tilde{\boldsymbol{\mu}}_S^{(0)}, \tilde{\boldsymbol{\mu}}_S^{(1)}, \tilde{\Sigma}_{S,S}, \boldsymbol{x}_i) \right\|_2 \\
&\leq 2 \left\| \tilde{\Sigma}_{S,S}^{-1} (\tilde{\boldsymbol{\mu}}_S^{(0)} - \tilde{\boldsymbol{\mu}}_S^{(1)}) \right\|_2 \cdot \left\| \frac{1}{n} \sum_{i=1}^n \left[\boldsymbol{x}_{i,S} - \frac{1}{2} (\tilde{\boldsymbol{\mu}}_S^{(0)} + \tilde{\boldsymbol{\mu}}_S^{(1)}) \right] \right\|_2 \cdot \left\| \tilde{\Sigma}_{S,S}^{-1} \right\|_2^2 \\
&\leq 2 \left\| \tilde{\Sigma}_{S,S}^{-1} \right\|_2^3 \cdot \left\| \tilde{\boldsymbol{\mu}}_S^{(0)} - \tilde{\boldsymbol{\mu}}_S^{(1)} \right\|_2 \cdot \left\| \frac{1}{n} \sum_{i=1}^n \left[\boldsymbol{x}_{i,S} - \frac{1}{2} (\tilde{\boldsymbol{\mu}}_S^{(0)} + \tilde{\boldsymbol{\mu}}_S^{(1)}) \right] \right\|_2.
\end{aligned}$$

When $\|\tilde{\boldsymbol{\mu}}_S^{(0)} - \boldsymbol{\mu}_S^{(0)}\|_2, \|\tilde{\boldsymbol{\mu}}_S^{(1)} - \boldsymbol{\mu}_S^{(1)}\|_2, \|\tilde{\Sigma}_{S,S} - \Sigma_{S,S}\|_F \leq \zeta < m$, due to (28), it follows that

$$\|\tilde{\Sigma}_{S,S}^{-1} - \Sigma_{S,S}^{-1}\|_2 \leq \frac{\frac{1}{m^2}\|\tilde{\Sigma}_{S,S} - \Sigma_{S,S}\|_2}{1 - \frac{1}{m}\|\tilde{\Sigma}_{S,S} - \Sigma_{S,S}\|_2} \leq \frac{\frac{1}{m^2}\|\tilde{\Sigma}_{S,S} - \Sigma_{S,S}\|_F}{1 - \frac{1}{m}\|\tilde{\Sigma}_{S,S} - \Sigma_{S,S}\|_F} \leq \frac{\zeta}{m^2 - m\zeta},$$

which leads to

$$\|\tilde{\Sigma}_{S,S}^{-1}\|_2 \leq \|\Sigma_{S,S}^{-1}\|_2 + \|\tilde{\Sigma}_{S,S}^{-1} - \Sigma_{S,S}^{-1}\|_2 \leq \frac{1}{m - \zeta}.$$

In addition, we have

$$\|\tilde{\boldsymbol{\mu}}_S^{(0)} - \tilde{\boldsymbol{\mu}}_S^{(1)}\|_2 \leq \|\tilde{\boldsymbol{\mu}}_S^{(0)} - \boldsymbol{\mu}_S^{(0)}\|_2 + \|\tilde{\boldsymbol{\mu}}_S^{(1)} - \boldsymbol{\mu}_S^{(1)}\|_2 + \|\boldsymbol{\mu}_S^{(0)} - \boldsymbol{\mu}_S^{(1)}\|_2 \lesssim D^{\frac{1}{2}} M'.$$

Without loss of generality, consider $\boldsymbol{x}_{i,S} \overset{i.i.d.}{\sim} f_S^{(0)}$, it holds that

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^n \left[\boldsymbol{x}_{i,S} - \frac{1}{2} (\tilde{\boldsymbol{\mu}}_S^{(0)} + \tilde{\boldsymbol{\mu}}_S^{(1)}) \right] \right\|_2 &\leq \left\| \frac{1}{n} \sum_{i=1}^n \left(\boldsymbol{x}_{i,S} - \boldsymbol{\mu}_S^{(0)} \right) \right\|_2 + \frac{1}{2} \|\tilde{\boldsymbol{\mu}}_S^{(0)} - \boldsymbol{\mu}_S^{(0)}\|_2 \\
&\quad + \frac{1}{2} \|\tilde{\boldsymbol{\mu}}_S^{(1)} - \boldsymbol{\mu}_S^{(1)}\|_2 + \frac{1}{2} \|\boldsymbol{\mu}_S^{(0)} - \boldsymbol{\mu}_S^{(1)}\|_2 \\
&\lesssim \left\| \frac{1}{n} \sum_{i=1}^n \left(\boldsymbol{x}_{i,S} - \boldsymbol{\mu}_S^{(0)} \right) \right\|_2 + D^{\frac{1}{2}} M'.
\end{aligned}$$

By Proposition 1 in Hsu et al. (2012), since $\frac{1}{n} \sum_{i=1}^n (\boldsymbol{x}_{i,S} - \boldsymbol{\mu}_S^{(0)}) \sim N(\mathbf{0}, \frac{1}{n} \Sigma_{S,S})$, it follows that

$$\mathbb{P} \left(\left\| \frac{1}{n} \sum_{i=1}^n \left(\boldsymbol{x}_{i,S} - \boldsymbol{\mu}_S^{(0)} \right) \right\|_2 > \sqrt{\frac{1}{n} \mathrm{Tr}(\Sigma_{S,S}) + 2 \sqrt{\mathrm{Tr}(\Sigma_{S,S}^2) \cdot \frac{\epsilon}{n}} + 2 \|\Sigma_{S,S}\|_2 \epsilon} \right) \leq \exp\{-n\epsilon\}.$$

And due to Assumption 4, we have

$$\begin{aligned}
&\mathrm{Tr}(\Sigma_{S,S}) \le D\|\Sigma_{S,S}\|_{\max} \le DM,\\
&\mathrm{Tr}(\Sigma_{S,S}^2) = \|\Sigma_{S,S}\|_F^2 \le D^2\|\Sigma_{S,S}\|_{\max}^2 \le D^2M^2,\\
&\|\Sigma_{S,S}\|_2 \le D\|\Sigma_{S,S}\|_{\max} \le DM,
\end{aligned}$$

yielding

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{x}_{i,S}-\boldsymbol{\mu}_S^{(0)}\right)\right\|_2 > \sqrt{\frac{1}{n}DM+2DM\sqrt{\frac{\epsilon}{n}}+2DM\epsilon}\right) \le \exp\{-n\epsilon\}.$$

Therefore when $\|\tilde{\boldsymbol{\mu}}_S^{(0)}-\boldsymbol{\mu}_S^{(0)}\|_2, \|\tilde{\boldsymbol{\mu}}_S^{(1)}-\boldsymbol{\mu}_S^{(1)}\|_2, \|\tilde{\Sigma}_{S,S}-\Sigma_{S,S}\|_2 \le \zeta$, we have

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 R_S^{0|1}}{\partial\Sigma_{S,S}\partial\Sigma_{S,S}}(\tilde{\boldsymbol{\mu}}_S^{(0)},\tilde{\boldsymbol{\mu}}_S^{(1)},\tilde{\Sigma}_{S,S},\boldsymbol{x}_i)\right\|_{\max} &\lesssim D^{\frac{1}{2}}M\left(\left\|\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{x}_{i,S}-\boldsymbol{\mu}_S^{(0)}\right)\right\|_2 + D^{\frac{1}{2}}M\right)\\
&:= V_S(\{\boldsymbol{x}_{i,S}\}_{i=1}^n),
\end{aligned}$$

and

$$\mathbb{P}\left(V_S(\{\boldsymbol{x}_{i,S}\}_{i=1}^n) > CD\right) \lesssim \exp\{-Cn\}.$$

Thus we can set $\kappa_2 = 1$.

(iv) $\Sigma_{S,S}^{-1}\left(\boldsymbol{x}_S-\boldsymbol{\mu}_S^{(0)}\right) \sim N(\boldsymbol{0},\Sigma_{S,S}^{-1}), \Sigma_{S,S}^{-1}\left(\boldsymbol{x}_S-\boldsymbol{\mu}_S^{(1)}\right) \sim N(\Sigma_{S,S}^{-1}(\boldsymbol{\mu}_S^{(0)}-\boldsymbol{\mu}_S^{(1)}),\Sigma_{S,S}^{-1})$ when $\boldsymbol{x}_S \sim f_S^{(0)}$. And also we have

$$\left\|\Sigma_{S,S}^{-1}\right\|_{\max} \le \left\|\Sigma_{S,S}^{-1}\right\|_2 \le \left\|\Sigma^{-1}\right\|_2 \le m^{-1}.$$

Then each component of $\frac{\partial R_S^{0|1}}{\partial\boldsymbol{\mu}_S^{(0)}}$ and $\frac{\partial R_S^{0|1}}{\partial\boldsymbol{\mu}_S^{(1)}}$ is $\sqrt{m^{-1}}$-subGaussian. On the other hand, since $\Sigma_{S,S}^{-1}\left(\boldsymbol{x}_S-\frac{1}{2}(\boldsymbol{\mu}_S^{(0)}+\boldsymbol{\mu}_S^{(1)})\right) \sim N\left(\frac{1}{2}\Sigma_{S,S}^{-1}(\boldsymbol{\mu}_S^{(0)}-\boldsymbol{\mu}_S^{(1)}),\Sigma_{S,S}^{-1}\right)$ and

$$\left\|\Sigma_{S,S}^{-1}(\boldsymbol{\mu}_S^{(0)}-\boldsymbol{\mu}_S^{(1)})\right\|_\infty \le D^{\frac{1}{2}}\left\|\Sigma_{S,S}^{-1}\right\|_2\cdot\left\|\boldsymbol{\mu}_S^{(0)}-\boldsymbol{\mu}_S^{(1)}\right\|_\infty \le m^{-1}M'D^{\frac{1}{2}}, \tag{27}$$

it follows that each component of $\frac{\partial R_S^{0|1}}{\partial\Sigma_{S,S}}$ is $m^{-1}\sqrt{M'D}$-subGaussian. Therefore $\kappa_3 = \frac{1}{2}$.

(v) Notice that because of (27), we have

$$\left\|\mathbb{E}_{\boldsymbol{x}_S\sim f_S^{(0)}}\left[\Sigma_{S,S}^{-1}(\boldsymbol{x}_S-\boldsymbol{\mu}_S^{(1)})\right]\right\|_\infty = \left\|\Sigma_{S,S}^{-1}(\boldsymbol{\mu}_S^{(0)}-\boldsymbol{\mu}_S^{(1)})\right\|_\infty \le m^{-1}M'D^{\frac{1}{2}},$$

$$\left\|\mathbb{E}_{\boldsymbol{x}_S\sim f_S^{(0)}}\left[\Sigma_{S,S}^{-1}\left(\boldsymbol{\mu}_S^{(0)}-\boldsymbol{\mu}_S^{(1)}\right)\left(\boldsymbol{x}_S-\frac{1}{2}(\boldsymbol{\mu}_S^{(0)}+\boldsymbol{\mu}_S^{(1)})\right)^T\Sigma_{S,S}^{-1}\right]\right\|_{\max}$$

$$
\begin{aligned}
&= \frac{1}{2}\left\|\Sigma_{S,S}^{-1}\left(\boldsymbol{\mu}_S^{(0)}-\boldsymbol{\mu}_S^{(1)}\right)\left(\boldsymbol{\mu}_S^{(0)}-\boldsymbol{\mu}_S^{(1)}\right)^T\Sigma_{S,S}^{-1}\right\|_{\max}\\
&\le \frac{1}{2}\left\|\Sigma_{S,S}^{-1}(\boldsymbol{\mu}_S^{(0)}-\boldsymbol{\mu}_S^{(1)})\right\|_\infty^2\\
&\le \frac{1}{2}(m^{-1}M')^2D.
\end{aligned}
$$

Therefore $\kappa_4 = 1$.

(vi) It's easy to see that $\left(\boldsymbol{\mu}_S^{(1)}-\boldsymbol{\mu}_S^{(0)}\right)^T\Sigma_{S,S}^{-1}\left[\boldsymbol{x}_S-\frac{1}{2}(\boldsymbol{\mu}_S^{(0)}+\boldsymbol{\mu}_S^{(1)})\right] \sim N((\boldsymbol{\mu}_S^{(1)}-\boldsymbol{\mu}_S^{(0)})^T\Sigma_{S,S}^{-1}(\boldsymbol{\mu}_S^{(1)}-\boldsymbol{\mu}_S^{(0)}), (\boldsymbol{\mu}_S^{(1)}-\boldsymbol{\mu}_S^{(0)})^T\Sigma_{S,S}^{-1}(\boldsymbol{\mu}_S^{(1)}-\boldsymbol{\mu}_S^{(0)}))$. And there holds

$$
(\boldsymbol{\mu}_S^{(1)}-\boldsymbol{\mu}_S^{(0)})^T\Sigma_{S,S}^{-1}(\boldsymbol{\mu}_S^{(1)}-\boldsymbol{\mu}_S^{(0)}) \le \left\|\boldsymbol{\mu}_S^{(1)}-\boldsymbol{\mu}_S^{(0)}\right\|_2^2\cdot\left\|\Sigma_{S,S}^{-1}\right\|_2 \le m^{-1}(M')^2D,
$$

which yields that $\log\left(\frac{f_S^{(0)}(\boldsymbol{x}_S)}{f_S^{(1)}(\boldsymbol{x}_S)}\right)$ is $M'\sqrt{m^{-1}D}$-subGaussian. So $\kappa_5 = \frac{1}{2}$.

(vii) This can be easily derived from Lemma 28 and its proof.

(viii) This is obvious because of Lemma 31.(iii) and Assumption 4.(iii).

B.9 Proof of Theorem 15

First suppose that we have n_0 observations of class 0 $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^{n_0}$ and n_1 observations of class 1 $\{\boldsymbol{x}_i^{(1)}\}_{i=1}^{n_1}$, where $n_0+n_1=n$. Define

$$
G_{n_0,S}^{(0)}(\boldsymbol{\theta}_S') = \frac{1}{n_0}\sum_{i=1}^{n_0}\log\left[\frac{f_S^{(0)}(\boldsymbol{x}_{i,S}^{(0)}|\boldsymbol{\theta}_S')}{f_S^{(1)}(\boldsymbol{x}_{i,S}^{(0)}|\boldsymbol{\theta}_S')}\right], G_{n_1,S}^{(1)}(\boldsymbol{\theta}_S') = \frac{1}{n_1}\sum_{i=1}^{n_1}\log\left[\frac{f_S^{(1)}(\boldsymbol{x}_{i,S}^{(1)}|\boldsymbol{\theta}_S')}{f_S^{(0)}(\boldsymbol{x}_{i,S}^{(1)}|\boldsymbol{\theta}_S')}\right],
$$

for any $\boldsymbol{\theta}_S'$. Denote $\widehat{\mathrm{RIC}}(S) = -2[\hat{\pi}_0 G_{n_0,S}^{(0)}(\hat{\boldsymbol{\theta}}_S) + \hat{\pi}_1 G_{n_1,S}^{(1)}(\hat{\boldsymbol{\theta}}_S)]$, then $\mathrm{RIC}_n(S) = \widehat{\mathrm{RIC}}(S) + c_n\deg(S)$.

Lemma 27 *If Assumptions 1-3 holds, then we have*

$$
\begin{aligned}
\mathbb{P}\left(\sup_{S:|S|\le D}|\widehat{\mathrm{RIC}}(S)-\mathrm{RIC}(S)|>\epsilon\right) \lesssim\ & p^D D^{\kappa_1}\exp\left\{-Cn\left(\frac{\epsilon}{D^{\kappa_1}(D^{\kappa_3}+D^{\kappa_4})}\right)^2\right\}\\
&+ p^D D^{\kappa_1}\exp\left\{-Cn\cdot\frac{\epsilon}{D^{2\kappa_1+\kappa_2}}\right\}\\
&+ p^D\exp\left\{-Cn\left(\frac{\epsilon}{D^{\kappa_5}}\right)^2\right\}.
\end{aligned}
$$

Proof [Proof of Lemma 27] By Taylor expansion and mean-value theorem, there exists $\lambda\in(0,1)$ and $\tilde{\boldsymbol{\theta}}_S = \lambda\boldsymbol{\theta}_S + (1-\lambda)\hat{\boldsymbol{\theta}}_S$ satisfying that

$$
G_{n_0,S}^{(0)}(\hat{\boldsymbol{\theta}}_S) = G_{n_0,S}^{(0)}(\boldsymbol{\theta}_S) + \nabla G_{n_0,S}^{(0)}(\boldsymbol{\theta}_S)^T(\hat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}_S) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}_S)^T\nabla^2 G_{n_0,S}^{(0)}(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}_S).
$$

Notice that

$$\left\|\nabla G^{(0)}_{n_0,S}(\boldsymbol{\theta}_S)^T(\hat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}_S)\right\|_2 \le D^{\kappa_1}\left\|\nabla G^{(0)}_{n_0,S}(\boldsymbol{\theta}_S)\right\|_\infty \|\hat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}_S\|_\infty,$$

and when $\|\hat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}_S\|_2\le\zeta$, we also have

$$\begin{aligned}\left|(\hat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}_S)^T\nabla^2 G^{(0)}_{n_0,S}(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}_S)\right| &\lesssim D^{\kappa_1}\|\hat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}_S\|_\infty^2\|\nabla^2 G^{(0)}_{n_0,S}(\tilde{\boldsymbol{\theta}})\|_{\max}\\ &\le D^{2\kappa_1}\|\hat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}_S\|_\infty^2\cdot\left|V_S(\{\boldsymbol{x}^{(0)}_{i,S}\}_{i=1}^{n_0})\right|.\end{aligned}$$

Since $n_r\sim Bin(n,\pi_r), r=0,1$, by Hoeffding's inequality, we have

$$\mathbb{P}(|n_r-n\pi_r|>0.5n\pi_r)\lesssim\exp\{-Cn\}, r=0,1.$$

Because of Assumption 3.(iv), given n_0, each component of $n_0\nabla G^{(0)}_{n_0,S}(\boldsymbol{\theta}_S)$ is $n_0\sqrt{2C_3}D^{\kappa_3}$-subGaussian, then the tail bound in the below holds:

$$\mathbb{P}\left(\left\|\nabla G^{(0)}_{n_0,S}(\boldsymbol{\theta}_S)-\mathbb{E}_{\boldsymbol{x}_S\sim f_S^{(0)}}\nabla_{\boldsymbol{\theta}_S}L_S(\boldsymbol{x};\boldsymbol{\theta})\right\|_\infty>\epsilon\middle| n_0\right)\lesssim D^{\kappa_1}\exp\left\{-Cn_0\left(\frac{\epsilon}{D^{\kappa_3}}\right)^2\right\}.$$

Then according to all conclusions above, we have

$$\begin{aligned}&\mathbb{P}(|G^{(0)}_{n_0,S}(\hat{\boldsymbol{\theta}}_S)-\mathrm{KL}(f_S^{(0)}||f_S^{(1)})|>\epsilon)\\ &\le \mathrm{E}_{n_0}[\mathbb{P}(|G^{(0)}_{n_0,S}(\hat{\boldsymbol{\theta}}_S)-\mathrm{KL}(f_S^{(0)}||f_S^{(1)})|>\epsilon||n_0-n\pi_0|\le 0.5n\pi_0)]+\mathbb{P}(|n_0-n\pi_0|>0.5n\pi_0)\\ &\le \mathrm{E}_{n_0}\mathbb{P}\left(\left|G^{(0)}_{n_0,S}(\boldsymbol{\theta}_S)-\mathrm{KL}(f_S^{(0)}||f_S^{(1)})\right|>\epsilon/3\middle||n_0-n\pi_0|\le 0.5n\pi_0\right)+\\ &\quad+\mathrm{E}_{n_0}\mathbb{P}\left(\left\|\nabla G^{(0)}_{n_0,S}(\boldsymbol{\theta}_S)-\mathbb{E}_{\boldsymbol{x}_S\sim f_S^{(0)}}\nabla G^{(0)}_{n_0,S}(\boldsymbol{\theta}_S)\right\|_\infty>CD^{\kappa_3}\middle||n_0-n\pi_0|\le 0.5n\pi_0\right)\\ &\quad+\mathbb{P}\left(CD^{\kappa_1}(D^{\kappa_3}+D^{\kappa_4})\|\hat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}_S\|_\infty>\epsilon/3\right)+\mathbb{P}\left(0.5CD^{2\kappa_1+\kappa_2}\|\hat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}_S\|_\infty^2>\epsilon/3\right)\\ &\quad+\mathrm{E}_{n_0}\mathbb{P}\left(V_S(\{\boldsymbol{x}^{(0)}_{i,S}\}_{i=1}^{n_0})>CD^{\kappa_2}\middle||n_0-n\pi_0|\le 0.5n\pi_0\right)\\ &\quad+\mathbb{P}(|n_0-n\pi_0|>0.5n\pi_0)+\mathbb{P}(D^{\frac{1}{2}\kappa_1}\|\hat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}_S\|_\infty>\zeta)\\ &\lesssim D^{\kappa_1}\exp\left\{-Cn\left(\frac{\epsilon}{D^{\kappa_1}(D^{\kappa_3}+D^{\kappa_4})}\right)^2\right\}+D^{\kappa_1}\exp\left\{-Cn\cdot\frac{\epsilon}{D^{2\kappa_1+\kappa_2}}\right\}+\exp\left\{-Cn\left(\frac{\epsilon}{D^{\kappa_5}}\right)^2\right\},\end{aligned}$$

which yields

$$\begin{aligned}\mathbb{P}(|\hat{\pi}_0G^{(0)}_{n_0,S}(\hat{\boldsymbol{\theta}}_S)-\pi_0\mathrm{KL}(f_S^{(0)}||f_S^{(1)})|>\epsilon)\lesssim{}& D^{\kappa_1}\exp\left\{-Cn\left(\frac{\epsilon}{D^{\kappa_1}(D^{\kappa_3}+D^{\kappa_4})}\right)^2\right\}\\ &+D^{\kappa_1}\exp\left\{-Cn\cdot\frac{\epsilon}{D^{2\kappa_1+\kappa_2}}\right\}+\exp\left\{-Cn\left(\frac{\epsilon}{D^{\kappa_5}}\right)^2\right\}.\end{aligned}$$

Similarly, there holds

$$\mathbb{P}(|\hat{\pi}_1 G^{(1)}_{n_1,S}(\hat{\boldsymbol{\theta}}_S) - \pi_1 \mathrm{KL}(f^{(1)}_S || f^{(0)}_S)| > \epsilon) \lesssim D^{\kappa_1} \exp\left\{-Cn\left(\frac{\epsilon}{D^{\kappa_1}(D^{\kappa_3}+D^{\kappa_4})}\right)^2\right\} + D^{\kappa_1}\exp\left\{-Cn\cdot\frac{\epsilon}{D^{2\kappa_1+\kappa_2}}\right\} + \exp\left\{-Cn\left(\frac{\epsilon}{D^{\kappa_5}}\right)^2\right\}.$$

Since $\widehat{\mathrm{RIC}}(S) = -2[\hat{\pi}_0 G^{(0)}_{n_0,S}(\hat{\boldsymbol{\theta}}_S) + \hat{\pi}_1 G^{(1)}_{n_1,S}(\hat{\boldsymbol{\theta}}_S)]$, we obtain that

$$\mathbb{P}\left(|\widehat{\mathrm{RIC}}(S) - \mathrm{RIC}(S)| > \epsilon\right) \lesssim D^{\kappa_1} \exp\left\{-Cn\left(\frac{\epsilon}{D^{\kappa_1}(D^{\kappa_3}+D^{\kappa_4})}\right)^2\right\} + D^{\kappa_1}\exp\left\{-Cn\cdot\frac{\epsilon}{D^{2\kappa_1+\kappa_2}}\right\} + \exp\left\{-Cn\left(\frac{\epsilon}{D^{\kappa_5}}\right)^2\right\}.$$

Due to the union bound over all $\binom{p}{D} = O(p^D)$ possible subsets, we obtain the conclusion. ■

Let's now prove Theorem 15.

(i) It's easy to see that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\substack{S:S\supseteq S^*\\|S|\le D}} \mathrm{RIC}_n(S) \ge \inf_{\substack{S:S\not\supseteq S^*\\|S|\le D}} \mathrm{RIC}_n(S)\right)\\
&\le \mathbb{P}\left(\mathrm{RIC}(S^*) + c_n \sup_{S:|S|\le D}\deg(S) + \Delta/3 \ge \inf_{\substack{S:S\not\supseteq S^*\\|S|\le D}} \mathrm{RIC}(S) - \Delta/3\right)\\
&\quad + \mathbb{P}\left(\sup_{S:|S|\le D}|\widehat{\mathrm{RIC}}(S) - \mathrm{RIC}(S)| > \Delta/3\right)\\
&\le \mathbb{P}(c_n \sup_{S:|S|\le D}\deg(S) \ge \Delta/3) + \mathbb{P}\left(\sup_{S:|S|\le D}|\widehat{\mathrm{RIC}}(S) - \mathrm{RIC}(S)| > \Delta/3\right)\\
&\lesssim p^D D^{\kappa_1}\exp\left\{-Cn\left(\frac{\Delta}{D^{\kappa_1}(D^{\kappa_3}+D^{\kappa_4})}\right)^2\right\} + p^D D^{\kappa_1}\exp\left\{-Cn\left(\frac{\Delta}{D^{2\kappa_1+\kappa_2}}\right)\right\}\\
&\quad + p^D\exp\left\{-Cn\left(\frac{\Delta}{D^{\kappa_5}}\right)^2\right\}\\
&\to 0.
\end{aligned}$$

(ii) Similar to above, we have

$$
\begin{aligned}
&\mathbb{P}\left(\mathrm{RIC}_n(S^*) \geq \inf_{\substack{S:S\not\supseteq S^* \\ |S|\leq D}} \mathrm{RIC}_n(S)\right) \\
&\leq \mathbb{P}\left(\mathrm{RIC}(S^*) + c_n \deg(S^*) + \Delta/3 \geq \inf_{\substack{S:S\not\supseteq S^* \\ |S|\leq D}} \mathrm{RIC}(S) - \Delta/3\right) \\
&\quad + \mathbb{P}\left(\sup_{S:|S|\leq D} |\widehat{\mathrm{RIC}}(S) - \mathrm{RIC}(S)| > \Delta/3\right) \\
&\lesssim p^D D^{\kappa_1} \exp\left\{-Cn\left(\frac{\Delta}{D^{\kappa_1}(D^{\kappa_3}+D^{\kappa_4})}\right)^2\right\} + p^D D^{\kappa_1} \exp\left\{-Cn\left(\frac{\Delta}{D^{2\kappa_1+\kappa_2}}\right)\right\} \\
&\quad + p^D \exp\left\{-Cn\left(\frac{\Delta}{D^{\kappa_5}}\right)^2\right\} \\
&\to 0.
\end{aligned}
$$

Besides, it holds

$$
\begin{aligned}
&\mathbb{P}\left(\mathrm{RIC}_n(S^*) \geq \inf_{\substack{S:S\supseteq S^* \\ |S|\leq D}} \mathrm{RIC}_n(S)\right) \\
&\leq \mathbb{P}\left(\mathrm{RIC}(S^*) + c_n \deg(S^*) + c_n/3 \geq \mathrm{RIC}(S^*) + c_n(\deg(S^*)+1) - c_n/3\right) \\
&\quad + \mathbb{P}\left(\sup_{S:|S|\leq D} |\widehat{\mathrm{RIC}}(S) - \mathrm{RIC}(S)| > c_n/3\right) \\
&\lesssim p^D D^{\kappa_1} \exp\left\{-Cn\left(\frac{c_n}{D^{\kappa_1}(D^{\kappa_3}+D^{\kappa_4})}\right)^2\right\} + p^D D^{\kappa_1} \exp\left\{-Cn\left(\frac{c_n}{D^{2\kappa_1+\kappa_2}}\right)\right\} \\
&\quad + p^D \exp\left\{-Cn\left(\frac{c_n}{D^{\kappa_5}}\right)^2\right\} \\
&\to 0.
\end{aligned}
$$

Therefore we have

$$
\begin{aligned}
&\mathbb{P}\left(\mathrm{RIC}_n(S^*) \neq \inf_{S:|S|\leq D} \mathrm{RIC}_n(S)\right) \\
&\leq \mathbb{P}\left(\mathrm{RIC}_n(S^*) \geq \inf_{\substack{S:S\not\supseteq S^* \\ |S|\leq D}} \mathrm{RIC}_n(S)\right) \\
&\quad + \mathbb{P}\left(\mathrm{RIC}_n(S^*) \geq \inf_{\substack{S:S\supseteq S^* \\ |S|\leq D}} \mathrm{RIC}_n(S)\right)
\end{aligned}
$$

$$\lesssim p^D D^{\kappa_1} \exp\left\{-Cn\left(\frac{c_n}{D^{\kappa_1}(D^{\kappa_3}+D^{\kappa_4})}\right)^2\right\} + p^D D^{\kappa_1}\exp\left\{-Cn\left(\frac{c_n}{D^{2\kappa_1+\kappa_2}}\right)\right\}$$
$$+ p^D \exp\left\{-Cn\left(\frac{c_n}{D^{\kappa_5}}\right)^2\right\}$$
$$\to 0,$$

which is because $c_n \ll \Delta$. This completes the proof.

B.10 Proof of Theorem 18

Lemma 28 *For arbitrary $\epsilon \in (0, m^{-1})$, we have the following conclusions:*

(i) $\mathbb{P}(\|(\hat{\boldsymbol{\mu}}^{(1)} - \hat{\boldsymbol{\mu}}^{(0)}) - (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)})\|_\infty > \epsilon) \lesssim p\exp\{-Cn\epsilon^2\}, r = 0, 1;$

(ii) $\mathbb{P}\left(\sup_{\substack{S_1:|S_1|\le D\\ S_2:|S_2|\le D}} \|\hat{\Sigma}_{S_1,S_2} - \Sigma_{S_1,S_2}\|_\infty > \epsilon\right) \lesssim p^2\exp\left\{-Cn\cdot\left(\frac{\epsilon}{D}\right)^2\right\} + p\exp\left\{-Cn\cdot\frac{\epsilon}{D}\right\};$

(iii) $\mathbb{P}\left(\sup_{S:|S|\le D} \|\hat{\Sigma}_{S,S} - \Sigma_{S,S}\|_2 > \epsilon\right) \lesssim p^2\exp\left\{-Cn\cdot\left(\frac{\epsilon}{D}\right)^2\right\} + p\exp\left\{-Cn\cdot\frac{\epsilon}{D}\right\};$

(iv) $\mathbb{P}\left(\sup_{S:|S|\le D} \|\hat{\Sigma}^{-1}_{S,S} - \Sigma^{-1}_{S,S}\|_2 > \epsilon\right) \lesssim p^2\exp\left\{-Cn\cdot\left(\frac{\epsilon}{D}\right)^2\right\} + p\exp\left\{-Cn\cdot\frac{\epsilon}{D}\right\}.$

Proof [Proof of Lemma 28] For (i), because $\|\Sigma\|_{\max} \le M$, for any $j = 1, \ldots, p$ and $r = 0, 1$, $\hat{\mu}_j^{(r)} - \mu_j^{(r)}$ is a $\sqrt{M}$-subGaussian variable. By the tail bound and union bound, we have

$$\mathbb{P}(\|\hat{\boldsymbol{\mu}}^{(r)} - \boldsymbol{\mu}^{(r)}\|_\infty > \epsilon) \le \sum_{j=1}^p \mathbb{P}(|\hat{\mu}_j^{(r)} - \mu_j^{(r)}| > \epsilon) \lesssim p\exp\{-Cn\epsilon^2\}, r = 0, 1,$$

which leads to (i). Denote $\Sigma = (\sigma_{ij})_{p\times p}, \hat{\Sigma} = (\hat{\sigma}_{ij})_{p\times p}$. To show (ii), similar to Bickel et al. (2008), we have

$$\mathbb{P}\left(\max_{i,j}|\hat{\sigma}_{ij} - \sigma_{ij}| > \epsilon\right) \lesssim p^2\exp\{-Cn\epsilon^2\} + p\exp\{-Cn\epsilon\}.$$

And it yields

$$\mathbb{P}\left(\sup_{\substack{S_1:|S_1|\le D\\ S_2:|S_2|\le D}} \|\hat{\Sigma}_{S_1,S_2} - \Sigma_{S_1,S_2}\|_\infty > \epsilon\right) = \mathbb{P}\left(\sup_{\substack{S_1:|S_1|\le D\\ S_2:|S_2|\le D}} \sup_{i\in S_1}\sum_{j\in S_2}|\hat{\sigma}_{ij} - \sigma_{ij}| > \epsilon\right)$$
$$\le \mathbb{P}\left(D\cdot\max_{i,j}|\hat{\sigma}_{ij} - \sigma_{ij}| > \epsilon\right)$$
$$\lesssim p^2\exp\left\{-Cn\cdot\left(\frac{\epsilon}{D}\right)^2\right\} + p\exp\left\{-Cn\cdot\frac{\epsilon}{D}\right\}.$$

Since $\hat{\Sigma}_{S,S} - \Sigma_{S,S}$ is symmetric, we have $\|\hat{\Sigma}_{S,S} - \Sigma_{S,S}\|_2 \leq \|\hat{\Sigma}_{S,S} - \Sigma_{S,S}\|_\infty$ (Bickel et al. (2008)). For (iv), firstly because the operator norm is sub-multiplicative, we have

$$\begin{aligned}\|\hat{\Sigma}_{S,S}^{-1} - \Sigma_{S,S}^{-1}\|_2 &= \|\hat{\Sigma}_{S,S}^{-1}(\hat{\Sigma}_{S,S} - \Sigma_{S,S})\Sigma_{S,S}^{-1}\|_2 \\ &\leq \|\hat{\Sigma}_{S,S}^{-1}\|_2 \cdot \|\hat{\Sigma}_{S,S} - \Sigma_{S,S}\|_2 \cdot \|\Sigma_{S,S}^{-1}\|_2 \\ &\leq (\|\hat{\Sigma}_{S,S}^{-1} - \Sigma_{S,S}^{-1}\|_2 + \|\Sigma_{S,S}^{-1}\|_2) \cdot \|\hat{\Sigma}_{S,S} - \Sigma_{S,S}\|_2 \cdot \|\Sigma_{S,S}^{-1}\|_2,\end{aligned}$$

leading to

$$\|\hat{\Sigma}_{S,S}^{-1} - \Sigma_{S,S}^{-1}\|_2 \leq \frac{\|\Sigma_{S,S}^{-1}\|_2^2 \cdot \|\hat{\Sigma}_{S,S} - \Sigma_{S,S}\|_2}{1 - \|\hat{\Sigma}_{S,S} - \Sigma_{S,S}\|_2 \cdot \|\Sigma_{S,S}^{-1}\|_2} \leq \frac{\frac{1}{m^2}\|\hat{\Sigma}_{S,S} - \Sigma_{S,S}\|_2}{1 - \frac{1}{m}\|\hat{\Sigma}_{S,S} - \Sigma_{S,S}\|_2}. \tag{28}$$

Then we obtain that

$$\begin{aligned}\mathbb{P}\left(\sup_{S:|S|\leq D} \|\hat{\Sigma}_{S,S}^{-1} - \Sigma_{S,S}^{-1}\|_2 > \epsilon\right) &\leq \mathbb{P}\left(\|\hat{\Sigma}_{S,S} - \Sigma_{S,S}\|_2 > \frac{1}{2}m\right) \\ &\quad + \mathbb{P}\left(\frac{2}{m^2}\|\hat{\Sigma}_{S,S} - \Sigma_{S,S}\|_2 > \epsilon\right) \\ &\leq 2\mathbb{P}\left(\sup_{S:|S|\leq D} \|\hat{\Sigma}_{S,S} - \Sigma_{S,S}\|_2 > \frac{m^2}{2}\epsilon\right).\end{aligned}$$

Then by applying (iii), we get (iv) immediately. ■

Lemma 29 *Define* $\boldsymbol{\delta}_S = \Sigma_{S,S}^{-1}(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)})$ *for any subset* S. *For* $\forall \tilde{S} = S \cup S^*$, *where* $S \cap S^* = \emptyset$, *we have*

$$\boldsymbol{\delta}_{\tilde{S}} = \Sigma_{\tilde{S},\tilde{S}}^{-1}(\boldsymbol{\mu}_{\tilde{S}}^{(1)} - \boldsymbol{\mu}_{\tilde{S}}^{(0)}) = \Sigma_{\tilde{S},\tilde{S}}^{-1}\begin{pmatrix}\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)} \\ \boldsymbol{\mu}_{S^*}^{(1)} - \boldsymbol{\mu}_{S^*}^{(0)}\end{pmatrix} = \begin{pmatrix}\mathbf{0} \\ \boldsymbol{\delta}_{S^*}\end{pmatrix}.$$

Proof [Proof of Lemma 29] First, when $\tilde{S} = S_{\text{Full}}$, we know that

$$\boldsymbol{\delta}_{\tilde{S}} = \Sigma_{\tilde{S},\tilde{S}}^{-1}(\boldsymbol{\mu}_{\tilde{S}}^{(1)} - \boldsymbol{\mu}_{\tilde{S}}^{(0)}) = \Sigma_{\tilde{S},\tilde{S}}^{-1}\begin{pmatrix}\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)} \\ \boldsymbol{\mu}_{S^*}^{(1)} - \boldsymbol{\mu}_{S^*}^{(0)}\end{pmatrix} = \begin{pmatrix}\mathbf{0} \\ *\end{pmatrix}.$$

Then combined with the matrix decomposition

$$\Sigma_{\tilde{S},\tilde{S}}^{-1} = \begin{pmatrix}(\Sigma_{S,S} - \Sigma_{S,S^*}\Sigma_{S^*,S^*}^{-1}\Sigma_{S^*,S})^{-1} & -\Sigma_{S,S}^{-1}\Sigma_{S,S^*}(\Sigma_{S^*,S^*} - \Sigma_{S^*,S}\Sigma_{S,S}^{-1}\Sigma_{S,S^*})^{-1} \\ -\Sigma_{S^*,S^*}^{-1}\Sigma_{S^*,S}(\Sigma_{S,S} - \Sigma_{S,S^*}\Sigma_{S^*,S^*}^{-1}\Sigma_{S^*,S})^{-1} & (\Sigma_{S^*,S^*} - \Sigma_{S^*,S}\Sigma_{S,S}^{-1}\Sigma_{S,S^*})^{-1}\end{pmatrix},$$

it can be noticed that

$$(\Sigma_{S,S} - \Sigma_{S,S^*}\Sigma_{S^*,S^*}^{-1}\Sigma_{S^*,S})^{-1}(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}) = \Sigma_{S,S}^{-1}\Sigma_{S,S^*}(\Sigma_{S^*,S^*} - \Sigma_{S^*,S}\Sigma_{S,S}^{-1}\Sigma_{S,S^*})^{-1}(\boldsymbol{\mu}_{S^*}^{(1)} - \boldsymbol{\mu}_{S^*}^{(0)}).$$

Also since there holds that

$$(\Sigma_{S^*,S^*}-\Sigma_{S^*,S}\Sigma_{S,S}^{-1}\Sigma_{S,S^*})^{-1} = \Sigma_{S^*,S^*}^{-1}-\Sigma_{S^*,S^*}^{-1}\Sigma_{S^*,S}(-\Sigma_{S,S}+\Sigma_{S,S^*}\Sigma_{S^*,S^*}^{-1}\Sigma_{S^*,S})^{-1}\Sigma_{S,S^*}\Sigma_{S^*,S^*}^{-1},$$

it can be easily verified that the "*" part is actually $\boldsymbol{\delta}_{S^*}$. Therefore the conclusion holds with $\tilde{S} = S_{\text{Full}}$.

For general $\tilde{S} \supseteq S^*$, notice that $S_{\text{Full}} = (S_{\text{Full}} \backslash \tilde{S}) \cup \tilde{S}$, with the same procedure, we can obtain that

$$\boldsymbol{\delta}_{S_{\text{Full}}} = \Sigma^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)}) = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\delta}_{\tilde{S}} \end{pmatrix}.$$

And since

$$\boldsymbol{\delta}_{S_{\text{Full}}} = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\delta}_{S^*} \end{pmatrix},$$

we reach the conclusion. ■

Lemma 30 *For $\forall S$ which satisfies $S \not\supseteq S^*$, let $\tilde{S}^* = S^* \backslash S$, then*

$$\left\| \Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\left(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}\right) - \left(\boldsymbol{\mu}_{S^*}^{(1)} - \boldsymbol{\mu}_{S^*}^{(0)}\right) \right\|_2 \geq \gamma m.$$

Proof [Proof of Lemma 30] Let $\tilde{S} = S \cup \tilde{S}^*$, due to Lemma 29, it's easy to see that there holds

$$\Sigma_{\tilde{S},\tilde{S}}^{-1} \begin{pmatrix} \boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)} \\ \boldsymbol{\mu}_{\tilde{S}^*}^{(1)} - \boldsymbol{\mu}_{\tilde{S}^*}^{(0)} \end{pmatrix} = \begin{pmatrix} * \\ \boldsymbol{\delta}'_{\tilde{S}^*} \end{pmatrix}, \tag{29}$$

where

$$\Sigma_{\tilde{S},\tilde{S}}^{-1} = \begin{pmatrix} (\Sigma_{S,S} - \Sigma_{S,\tilde{S}^*}\Sigma_{\tilde{S}^*,\tilde{S}^*}^{-1}\Sigma_{\tilde{S}^*,S})^{-1} & -\Sigma_{S,S}^{-1}\Sigma_{S,\tilde{S}^*}(\Sigma_{\tilde{S}^*,\tilde{S}^*} - \Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\Sigma_{S,\tilde{S}^*})^{-1} \\ -\Sigma_{\tilde{S}^*,\tilde{S}^*}^{-1}\Sigma_{\tilde{S}^*,S}(\Sigma_{S,S} - \Sigma_{S,\tilde{S}^*}\Sigma_{\tilde{S}^*,\tilde{S}^*}^{-1}\Sigma_{\tilde{S}^*,S})^{-1} & (\Sigma_{\tilde{S}^*,\tilde{S}^*} - \Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\Sigma_{S,\tilde{S}^*})^{-1} \end{pmatrix}, \tag{30}$$

and $\boldsymbol{\delta}'_{\tilde{S}^*}$ consists of several components of $\boldsymbol{\delta}_{S^*}$. Denote $c = \Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\left(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}\right) - \left(\boldsymbol{\mu}_{\tilde{S}^*}^{(1)} - \boldsymbol{\mu}_{\tilde{S}^*}^{(0)}\right)$. Since $\Sigma_{\tilde{S},\tilde{S}}^{-1}$ is symmetric, combining with (29) we have

$$(\Sigma_{\tilde{S}^*,\tilde{S}^*} - \Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\Sigma_{S,\tilde{S}^*})^{-1}c = -\boldsymbol{\delta}'_{\tilde{S}^*},$$

leading to

$$\|(\Sigma_{\tilde{S}^*,\tilde{S}^*} - \Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\Sigma_{S,\tilde{S}^*})^{-1}c\|_2 = \|\boldsymbol{\delta}'_{\tilde{S}^*}\|_2 \geq \gamma,$$

by Assumption 4.(iii). For the left-hand side:

$$\begin{aligned} &\|(\Sigma_{\tilde{S}^*,\tilde{S}^*} - \Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\Sigma_{S,\tilde{S}^*})^{-1}c\|_2 \\ &\leq \|c\|_2 \cdot \|(\Sigma_{\tilde{S}^*,\tilde{S}^*} - \Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\Sigma_{S,\tilde{S}^*})^{-1}\|_2 \\ &\leq \|c\|_2 \cdot \|\Sigma_{\tilde{S},\tilde{S}}^{-1}\|_2 \\ &\leq \|c\|_2 \cdot \lambda_{\min}^{-1}(\Sigma_{\tilde{S},\tilde{S}}) \end{aligned}$$

$$\leq m^{-1}\|c\|_2,$$

which leads to $\|c\|_2 \geq \gamma m$. ■

Lemma 31 *For LDA,* $\mathrm{RIC}(S) = -\left(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}\right)^T \Sigma_{S,S}^{-1}\left(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}\right)$. *It satisfies the following conclusions:*

(i) If $S \supseteq S^*$, *then* $\mathrm{RIC}(S) = \mathrm{RIC}(S^*)$;

(ii) For $\tilde{S} = \tilde{S}^* \cup S$, *we have*

$$\begin{aligned}\mathrm{RIC}(\tilde{S}) = \mathrm{RIC}(S) &- \left[\Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\left(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}\right) - \left(\boldsymbol{\mu}_{\tilde{S}^*}^{(1)} - \boldsymbol{\mu}_{\tilde{S}^*}^{(0)}\right)\right]^T \\ &\left(\Sigma_{\tilde{S}^*,\tilde{S}^*} - \Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\Sigma_{S,\tilde{S}^*}\right)\left[\Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\left(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}\right) - \left(\boldsymbol{\mu}_{\tilde{S}^*}^{(1)} - \boldsymbol{\mu}_{\tilde{S}^*}^{(0)}\right)\right]\end{aligned}$$

(iii) It holds

$$\inf_{\substack{S:|S|\leq D \\ S\not\supseteq S^*}} \mathrm{RIC}(S) - \mathrm{RIC}(S^*) \geq m^3\gamma^2.$$

Proof [Proof of Lemma 31]

(i) First let's suppose (ii) is correct. For $S \supseteq S^*$, consider $S_{\mathrm{Full}} = \tilde{S}^* \cup S$ and $\tilde{S}^* \cap S = \emptyset$, then by sparsity condition, we have

$$\begin{pmatrix} * \\ \mathbf{0} \end{pmatrix} = \Sigma^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)}) = \Sigma^{-1}\begin{pmatrix} \boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)} \\ \boldsymbol{\mu}_{\tilde{S}^*}^{(1)} - \boldsymbol{\mu}_{\tilde{S}^*}^{(0)} \end{pmatrix}.$$

Also by basic algebra, there holds

$$(\Sigma_{S,S} - \Sigma_{S,\tilde{S}^*}\Sigma_{\tilde{S}^*,\tilde{S}^*}^{-1}\Sigma_{\tilde{S}^*,S})^{-1} = \Sigma_{S,S}^{-1} - \Sigma_{S,S}^{-1}\Sigma_{S,\tilde{S}^*}(-\Sigma_{\tilde{S}^*,\tilde{S}^*} + \Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\Sigma_{S,\tilde{S}^*})^{-1}\Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}. \tag{31}$$

Combined with (30) and (31), it can be obtained that

$$\Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\left(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}\right) - \left(\boldsymbol{\mu}_{\tilde{S}^*}^{(1)} - \boldsymbol{\mu}_{\tilde{S}^*}^{(0)}\right) = \mathbf{0},$$

yielding that $\mathrm{RIC}(S_{\mathrm{Full}}) = \mathrm{RIC}(S)$. Since the same procedure can be conducted for arbitrary $S \supseteq S^*$, we complete the proof.

(ii) This can be directly calculated and simplified by applying (30) and (31).

(iii) It's easy to see that

$$\begin{aligned}\lambda_{\min}(\Sigma_{\tilde{S}^*,\tilde{S}^*} - \Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\Sigma_{S,\tilde{S}^*}) &= \lambda_{\max}^{-1}((\Sigma_{\tilde{S}^*,\tilde{S}^*} - \Sigma_{\tilde{S}^*,S}\Sigma_{S,S}^{-1}\Sigma_{S,\tilde{S}^*})^{-1}) \\ &\geq \lambda_{\max}^{-1}(\Sigma_{\tilde{S},\tilde{S}}^{-1}) \\ &= \lambda_{\min}(\Sigma_{\tilde{S},\tilde{S}})\end{aligned}$$

$$\geq m.$$

Then by (ii) and Lemma 30, it holds that

$$\begin{aligned}
&\mathrm{RIC}(\tilde{S}) - \mathrm{RIC}(S) \\
&\geq \left\| \Sigma_{\tilde{S}^*,S} \Sigma_{S,S}^{-1} \left(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)} \right) - \left(\boldsymbol{\mu}_{\tilde{S}^*}^{(1)} - \boldsymbol{\mu}_{\tilde{S}^*}^{(0)} \right) \right\|_2^2 \cdot \lambda_{\min}(\Sigma_{\tilde{S}^*,\tilde{S}^*} - \Sigma_{\tilde{S}^*,S} \Sigma_{S,S}^{-1} \Sigma_{S,\tilde{S}^*}) \\
&\geq m^3 \gamma^2,
\end{aligned}$$

which completes the proof.

■

Lemma 32 *If Assumption 4 holds, for ϵ smaller than some constant and n, p larger than some constants, we have*

$$\mathbb{P}\left(\sup_{S:|S|\leq D} |\widehat{\mathrm{RIC}}(S) - \mathrm{RIC}(S)| > \epsilon \right) \lesssim p^2 \exp\left\{ -Cn \left(\frac{\epsilon}{D^2} \right)^2 \right\}.$$

Proof [Proof of Lemma 32] By Lemma 28, when $\epsilon < m^{-1}$, there holds that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{S:|S|\leq D} |\widehat{\mathrm{RIC}}(S) - \mathrm{RIC}(S)| > \epsilon \right) \\
&= \mathbb{P}\left(\sup_{S:|S|\leq D} \left| \left(\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)} \right)^T \hat{\Sigma}_{S,S}^{-1} \left(\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)} \right) - \left(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)} \right)^T \Sigma_{S,S}^{-1} \left(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)} \right) \right| > \epsilon \right) \\
&\leq \mathbb{P}\left(\sup_{S:|S|\leq D} \left| \left(\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)} \right)^T \left(\hat{\Sigma}_{S,S}^{-1} - \Sigma_{S,S}^{-1} \right) \left(\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)} \right) \right| \right. \\
&\quad + \left| \left(\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)} - \boldsymbol{\mu}_S^{(1)} + \boldsymbol{\mu}_S^{(0)} \right)^T \Sigma_{S,S}^{-1} \left(\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)} \right) \right| \\
&\quad \left. + \left| \left(\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)} - \boldsymbol{\mu}_S^{(1)} + \boldsymbol{\mu}_S^{(0)} \right)^T \Sigma_{S,S}^{-1} \left(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)} \right) \right| > \epsilon \right) \\
&\leq \mathbb{P}\left(\sup_{S:|S|\leq D} \left[\left\| \hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)} \right\|_2^2 \left\| \hat{\Sigma}_{S,S}^{-1} - \Sigma_{S,S}^{-1} \right\|_2 + \left(\left\| (\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)}) - (\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}) \right\|_2 \right) \right. \right. \\
&\quad \left. \left. \cdot \left\| \Sigma_{S,S}^{-1} \right\|_2 \cdot \left(\left\| \hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)} \right\|_2 + \left\| \boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)} \right\|_2 \right) \right] > \epsilon \right) \\
&\leq \mathbb{P}\left((3M')^2 D \cdot \sup_{S:|S|\leq D} \left\| \hat{\Sigma}_{S,S}^{-1} - \Sigma_{S,S}^{-1} \right\|_2 > \frac{\epsilon}{3} \right) \\
&\quad + \mathbb{P}\left(\sup_{S:|S|\leq D} \left\| (\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)}) - (\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}) \right\|_\infty > \frac{M'}{2} \right)
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{P}\left(\sup_{S:|S|\leq D}\left\|(\hat{\boldsymbol{\mu}}_S^{(1)}-\hat{\boldsymbol{\mu}}_S^{(0)})-(\boldsymbol{\mu}_S^{(1)}-\boldsymbol{\mu}_S^{(0)})\right\|_2\cdot m^{-1}\left(\frac{3}{2}M'+M'\right)\sqrt{D}>\frac{\epsilon}{3}\right)\\
&\lesssim p^2\exp\left\{-Cn\left(\frac{\epsilon}{D^2}\right)^2\right\}+p\exp\left\{-Cn\cdot\frac{\epsilon}{D}\right\}+p\exp\left\{-Cn\right\}+p\exp\left\{-Cn\left(\frac{\epsilon}{D}\right)^2\right\}\\
&\lesssim p^2\exp\left\{-Cn\left(\frac{\epsilon}{D^2}\right)^2\right\}.
\end{aligned}
$$

■

Then the following steps to prove Theorem 18 are analogous to what we did to prove Theorem 15.

B.11 Proof of Theorem 20

Denote $\hat{\Omega}_{S,S}^{(r)}=(\hat{\Sigma}_{S,S}^{(r)})^{-1}, r=0,1$. Then we denote

$$
\begin{aligned}
T(S)&=\mathrm{Tr}\left[(\Omega_{S,S}^{(1)}-\Omega_{S,S}^{(0)})(\pi_1\Sigma_{S,S}^{(1)}-\pi_0\Sigma_{S,S}^{(0)})\right]+(\pi_1-\pi_0)(\log|\Sigma_{S,S}^{(1)}|-\log|\Sigma_{S,S}^{(0)}|),\\
D(S)&=(\boldsymbol{\mu}_S^{(1)}-\boldsymbol{\mu}_S^{(0)})^T\left[\pi_1\Omega_{S,S}^{(0)}+\pi_0\Omega_{S,S}^{(1)}\right](\boldsymbol{\mu}_S^{(1)}-\boldsymbol{\mu}_S^{(0)}),\\
\mathrm{RIC}(S)&=2[T(S)-D(S)].
\end{aligned}
$$

And their sample versions by MLEs are

$$
\begin{aligned}
\hat{T}(S)&=\mathrm{Tr}\left[(\hat{\Omega}_{S,S}^{(1)}-\hat{\Omega}_{S,S}^{(0)})(\hat{\pi}_1\hat{\Sigma}_{S,S}^{(1)}-\hat{\pi}_0\hat{\Sigma}_{S,S}^{(0)})\right]+(\hat{\pi}_1-\hat{\pi}_0)(\log|\hat{\Sigma}_{S,S}^{(1)}|-\log|\hat{\Sigma}_{S,S}^{(0)}|),\\
\hat{D}(S)&=(\hat{\boldsymbol{\mu}}_S^{(1)}-\hat{\boldsymbol{\mu}}_S^{(0)})^T\left[\hat{\pi}_1\Omega_{S,S}^{(0)}+\hat{\pi}_0\Omega_{S,S}^{(1)}\right](\hat{\boldsymbol{\mu}}_S^{(1)}-\hat{\boldsymbol{\mu}}_S^{(0)}),\\
\widehat{\mathrm{RIC}}(S)&=2[\hat{T}(S)-\hat{D}(S)].
\end{aligned}
$$

Similar to Lemma 28, we have the following lemma holds.

Lemma 33 *For arbitrary $\epsilon\in(0,m^{-1})$, we have conclusions in the follows for $r=0,1$:*

(i) $\mathbb{P}(\|(\hat{\boldsymbol{\mu}}^{(1)}-\hat{\boldsymbol{\mu}}^{(0)})-(\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(0)})\|_\infty>\epsilon)\lesssim p\exp\{-Cn\epsilon^2\}$;

(ii) $\mathbb{P}\left(\sup_{S:|S|\leq D}\left\|\hat{\Sigma}_{S,S}^{(r)}-\Sigma_{S,S}^{(r)}\right\|_2>\epsilon\right)\lesssim p^2\exp\left\{-Cn\cdot\left(\frac{\epsilon}{D}\right)^2\right\}+p\exp\left\{-Cn\cdot\frac{\epsilon}{D}\right\}$;

(iii) $\mathbb{P}\left(\sup_{S:|S|\leq D}\left\|\hat{\Omega}_{S,S}^{(r)}-\Omega_{S,S}^{(r)}\right\|_2>\epsilon\right)\lesssim p^2\exp\left\{-Cn\cdot\left(\frac{\epsilon}{D}\right)^2\right\}+p\exp\left\{-Cn\cdot\frac{\epsilon}{D}\right\}$.

Also, the lemmas in the follows are useful to prove Theorem 20 as well.

Lemma 34 *There hold the following conditions:*

(i) For $\tilde{S} = S \cup \{j\}, k_{j,S}^{(r)} = \Sigma_{j,j}^{(r)} - \Sigma_{j,S}^{(r)}(\Sigma_{S,S}^{(r)})^{-1}\Sigma_{S,j}^{(r)}, r = 0, 1$*, where* $j \notin S$*, we have:*

$$\begin{aligned} T(\tilde{S}) - T(S) = &- \left[\Omega_{S,S}^{(1)}\Sigma_{S,j}^{(1)} - \Omega_{S,S}^{(0)}\Sigma_{S,j}^{(0)}\right]^T \left(\frac{\pi_0}{k_{j,S}^{(0)}}\Sigma_{S,S}^{(0)} + \frac{\pi_1}{k_{j,S}^{(1)}}\Sigma_{S,S}^{(1)}\right) \\ &\left[\Omega_{S,S}^{(1)}\Sigma_{S,j}^{(1)} - \Omega_{S,S}^{(0)}\Sigma_{S,j}^{(0)}\right] + 1 - \frac{\pi_0}{k_{j,S}^{(1)}} \cdot k_{j,S}^{(0)} - \frac{\pi_1}{k_{j,S}^{(0)}} \cdot k_{j,S}^{(1)} \\ &+ (\pi_1 - \pi_0)\log\left(\frac{k_{j,S}^{(1)}}{k_{j,S}^{(0)}}\right). \end{aligned} \tag{32}$$

And for $\tilde{S} = S \cup \tilde{S}^*$ *where* $S \cap \tilde{S}^* = \emptyset$*, there holds*

$$\begin{aligned} D(\tilde{S}) - D(S) = &\pi_1 \left[\Sigma_{\tilde{S}^*,S}^{(0)}\Omega_{S,S}^{(0)}(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}) - (\boldsymbol{\mu}_{\tilde{S}^*}^{(1)} - \boldsymbol{\mu}_{\tilde{S}^*}^{(0)})\right]^T \\ &\left(\Sigma_{\tilde{S}^*,\tilde{S}^*}^{(0)} - \Sigma_{\tilde{S}^*,S}^{(0)}\Omega_{S,S}^{(0)}\Sigma_{S,\tilde{S}^*}^{(0)}\right)\left[\Sigma_{\tilde{S}^*,S}^{(0)}\Omega_{S,S}^{(0)}(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}) - (\boldsymbol{\mu}_{\tilde{S}^*}^{(1)} - \boldsymbol{\mu}_{\tilde{S}^*}^{(0)})\right] \\ &+ \pi_0 \left[\Sigma_{\tilde{S}^*,S}^{(1)}\Omega_{S,S}^{(1)}(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}) - (\boldsymbol{\mu}_{\tilde{S}^*}^{(1)} - \boldsymbol{\mu}_{\tilde{S}^*}^{(0)})\right]^T \\ &\left(\Sigma_{\tilde{S}^*,\tilde{S}^*}^{(1)} - \Sigma_{\tilde{S}^*,S}^{(1)}\Omega_{S,S}^{(1)}\Sigma_{S,\tilde{S}^*}^{(1)}\right)\left[\Sigma_{\tilde{S}^*,S}^{(1)}\Omega_{S,S}^{(1)}(\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}) - (\boldsymbol{\mu}_{\tilde{S}^*}^{(1)} - \boldsymbol{\mu}_{\tilde{S}^*}^{(0)})\right]. \end{aligned} \tag{33}$$

(ii) Further, (i) implies the monotonicity of T, D, RIC *in the following sense: If* $S_1 \supseteq S_2$*, then* $T(S_1) \le T(S_2), D(S_1) \ge D(S_2)$*, which leads to* $\mathrm{RIC}(S_1) \le \mathrm{RIC}(S_2)$*.*

(iii) Define $S_l^* = \{j : [(\Sigma^{(0)})^{-1}\boldsymbol{\mu}^{(0)} - (\Sigma^{(1)})^{-1}\boldsymbol{\mu}^{(1)}]_j \neq 0\}$*,* $S_q^* = \{j : [(\Sigma^{(1)})^{-1} - (\Sigma^{(0)})^{-1}]_{ij} \neq 0, \exists i\}$*, then:*

(a) If $S \supseteq S_q^*$*, then* $T(S) = T(S_q^*)$*;*

(b) If $S \supseteq S^*$*, then* $D(S) = D(S^*) = D(S_l^*)$*.*

Proof [Proof of Lemma 34]

(i) (33) is obvious due to Lemma 31. Now let's prove (32). It's easy to see that

$$\begin{aligned} T(\tilde{S}) &= \mathrm{Tr}\left[(\Omega_{\tilde{S},\tilde{S}}^{(1)} - \Omega_{\tilde{S},\tilde{S}}^{(0)})(\pi_1\Sigma_{\tilde{S},\tilde{S}}^{(1)} - \pi_0\Sigma_{\tilde{S},\tilde{S}}^{(0)})\right] + (\pi_1 - \pi_0)(\log|\Sigma_{\tilde{S},\tilde{S}}^{(1)}| - \log|\Sigma_{\tilde{S},\tilde{S}}^{(0)}|) \\ &= |S| - \pi_1\mathrm{Tr}\left[\Omega_{\tilde{S},\tilde{S}}^{(0)}\Sigma_{\tilde{S},\tilde{S}}^{(1)}\right] - \pi_0\mathrm{Tr}\left[\Omega_{\tilde{S},\tilde{S}}^{(1)}\Sigma_{\tilde{S},\tilde{S}}^{(0)}\right] + (\pi_1 - \pi_0)(\log|\Sigma_{\tilde{S},\tilde{S}}^{(1)}| - \log|\Sigma_{\tilde{S},\tilde{S}}^{(0)}|). \end{aligned} \tag{34}$$

Because

$$\Omega_{\tilde{S},\tilde{S}}^{(r)} = \begin{pmatrix} \Omega_{S,S}^{(r)} + \frac{1}{k_{j,S}^{(r)}}\Omega_{S,S}^{(r)}\Sigma_{S,j}^{(r)}\Sigma_{j,S}^{(r)}\Omega_{S,S}^{(r)} & -\frac{1}{k_{j,S}^{(r)}}\Omega_{S,S}^{(r)}\Sigma_{S,j}^{(r)} \\ -\frac{1}{k_{j,S}^{(r)}}\Sigma_{j,S}^{(r)}\Omega_{S,S}^{(r)} & \frac{1}{k_{j,S}^{(r)}} \end{pmatrix}, \tag{35}$$

for $r = 0, 1$, we have

$$\begin{aligned}
&\mathrm{Tr}\left(\Omega^{(1)}_{\tilde{S},\tilde{S}}\Sigma^{(0)}_{\tilde{S},\tilde{S}}\right)\\
&= \mathrm{Tr}\left[\Omega^{(1)}_{S,S}\Sigma^{(0)}_{S,S} + \frac{1}{k^{(1)}_{j,S}}\Omega^{(1)}_{S,S}\Sigma^{(1)}_{S,j}\Sigma^{(1)}_{j,S}\Omega^{(1)}_{S,S}\Sigma^{(0)}_{S,S} - \frac{1}{k^{(1)}_{j,S}}\Omega^{(1)}_{S,S}\Sigma^{(1)}_{S,j}\Sigma^{(0)}_{j,S}\right]\\
&\quad - \frac{1}{k^{(1)}_{j,S}}\Sigma^{(1)}_{j,S}\Omega^{(1)}_{S,S}\Sigma^{(0)}_{S,j} + \frac{\Sigma^{(0)}_{j,j}}{k^{(1)}_{j,S}}\\
&= \mathrm{Tr}[\Omega^{(1)}_{S,S}\Sigma^{(0)}_{S,S}] + \frac{1}{k^{(1)}_{j,S}}\Sigma^{(1)}_{j,S}\Omega^{(1)}_{S,S}\Sigma^{(0)}_{S,S}\Omega^{(1)}_{S,S}\Sigma^{(1)}_{S,j} - \frac{2}{k^{(1)}_{j,S}}\Sigma^{(1)}_{j,S}\Omega^{(1)}_{S,S}\Sigma^{(0)}_{S,j} + \frac{\Sigma^{(0)}_{j,j}}{k^{(1)}_{j,S}}, \quad (36)
\end{aligned}$$

where the last equality follows from the fact that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ if A and B are square matrices with the same dimension.

Similarly we have

$$\mathrm{Tr}\left(\Omega^{(0)}_{\tilde{S},\tilde{S}}\Sigma^{(1)}_{\tilde{S},\tilde{S}}\right) = \mathrm{Tr}[\Omega^{(0)}_{S,S}\Sigma^{(1)}_{S,S}] + \frac{1}{k^{(0)}_{j,S}}\Sigma^{(0)}_{j,S}\Omega^{(0)}_{S,S}\Sigma^{(1)}_{S,S}\Omega^{(0)}_{S,S}\Sigma^{(0)}_{S,j} - \frac{2}{k^{(0)}_{j,S}}\Sigma^{(0)}_{j,S}\Omega^{(0)}_{S,S}\Sigma^{(1)}_{S,j} + \frac{\Sigma^{(1)}_{j,j}}{k^{(0)}_{j,S}}. \quad (37)$$

Combining (34), (36), (37), the fact that $\Sigma^{(r)}_{j,j} = k^{(r)}_{j,S} + \Sigma^{(r)}_{j,S}\Omega^{(r)}_{S,S}\Sigma^{(r)}_{S,j}$, $|\Sigma^{(r)}_{\tilde{S},\tilde{S}}| = |\Sigma^{(r)}_{S,S}| \cdot |k^{(r)}_{j,S}|$, $r = 0, 1$ and $T(S) = |S| - \pi_1\mathrm{Tr}\left(\Omega^{(0)}_{S,S}\Sigma^{(1)}_{S,S}\right) - \pi_0\mathrm{Tr}\left(\Omega^{(1)}_{S,S}\Sigma^{(0)}_{S,S}\right) + (\pi_1 - \pi_0)(\log|\Sigma^{(1)}_{S,S}| - \log|\Sigma^{(0)}_{S,S}|)$, (32) is obtained.

(ii) For the monotonicity, since $-\pi_0\log\left(\frac{k^{(1)}_{j,S}}{k^{(0)}_{j,S}}\right) = \pi_0\log\left(\frac{k^{(0)}_{j,S}}{k^{(1)}_{j,S}}\right) \le \pi_0\left(\frac{k^{(0)}_{j,S}}{k^{(1)}_{j,S}} - 1\right)$ and $\pi_1\log\left(\frac{k^{(1)}_{j,S}}{k^{(0)}_{j,S}}\right) \le \pi_1\left(\frac{k^{(1)}_{j,S}}{k^{(0)}_{j,S}} - 1\right)$, we have

$$\begin{aligned}
&1 + (\pi_1 - \pi_0)\log\left(\frac{k^{(1)}_{j,S}}{k^{(0)}_{j,S}}\right) - \frac{\pi_0}{k^{(1)}_{j,S}}\cdot k^{(0)}_{j,S} - \frac{\pi_1}{k^{(0)}_{j,S}}\cdot k^{(1)}_{j,S}\\
&\le 1 + \pi_0\left(\frac{k^{(0)}_{j,S}}{k^{(1)}_{j,S}} - 1\right) + \pi_1\left(\frac{k^{(1)}_{j,S}}{k^{(0)}_{j,S}} - 1\right) - \frac{\pi_0}{k^{(1)}_{j,S}}\cdot k^{(0)}_{j,S} - \frac{\pi_1}{k^{(0)}_{j,S}}\cdot k^{(1)}_{j,S}\\
&\le 0,
\end{aligned}$$

implying that

$$T(\tilde{S}) \le T(S).$$

And it's easy to see that $\Sigma^{(r)}_{\tilde{S}^*,\tilde{S}^*} - \Sigma^{(r)}_{\tilde{S}^*,S}\Omega^{(r)}_{S,S}\Sigma^{(r)}_{S,\tilde{S}^*}$ is positive-definite, thus

$$D(\tilde{S}) \ge D(S).$$

And we also have

$$\mathrm{RIC}(\tilde{S}) \leq \mathrm{RIC}(S).$$

By induction we will obtain the monotonicity.

(iii) Denote $\Omega^{(r)} = (\Sigma^{(r)})^{-1}, r = 0, 1$. Consider full feature space $S_{\text{Full}} \supseteq S^*$. Let's remove one feature which does not belong to S_q^* from S_{Full}. Again, without loss of generality (in fact, we can always switch this feature with the last one), suppose we are removing the last feature j. That is, $S_{\text{Full}} = S \cup \{j\}$.

By sparsity:

$$\Omega^{(1)} - \Omega^{(0)} = \begin{pmatrix} * & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}. \tag{38}$$

Plugging (35) into above and by simplification we can obtain that

$$k_{j,S}^{(0)} = k_{j,S}^{(1)}, \tag{39}$$

$$\Omega_{S,S}^{(0)} \Sigma_{S,j}^{(0)} = \Omega_{S,S}^{(1)} \Sigma_{S,j}^{(1)}, \tag{40}$$

By Lemma 34, $T(S_{\text{Full}}) = T(S)$. Besides, this also implies that

$$\Omega^{(1)} - \Omega^{(0)} = \begin{pmatrix} \Omega_{S,S}^{(1)} - \Omega_{S,S}^{(0)} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix}.$$

By induction, it can be seen that for all subspace $S \supseteq S_q^*, T(S) = T(S_q^*)$.

Then again consider $S_{\text{Full}} = S \cup \{j\}$ but $j \notin S^*$, therefore by sparsity, in addition to (38), we also have

$$\Omega^{(1)}\boldsymbol{\mu}^{(1)} - \Omega^{(0)}\boldsymbol{\mu}^{(0)} = \begin{pmatrix} * \\ 0 \end{pmatrix},$$

which together with (38) leads to

$$\Sigma_{j,S}^{(0)} \Omega_{S,S}^{(0)} (\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}) - (\boldsymbol{\mu}_j^{(1)} - \boldsymbol{\mu}_j^{(0)}) = 0,$$
$$\Sigma_{j,S}^{(1)} \Omega_{S,S}^{(1)} (\boldsymbol{\mu}_S^{(1)} - \boldsymbol{\mu}_S^{(0)}) - (\boldsymbol{\mu}_j^{(1)} - \boldsymbol{\mu}_j^{(0)}) = 0.$$

By Lemma 34, $D(S_{\text{Full}}) = D(S)$. And it holds that

$$\Omega^{(1)}\boldsymbol{\mu}^{(1)} - \Omega^{(0)}\boldsymbol{\mu}^{(0)} = \begin{pmatrix} \Omega_{S,S}^{(1)}\boldsymbol{\mu}_S^{(1)} - \Omega_{S,S}^{(0)}\boldsymbol{\mu}_S^{(0)} \\ 0 \end{pmatrix}.$$

Again by induction, it can be seen that for any subspace $S \supseteq S^*, D(S) = D(S^*)$.

■

Lemma 35 *It holds that*

$$\inf_{j \in S_q^*} \inf_{\substack{S: S^* \setminus S = \{j\} \\ |S| \leq D + p^*}} T(S) - T(S^*) \geq \frac{m^2}{4M} \gamma_q \cdot \min \left\{ m\gamma_q, \frac{1}{3}, \frac{m^2}{4M} \gamma_q \right\}.$$

Proof [Proof of Lemma 35] Suppose $S^*\backslash S = \{j\}, \tilde{S} = S \cup \{j\} \supseteq S^*$ and $j \in S_q^*$. Due to Lemma 34, equation (35) and Assumption 5, we have either $\left|(k_{j,S}^{(1)})^{-1} - (k_{j,S}^{(0)})^{-1}\right| \geq \gamma_q$ or $\|(k_{j,S}^{(1)})^{-1}\Omega_{S,S}^{(1)}\Sigma_{S,j}^{(1)} - (k_{j,S}^{(0)})^{-1}\Omega_{S,S}^{(0)}\Sigma_{S,j}^{(0)}\|_\infty \geq \gamma_q$ holds. And it's easy to notice that for $r = 0, 1$,

$$(k_{j,S}^{(r)})^{-1} \leq \left\|\Omega_{\tilde{S},\tilde{S}}^{(r)}\right\|_2 = \lambda_{\min}^{-1}(\Sigma_{\tilde{S},\tilde{S}}^{(r)}) \leq \lambda_{\min}^{-1}(\Sigma^{(r)}) \leq m^{-1},$$
$$k_{j,S}^{(r)} = \Sigma_{j,j}^{(r)} - \Sigma_{j,S}^{(r)}\Omega_{S,S}^{(r)}\Sigma_{S,j}^{(r)} \leq \Sigma_{j,j}^{(r)} \leq M,$$

which implies that

$$m \leq (k_{j,S}^{(r)}) \leq M, r = 0, 1.$$

(i) If $\left|(k_{j,S}^{(1)})^{-1} - (k_{j,S}^{(0)})^{-1}\right| \geq \gamma_q$, then we have

$$\left|\frac{k_{j,S}^{(0)}}{k_{j,S}^{(1)}} - 1\right| = \left|k_{j,S}^{(0)}\right| \cdot \left|(k_{j,S}^{(1)})^{-1} - (k_{j,S}^{(0)})^{-1}\right| \geq m\gamma_q,$$

leading to

$$\frac{k_{j,S}^{(0)}}{k_{j,S}^{(1)}} - 1 - \log\left(\frac{k_{j,S}^{(0)}}{k_{j,S}^{(1)}}\right) \geq m\gamma_q - \log(1 + m\gamma_q).$$

When $m\gamma_q \leq 1$, we know that

$$m\gamma_q - \log(1 + m\gamma_q) \geq \frac{1}{2}(m\gamma_q)^2 - \frac{1}{3}(m\gamma_q)^3 \geq \frac{1}{6}(m\gamma_q)^2.$$

When $m\gamma_q > 1$, it holds that

$$m\gamma_q - \log(1 + m\gamma_q) \geq \frac{1}{6}m\gamma_q.$$

Thus, we have

$$\frac{k_{j,S}^{(0)}}{k_{j,S}^{(1)}} - 1 - \log\left(\frac{k_{j,S}^{(0)}}{k_{j,S}^{(1)}}\right) \geq \frac{1}{6}\min\{m\gamma_q, (m\gamma_q)^2\}.$$

And the same result holds for $\frac{k_{j,S}^{(1)}}{k_{j,S}^{(0)}}$ as well. Therefore by (32), we have

$$\begin{aligned} T(S) - T(\tilde{S}) &= T(S) - T(S^*) \\ &\geq \pi_0\left[\frac{k_{j,S}^{(0)}}{k_{j,S}^{(1)}} - 1 - \log\left(\frac{k_{j,S}^{(0)}}{k_{j,S}^{(1)}}\right)\right] + \pi_1\left[\frac{k_{j,S}^{(1)}}{k_{j,S}^{(0)}} - 1 - \log\left(\frac{k_{j,S}^{(1)}}{k_{j,S}^{(0)}}\right)\right] \\ &\geq \frac{1}{6}\min\{m\gamma_q, (m\gamma_q)^2\}. \end{aligned}$$

Since this holds for arbitrary $j \in S_q^*$ and S satisfying $S^*\backslash S = \{j\}$, we obtain that

$$\inf_{j \in S_q^*} \inf_{\substack{S: S^*\backslash S = \{j\} \\ |S| \leq D + p^*}} T(S) - T(S^*) \geq \frac{1}{6}\min\{m\gamma_q, (m\gamma_q)^2\}. \tag{41}$$

(ii) If $\left\|(k_{j,S}^{(1)})^{-1}\Omega_{S,S}^{(1)}\Sigma_{S,j}^{(1)} - (k_{j,S}^{(0)})^{-1}\Omega_{S,S}^{(0)}\Sigma_{S,j}^{(0)}\right\|_\infty \geq \gamma_q$, when $\left|(k_{j,S}^{(1)})^{-1} - (k_{j,S}^{(0)})^{-1}\right| \geq \frac{m}{2M}\gamma_q$, similar to (i), we have

$$T(S) - T(\tilde{S}) \geq \frac{1}{6}\min\left\{\frac{m^2}{2M}\gamma_q, \frac{m^4}{4M^2}\gamma_q^2\right\}. \tag{42}$$

Otherwise,

$$\begin{aligned}\gamma_q &\leq \left\|(k_{j,S}^{(1)})^{-1}\Omega_{S,S}^{(1)}\Sigma_{S,j}^{(1)} - (k_{j,S}^{(0)})^{-1}\Omega_{S,S}^{(0)}\Sigma_{S,j}^{(0)}\right\|_\infty \\ &\leq \left|(k_{j,S}^{(1)})^{-1} - (k_{j,S}^{(0)})^{-1}\right| \cdot \left\|\Omega_{S,S}^{(1)}\Sigma_{S,j}^{(1)}\right\|_\infty + (k_{j,S}^{(0)})^{-1} \cdot \left\|\Omega_{S,S}^{(1)}\Sigma_{S,j}^{(1)} - \Omega_{S,S}^{(0)}\Sigma_{S,j}^{(0)}\right\|_\infty.\end{aligned}$$

By (35), $(k_{j,S}^{(1)})^{-1}\Omega_{S,S}^{(1)}\Sigma_{S,j}^{(1)}$ is part of $\Omega_{\tilde{S},\tilde{S}}^{(1)}$, then

$$M^{-1}\left\|\Omega_{S,S}^{(1)}\Sigma_{S,j}^{(1)}\right\|_\infty \leq \left\|(k_{j,S}^{(1)})^{-1}\Omega_{S,S}^{(1)}\Sigma_{S,j}^{(1)}\right\|_\infty \leq \left\|\Omega_{\tilde{S},\tilde{S}}^{(1)}\right\|_{\max} \leq \left\|\Omega_{\tilde{S},\tilde{S}}^{(1)}\right\|_2 \leq m^{-1},$$

yielding

$$\left\|\Omega_{S,S}^{(1)}\Sigma_{S,j}^{(1)}\right\|_\infty \leq \frac{M}{m}.$$

Then we have

$$\left\|\Omega_{S,S}^{(1)}\Sigma_{S,j}^{(1)} - \Omega_{S,S}^{(0)}\Sigma_{S,j}^{(0)}\right\|_2 \geq \left\|\Omega_{S,S}^{(1)}\Sigma_{S,j}^{(1)} - \Omega_{S,S}^{(0)}\Sigma_{S,j}^{(0)}\right\|_\infty \geq \frac{1}{2}m\gamma_q,$$

leading to

$$\begin{aligned}T(S) - T(\tilde{S}) &\geq \left\|\Omega_{S,S}^{(1)}\Sigma_{S,j}^{(1)} - \Omega_{S,S}^{(0)}\Sigma_{S,j}^{(0)}\right\|_2^2 \cdot \lambda_{\min}\left(\pi_0(k_{j,S}^{(1)})^{-1}\Sigma_{S,S}^{(0)} + \pi_1(k_{j,S}^{(0)})^{-1}\Sigma_{S,S}^{(1)}\right) \\ &\geq \left(\frac{1}{2}m\gamma_q\right)^2 \cdot \left(\pi_0 M^{-1}\lambda_{\min}(\Sigma_{S,S}^{(0)}) + \pi_1 M^{-1}\lambda_{\min}(\Sigma_{S,S}^{(1)})\right) \\ &\geq \frac{m^3\gamma_q^2}{4M}.\end{aligned} \tag{43}$$

Combining (41), (42) and (43), we complete the proof.

■

Lemma 36 *It holds that*

$$\inf_{j\in S_l^*\backslash S_q^*}\inf_{\substack{S:S^*\backslash S=\{j\}\\|S|\leq D+p^*}}[D(S^*) - D(S)] \geq m^3\gamma_l^2.$$

Proof For any $j \in S_l^*\backslash S_q^*$ and S satisfying $S^*\backslash S = \{j\}$, let $\tilde{S} = S \cup \{j\}$, it's easy to see that there holds

$$\Omega_{\tilde{S},\tilde{S}}^{(1)}\begin{pmatrix}\boldsymbol{\mu}_S^{(1)}\\ \boldsymbol{\mu}_j^{(1)}\end{pmatrix} - \Omega_{\tilde{S},\tilde{S}}^{(0)}\begin{pmatrix}\boldsymbol{\mu}_S^{(0)}\\ \boldsymbol{\mu}_j^{(0)}\end{pmatrix} = \begin{pmatrix}*\\ \boldsymbol{\delta}_j'\end{pmatrix}, \tag{44}$$

Denote $c = \Sigma^{(1)}_{j,S}\Omega^{(1)}_{S,S}(\boldsymbol{\mu}^{(1)}_S - \boldsymbol{\mu}^{(0)}_S) - (\boldsymbol{\mu}^{(1)}_j - \boldsymbol{\mu}^{(0)}_j)$. By (39), (40), we can obtain

$$\Sigma^{(0)}_{j,S}\Omega^{(0)}_{S,S}(\boldsymbol{\mu}^{(1)}_S - \boldsymbol{\mu}^{(0)}_S) - (\boldsymbol{\mu}^{(1)}_j - \boldsymbol{\mu}^{(0)}_j) = \Sigma^{(1)}_{j,S}\Omega^{(1)}_{S,S}(\boldsymbol{\mu}^{(1)}_S - \boldsymbol{\mu}^{(0)}_S) - (\boldsymbol{\mu}^{(1)}_j - \boldsymbol{\mu}^{(0)}_j).$$

Combining with (44), we have

$$m^{-1}|c| \geq (k^{(1)}_{j,S})^{-1}|c| = |\boldsymbol{\delta}'_j| \geq \gamma_l,$$

because $\boldsymbol{\delta}'_j$ is the corresponding component of $\boldsymbol{\delta}_{S^*}$ with respect to feature j. We obtain that $|c| \geq m\gamma_l$, which leads to

$$D(S^*) - D(S) = D(\tilde{S}) - D(S) = k^{(1)}_{j,S}|c|^2 \geq m^3\gamma_l^2.$$

■

Lemma 37 *It holds that*

$$\inf_{\substack{S:S\not\supseteq S^* \\ |S|\leq D}} \mathrm{RIC}(S) - \mathrm{RIC}(S^*) \geq \min\left\{\frac{m^2}{4M}\gamma_q \cdot \min\left\{m\gamma_q, \frac{1}{3}, \frac{m^2}{4M}\gamma_q\right\}, m^3\gamma_l^2\right\}.$$

Proof [Proof of Lemma 37] Because of Lemmas 35 and 36, it suffices to prove

$$\begin{aligned}
&\inf_{\substack{S:S\not\supseteq S^* \\ |S|\leq D}} \mathrm{RIC}(S) - \mathrm{RIC}(S^*) \\
&\geq \inf_{j\in S^*} \inf_{\substack{S:S^*\backslash S=\{j\} \\ |S|\leq D+p^*}} \mathrm{RIC}(S) - \mathrm{RIC}(S^*) \\
&\geq \min\left\{\inf_{j\in S^*_q} \inf_{\substack{S:S^*\backslash S=\{j\} \\ |S|\leq D+p^*}} [T(S) - T(S^*)], \inf_{j\in S^*_l\backslash S^*_q} \inf_{\substack{S:j\notin S \\ |S|\leq D+p^*}} [D(S^*) - D(S)]\right\}.
\end{aligned}$$

For any subset S which does not cover S^* and $|S| \leq D$, consider feature $j \in S^*\backslash S$ and $\tilde{S} = S \cup S^*\backslash\{j\} \supseteq S$. It's easy to notice that $|\tilde{S}| \leq D + p^*$. By the monotonicity proved in Lemma 34, we know that $\mathrm{RIC}(\tilde{S}) \leq \mathrm{RIC}(S)$. In addition, we know that

$$\mathrm{RIC}(\tilde{S}) \geq \inf_{j\in S^*} \inf_{\substack{S:S^*\backslash S=\{j\} \\ |S|\leq D+p^*}} \mathrm{RIC}(S),$$

which yields the first inequality. For the second inequality, it directly comes from Lemma 34.(iii). ■

Lemma 38 *If Assumption 5 holds, for ϵ smaller than some constant and n, p larger than some constants, we have*

$$\mathbb{P}\left(\sup_{S:|S|\leq D} |\widehat{\mathrm{RIC}}(S) - \mathrm{RIC}(S)| > \epsilon\right) \lesssim p^2 \exp\left\{-Cn\left(\frac{\epsilon}{D^2}\right)^2\right\}.$$

Proof [Proof of Lemma 38] Recall that $\mathrm{RIC}(S) = 2[T(S) - D(S)]$, $\widehat{\mathrm{RIC}}(S) = 2[\hat{T}(S) - \hat{D}(S)]$. Denote $T_1(S) = \mathrm{Tr}\left[\left(\Omega_{S,S}^{(1)} - \Omega_{S,S}^{(0)}\right)\left(\pi_1\Sigma_{S,S}^{(1)} - \pi_0\Sigma_{S,S}^{(0)}\right)\right], T_2(S) = (\pi_1 - \pi_0)(\log|\Sigma_{S,S}^{(1)}| - \log|\Sigma_{S,S}^{(0)}|)$, then $T(S) = T_1(S) + T_2(S)$.

And since for any S with $|S| \le D$, we have

$$
\begin{aligned}
&\left|\hat{T}_1(S) - T_1(S)\right| \\
&= \left|\mathrm{Tr}\left[\left(\hat{\Omega}_{S,S}^{(1)} - \hat{\Omega}_{S,S}^{(0)}\right)\left(\hat{\pi}_1\hat{\Sigma}_{S,S}^{(1)} - \hat{\pi}_0\hat{\Sigma}_{S,S}^{(0)}\right) - \left(\Omega_{S,S}^{(1)} - \Omega_{S,S}^{(0)}\right)\left(\pi_1\Sigma_{S,S}^{(1)} - \pi_0\Sigma_{S,S}^{(0)}\right)\right]\right| \\
&\le D\left\|\left(\hat{\Omega}_{S,S}^{(1)} - \hat{\Omega}_{S,S}^{(0)}\right)\left(\hat{\pi}_1\hat{\Sigma}_{S,S}^{(1)} - \hat{\pi}_0\hat{\Sigma}_{S,S}^{(0)}\right) - \left(\Omega_{S,S}^{(1)} - \Omega_{S,S}^{(0)}\right)\left(\pi_1\Sigma_{S,S}^{(1)} - \pi_0\Sigma_{S,S}^{(0)}\right)\right\|_2 \\
&\le D\left\|\Omega_{S,S}^{(1)} - \Omega_{S,S}^{(0)}\right\|_2 \cdot \left[|\hat{\pi}_1 - \pi_1| \cdot \left(\left\|\Sigma_{S,S}^{(1)}\right\|_2 + \left\|\Sigma_{S,S}^{(0)}\right\|_2\right) + \left\|\hat{\Sigma}_{S,S}^{(1)} - \Sigma_{S,S}^{(1)}\right\|_2\right. \\
&\quad \left. + \left\|\hat{\Sigma}_{S,S}^{(0)} - \Sigma_{S,S}^{(0)}\right\|_2\right] + D\left\|\pi_1\Sigma_{S,S}^{(1)} - \pi_0\Sigma_{S,S}^{(0)}\right\|_2 \cdot \left(\left\|\hat{\Omega}_{S,S}^{(1)} - \Omega_{S,S}^{(1)}\right\|_2 + \left\|\hat{\Omega}_{S,S}^{(0)} - \Omega_{S,S}^{(0)}\right\|_2\right) \\
&\le 2Dm^{-1} \cdot \left[|\hat{\pi}_1 - \pi_1| \cdot 2M + \left\|\hat{\Sigma}_{S,S}^{(1)} - \Sigma_{S,S}^{(1)}\right\|_2 + \left\|\hat{\Sigma}_{S,S}^{(0)} - \Sigma_{S,S}^{(0)}\right\|_2\right] \\
&\quad + DM \cdot \left(\left\|\hat{\Omega}_{S,S}^{(1)} - \Omega_{S,S}^{(1)}\right\|_2 + \left\|\hat{\Omega}_{S,S}^{(0)} - \Omega_{S,S}^{(0)}\right\|_2\right),
\end{aligned}
$$

where the last inequality comes from $\left\|\Sigma_{S,S}^{(r)}\right\|_2 \le M$ and $\left\|\Omega_{S,S}^{(1)} - \Omega_{S,S}^{(0)}\right\|_2 \le 2m^{-1}$. And this yields

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{S:|S|\le D}\left|\hat{T}_1(S) - T_1(S)\right| > \epsilon\right) \\
&\le \mathbb{P}\left(4Dm^{-1}M|\hat{\pi}_1 - \pi_1| > \frac{\epsilon}{4}\right) + \mathbb{P}\left(2Dm^{-1} \cdot \sup_{S:|S|\le D}\left\|\hat{\Sigma}_{S,S}^{(1)} - \Sigma_{S,S}^{(1)}\right\|_2 > \frac{\epsilon}{8}\right) \\
&\quad + \mathbb{P}\left(2Dm^{-1} \cdot \sup_{S:|S|\le D}\left\|\hat{\Sigma}_{S,S}^{(0)} - \Sigma_{S,S}^{(0)}\right\|_2 > \frac{\epsilon}{8}\right) + \mathbb{P}\left(DM \sup_{S:|S|\le D}\left\|\hat{\Omega}_{S,S}^{(1)} - \Omega_{S,S}^{(1)}\right\|_2 > \frac{\epsilon}{4}\right) \\
&\quad + \mathbb{P}\left(DM \sup_{S:|S|\le D}\left\|\hat{\Omega}_{S,S}^{(0)} - \Omega_{S,S}^{(0)}\right\|_2 > \frac{\epsilon}{4}\right) \\
&\lesssim p^2 \exp\left\{-Cn\left(\frac{\epsilon}{D^2}\right)^2\right\}.
\end{aligned}
$$

On the other hand, we have

$$
\begin{aligned}
\left|\log\left(\frac{|\hat{\Sigma}_{S,S}^{(1)}|}{|\Sigma_{S,S}^{(1)}|}\right)\right| &\le \sum_{i=1}^{|S|}\left|\log\left(\frac{\lambda_i(\hat{\Sigma}_{S,S}^{(1)})}{\lambda_i(\Sigma_{S,S}^{(1)})}\right)\right| \\
&\le \sum_{i=1}^{|S|}\left|\frac{\lambda_i(\hat{\Sigma}_{S,S}^{(1)})}{\lambda_i(\Sigma_{S,S}^{(1)})} - 1\right| + o\left(\left|\frac{\lambda_i(\hat{\Sigma}_{S,S}^{(1)})}{\lambda_i(\Sigma_{S,S}^{(1)})} - 1\right|\right).
\end{aligned}
$$

By Weyl's inequality,

$$
\left|\frac{\lambda_i(\hat{\Sigma}_{S,S}^{(1)}) - \lambda_i(\Sigma_{S,S}^{(1)})}{\lambda_i(\Sigma_{S,S}^{(1)})}\right| \le \frac{1}{m}\left\|\hat{\Sigma}_{S,S}^{(1)} - \Sigma_{S,S}^{(1)}\right\|_2.
$$

Therefore there exists $\epsilon' > 0$ such that when $\left\|\hat{\Sigma}^{(1)}_{S,S} - \Sigma^{(1)}_{S,S}\right\|_2 \leq \epsilon'$, we have

$$\left|\log\left(\frac{|\hat{\Sigma}^{(1)}_{S,S}|}{|\Sigma^{(1)}_{S,S}|}\right)\right| \leq \frac{D}{m}\left\|\hat{\Sigma}^{(1)}_{S,S} - \Sigma^{(1)}_{S,S}\right\|_2 + o\left(D\left\|\hat{\Sigma}^{(1)}_{S,S} - \Sigma^{(1)}_{S,S}\right\|_2\right) \leq \frac{2D}{m}\left\|\hat{\Sigma}^{(1)}_{S,S} - \Sigma^{(1)}_{S,S}\right\|_2.$$

Therefore, there holds

$$\begin{aligned}
&\mathbb{P}\left(\sup_{S:|S|\leq D}\left|\hat{T}_2(S) - T_2(S)\right| > \epsilon\right) \\
&\leq \mathbb{P}\left(\sup_{S:|S|\leq D} 2\left|\hat{\pi}_1 - \pi_1\right| \cdot \left|\log|\Sigma^{(1)}_{S,S}| + \log|\Sigma^{(0)}_{S,S}|\right| + \log\left(\frac{|\hat{\Sigma}^{(1)}_{S,S}|}{|\Sigma^{(1)}_{S,S}|}\right) + \log\left(\frac{|\hat{\Sigma}^{(0)}_{S,S}|}{|\Sigma^{(0)}_{S,S}|}\right) > \epsilon\right) \\
&\leq \mathbb{P}\left(2\left|\hat{\pi}_1 - \pi_1\right| \cdot \sup_{S:|S|\leq D}\sum_{i=1}^{|S|}\left[\left|\log\left(\lambda_i\left(\Sigma^{(1)}_{S,S}\right)\right)\right| + \left|\log\left(\lambda_i\left(\Sigma^{(1)}_{S,S}\right)\right)\right|\right] > \frac{\epsilon}{3}\right) \\
&\quad + \mathbb{P}\left(\sup_{S:|S|\leq D}\log\left(\frac{|\hat{\Sigma}^{(1)}_{S,S}|}{|\Sigma^{(1)}_{S,S}|}\right) > \frac{\epsilon}{3}\right) + \mathbb{P}\left(\sup_{S:|S|\leq D}\log\left(\frac{|\hat{\Sigma}^{(0)}_{S,S}|}{|\Sigma^{(0)}_{S,S}|}\right) > \frac{\epsilon}{3}\right) \\
&\leq \mathbb{P}\left(2D\max\{|\log M|, |\log m|\} \cdot \left|\hat{\pi}_1 - \pi_1\right| > \frac{\epsilon}{3}\right) + \mathbb{P}\left(\frac{2D}{m} \cdot \sup_{S:|S|\leq D}\left\|\hat{\Sigma}^{(1)}_{S,S} - \Sigma^{(1)}_{S,S}\right\|_2 > \frac{\epsilon}{3}\right) \\
&\quad + \mathbb{P}\left(\frac{2D}{m} \cdot \sup_{S:|S|\leq D}\left\|\hat{\Sigma}^{(0)}_{S,S} - \Sigma^{(0)}_{S,S}\right\|_2 > \frac{\epsilon}{3}\right) + \mathbb{P}\left(\sup_{S:|S|\leq D}\left\|\hat{\Sigma}^{(1)}_{S,S} - \Sigma^{(1)}_{S,S}\right\|_2 > \epsilon'\right) \\
&\quad + \mathbb{P}\left(\sup_{S:|S|\leq D}\left\|\hat{\Sigma}^{(0)}_{S,S} - \Sigma^{(0)}_{S,S}\right\|_2 > \epsilon'\right) \\
&\lesssim p^2\exp\left\{-Cn\left(\frac{\epsilon}{D^2}\right)^2\right\}.
\end{aligned}$$

Thus we have

$$\mathbb{P}\left(\sup_{S:|S|\leq D}\left|\hat{T}(S) - T(S)\right| > \epsilon\right) \lesssim p^2\exp\left\{-Cn\left(\frac{\epsilon}{D^2}\right)^2\right\}. \tag{45}$$

Besides, following the same strategy in the proof of Lemma 32, it can be shown that

$$\mathbb{P}\left(\sup_{S:|S|\leq D}\left|\hat{D}(S) - D(S)\right| > \epsilon\right) \lesssim p^2\exp\left\{-Cn\left(\frac{\epsilon}{D^2}\right)^2\right\}. \tag{46}$$

By (45) and (46), we complete the proof. ■

By all the above lemmas, and following similar idea in the proof of Theorem 15, we can prove the consistency stated in the theorem.

B.12 Proof of Theorem 21

Firstly since taking subspace can also be seen as an axis-aligned projection, applying Theorem 2 in Cannings and Samworth (2017) we have

$$\mathbf{E}[R(C_n^{RaSE}) - R(C_{Bayes})] \leq \frac{\mathbf{E}[R(C_n^{S_{1*}}) - R(C_{Bayes})]}{\min(\alpha, 1-\alpha)}. \tag{47}$$

Then it can be easily noticed that

$$\begin{aligned}\mathbb{E}\{\mathbf{E}[R(C_n^{S_{1*}}) - R(C_{Bayes})]\} &\leq \mathbb{E}\{\mathbf{E}[(R(C_n^{S_{1*}}) - R(C_{Bayes}))\mathbb{1}(S_{1*} \supseteq S^*)]\} \\ &\quad + (1 - R(C_{Bayes}))\mathrm{P}(S_{1*} \not\supseteq S^*) \\ &\leq \mathbb{E} \sup_{\substack{S: S\supseteq S^* \\ |S|\leq D}} \left[R(C_n^S) - R(C_{Bayes})\right] + \mathrm{P}(S_{1*} \not\supseteq S^*). \end{aligned} \tag{48}$$

Combining (47) and (48), we obtain the conclusion.

B.13 Proof of Theorem 22

This conclusion is almost the same as Theorem 2 in Cannings and Samworth (2017). However, since we are studying the discrete space of subspaces and discriminative sets are ideal subspaces here, we can drop Assumptions 2 and 3 in their paper and get a similar upper bound. Firstly we have

$$\mathbf{E}[R(C_n^{RaSE}) - R(C_{Bayes})] \leq \frac{\mathbf{E}[R(C_n^{S_{1*}}) - R(C_{Bayes})]}{\min(\alpha, 1-\alpha)}, \tag{49}$$

by Cannings and Samworth (2017). Then write

$$\mathbf{E}[R(C_n^{S_{1*}})] = \mathbf{E}[R_n(C_n^{S_{1*}})] + \epsilon_n, \tag{50}$$

where $\epsilon_n = \mathbf{E}[R(C_n^{S_{1*}})] - \mathbf{E}[R_n(C_n^{S_{1*}})]$. Then we have

$$\begin{aligned}&\mathbf{E}[R_n(C_n^{S_{1*}})] \\ &\leq \sup_{\substack{S: S\supseteq S^* \\ |S|\leq D}} R_n(C_n^S) + \mathbf{E}\left[R(C_n^{S_{1*}}) \cdot \mathbb{1}\left(R_n(C_n^{S_{1*}}) > \sup_{\substack{S: S\supseteq S^* \\ |S|\leq D}} R_n(C_n^S)\right)\right] \\ &\leq \sup_{\substack{S: S\supseteq S^* \\ |S|\leq D}} R_n(C_n^S) + \mathbf{P}\left(R_n(C_n^{S_{1*}}) > \sup_{\substack{S: S\supseteq S^* \\ |S|\leq D}} R_n(C_n^S)\right) \\ &= \sup_{\substack{S: S\supseteq S^* \\ |S|\leq D}} R_n(C_n^S) + \left[\mathbf{P}\left(R_n(C_n^{S_{11}}) > \sup_{\substack{S: S\supseteq S^* \\ |S|\leq D}} R_n(C_n^S)\right)\right]^{B_2} \\ &\leq \sup_{\substack{S: S\supseteq S^* \\ |S|\leq D}} R_n(C_n^S) + (1 - p_{S^*})^{B_2}\end{aligned}$$

$$= \sup_{\substack{S:S\supseteq S^* \\ |S|\le D}} R(C_n^S) + \sup_{\substack{S:S\supseteq S^* \\ |S|\le D}} |\epsilon_n^S| + (1-p_{S^*})^{B_2}, \tag{51}$$

where $\epsilon_n^S = R(C_n^S) - R_n(C_n^S)$. Combining (49), (50) and (51), we obtain

$$\begin{aligned}
&\mathbb{E}\{\mathbf{E}[R(C_n^{RaSE}) - R(C_{Bayes})]\} \\
&\le \frac{\mathbb{E}\sup_{\substack{S:S\supseteq S^* \\ |S|\le D}} \left[R(C_n^S) - R(C_{Bayes})\right] + \mathbb{E}\sup_{\substack{S:S\supseteq S^* \\ |S|\le D}} |\epsilon_n^S| + \mathbb{E}(\epsilon_n) + (1-p_{S^*})^{B_2}}{\min(\alpha, 1-\alpha)}.
\end{aligned}$$

B.14 Proof of Proposition 23

First we prove the following lemma.

Lemma 39 (Devroye et al. (2013)) *For a non-negative random variable z satisfying*

$$\mathrm{P}(z > t) \le C_1 \exp\{-C_2 n t^\alpha\},$$

for any $t > 0$, where $C_1 > 1$, $C_2 > 0$ and $\alpha \ge 1$ are three fixed constants, then we have

$$\mathrm{E}z \le \left(\frac{\log C_1 + 1}{C_2 n}\right)^{\frac{1}{\alpha}}.$$

Proof [Proof of Lemma 39] It's easy to see

$$\mathrm{E}z^\alpha = \int_0^\infty \mathrm{P}(z^\alpha > t)dt = \epsilon + C_1 \int_\epsilon^\infty \exp\{-C_2 n t\}dt = \epsilon + \frac{C_1}{C_2 n}\cdot\exp\{-C_2 n\epsilon\},$$

leading to

$$\mathrm{E}z^\alpha \le \inf_{\epsilon>0}\left[\epsilon + \frac{C_1}{C_2 n}\cdot\exp\{-C_2 n\epsilon\}\right] = \frac{\log C_1 + 1}{C_2 n}.$$

Then by Jensen's inequality, it holds

$$\mathrm{E}z \le (\mathrm{E}z^\alpha)^{\frac{1}{\alpha}} \le \left(\frac{\log C_1 + 1}{C_2 n}\right)^{\frac{1}{\alpha}},$$

which completes the proof. ■

Then let's prove the proposition. Notice that

$$\Delta_S^2 = |\boldsymbol{\delta}_S^T \Sigma_{S,S} \boldsymbol{\delta}_S| \ge \|\boldsymbol{\delta}_S\|_2^2 \cdot \lambda_{\min}(\Sigma_{S,S}) \ge m p^* \gamma^2,$$

for any $S \supseteq S^*$. In addition, due to mean value theorem, it follows that

$$\Phi\left(-\frac{\hat{\Delta}_S}{2} + \hat{\tau}_S\right) - \Phi\left(-\frac{\Delta_S}{2} + \tau_S\right) \le \frac{1}{2}|\hat{\Delta}_S - \Delta_S| + |\hat{\tau}_S - \tau_S|.$$

By a similar argument, due to (10), we can obtain that

$$R(C_n^{S-LDA}) - R(C_{Bayes}) \leq \frac{1}{2}|\hat{\Delta}_S - \Delta_S| + |\hat{\tau}_S - \tau_S|.$$

By Lemma 32, we know that

$$\mathbb{P}\left(\sup_{S:|S|\leq D}|\hat{\Delta}_S^2 - \Delta_S^2| > \epsilon\right) \lesssim p^2 \exp\left\{-Cn\left(\frac{\epsilon}{D^2}\right)^2\right\}.$$

This yields that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\substack{S:S\supseteq S^*\\|S|\leq D}}\left[R(C_n^{S-LDA}) - R(C_{Bayes})\right] > \epsilon\right)\\
&\leq \mathbb{P}\left(\sup_{\substack{S:S\supseteq S^*\\|S|\leq D}}|\hat{\Delta}_S^2 - \Delta_S^2| > \frac{3}{4}mp^*\gamma^2\right)\\
&\quad + \mathbf{P}\left(\frac{1}{2}\cdot\sup_{\substack{S:S\supseteq S^*\\|S|\leq D}}\frac{|\hat{\Delta}_S^2 - \Delta_S^2|}{\hat{\Delta}_S + \Delta_S} > \frac{1}{2}\epsilon, \sup_{\substack{S:S\supseteq S^*\\|S|\leq D}}|\hat{\Delta}_S^2 - \Delta_S^2| \leq \frac{3}{4}mp^*\gamma^2\right)\\
&\quad + \mathbf{P}\left(\sup_{\substack{S:S\supseteq S^*\\|S|\leq D}}|\hat{\tau}_S - \tau_S| > \frac{1}{2}\epsilon, \sup_{\substack{S:S\supseteq S^*\\|S|\leq D}}|\hat{\Delta}_S^2 - \Delta_S^2| \leq \frac{3}{4}mp^*\gamma^2\right)\\
&\leq \mathbb{P}\left(\sup_{S:|S|\leq D}|\hat{\Delta}_S^2 - \Delta_S^2| > \frac{3}{4}mp^*\gamma^2\right) + \mathbb{P}\left(\sup_{S:|S|\leq D}|\hat{\Delta}_S^2 - \Delta_S^2| > \frac{3}{2}\epsilon\sqrt{mp^*\gamma^2}\right)\\
&\quad + \mathbb{P}\left(\log\left(\frac{\pi_1}{\pi_0}\right)\cdot\sup_{S:|S|\leq D}\frac{|\hat{\Delta}_S - \Delta_S|}{\hat{\Delta}_S\Delta_S} > \frac{1}{4}\epsilon, \sup_{\substack{S:S\supseteq S^*\\|S|\leq D}}|\hat{\Delta}_S^2 - \Delta_S^2| \leq \frac{3}{4}mp^*\gamma^2\right)\\
&\quad + \mathbb{P}\left(\left|\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_0}\right) - \log\left(\frac{\pi_1}{\pi_0}\right)\right|\cdot\sup_{S:|S|\leq D}\frac{1}{\hat{\Delta}_S} > \frac{1}{4}\epsilon, \sup_{\substack{S:S\supseteq S^*\\|S|\leq D}}|\hat{\Delta}_S^2 - \Delta_S^2| \leq \frac{3}{4}mp^*\gamma^2\right)\\
&\leq \mathbb{P}\left(\sup_{S:|S|\leq D}|\hat{\Delta}_S^2 - \Delta_S^2| > \frac{3}{4}mp^*\gamma^2\right) + \mathbb{P}\left(\sup_{S:|S|\leq D}|\hat{\Delta}_S^2 - \Delta_S^2| > \frac{3}{2}\epsilon\sqrt{mp^*\gamma^2}\right)\\
&\quad + \mathbb{P}\left(\sup_{S:|S|\leq D}|\hat{\Delta}_S^2 - \Delta_S^2| > C(p^*)^{\frac{3}{2}}\gamma^3\epsilon\right) + \mathbb{P}\left(|\hat{\pi}_0 - \pi_0| > C\epsilon\sqrt{mp^*\gamma^2}\right)\\
&\quad + \mathbb{P}\left(|\hat{\pi}_1 - \pi_1| > C\epsilon\sqrt{mp^*\gamma^2}\right)\\
&\lesssim p^2\exp\left\{-Cn\left(\frac{\epsilon(p^*)^{\frac{3}{2}}\gamma^3}{D^2}\right)^2\right\} + p^2\exp\left\{-Cn\left(\frac{\epsilon(p^*)^{\frac{1}{2}}\gamma}{D^2}\right)^2\right\},
\end{aligned}$$

which completes the proof by applying Lemma 39.

B.15 Proof of Proposition 24

Denote $\boldsymbol{\delta}_S = \Omega^{(1)}_{S,S}\boldsymbol{\mu}^{(1)}_S - \Omega^{(0)}_{S,S}\boldsymbol{\mu}^{(0)}_S$, $\Omega_{S,S} = \Omega^{(1)}_{S,S} - \Omega^{(0)}_{S,S}$, and $d_S(\boldsymbol{x}_S) = \log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}\boldsymbol{x}_S^T\Omega_{S,S}\boldsymbol{x}_S + \boldsymbol{\delta}_S^T\boldsymbol{x}_S - \frac{1}{2}(\boldsymbol{\mu}^{(1)}_S)^T\Omega^{(1)}_{S,S}\boldsymbol{\mu}^{(1)}_S + \frac{1}{2}(\boldsymbol{\mu}^{(0)}_S)^T\Omega^{(0)}_{S,S}\boldsymbol{\mu}^{(0)}_S$. Their estimators are correspondingly denoted as $\hat{\boldsymbol{\delta}}_S = \hat{\Omega}^{(1)}_{S,S}\hat{\boldsymbol{\mu}}^{(1)}_S - \hat{\Omega}^{(0)}_{S,S}\hat{\boldsymbol{\mu}}^{(0)}_S$, $\hat{\Omega}_{S,S} = \hat{\Omega}^{(1)}_{S,S} - \hat{\Omega}^{(0)}_{S,S}$, and $\hat{d}_S(\boldsymbol{x}_S) = \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_0}\right) - \frac{1}{2}\boldsymbol{x}_S^T\hat{\Omega}_{S,S}\boldsymbol{x}_S + \hat{\boldsymbol{\delta}}_S^T\boldsymbol{x}_S - \frac{1}{2}(\hat{\boldsymbol{\mu}}^{(1)}_S)^T\hat{\Omega}^{(1)}_{S,S}\hat{\boldsymbol{\mu}}^{(1)}_S + \frac{1}{2}(\hat{\boldsymbol{\mu}}^{(0)}_S)^T\hat{\Omega}^{(0)}_{S,S}\hat{\boldsymbol{\mu}}^{(0)}_S$. From the proof of Theorem 20, we know that $d_S(\boldsymbol{x}_S) = d_{S^*}(\boldsymbol{x}_{S^*})$ for any $S \supseteq S^*$. Denote the training data as D_{tr}, then $R(C_n^S) = \pi_0 \mathrm{P}^{(0)}(\hat{d}_S(\boldsymbol{x}_S) > 0|D_{tr}) + \pi_1 \mathrm{P}^{(1)}(\hat{d}_S(\boldsymbol{x}_S) \le 0|D_{tr}), R(C_{Bayes}) = \pi_0 \mathrm{P}^{(0)}(d_{S^*}(\boldsymbol{x}_{S^*}) > 0) + \pi_1 \mathrm{P}^{(1)}(d_{S^*}(\boldsymbol{x}_{S^*}) \le 0)$. Then we have

$$
\begin{aligned}
&\sup_{\substack{S:S\supseteq S^*\\|S|\le D}} [\mathrm{P}^{(0)}(\hat{d}_S(\boldsymbol{x}_S) > 0|D_{tr}) - \mathrm{P}^{(0)}(d_{S^*}(\boldsymbol{x}_{S^*}) > 0)] \\
&\le \sup_{\substack{S:S\supseteq S^*\\|S|\le D}} \mathrm{P}^{(0)}(d_S(\boldsymbol{x}_S) > d_S(\boldsymbol{x}_S) - \hat{d}_S(\boldsymbol{x}_S)|D_{tr}) - \mathrm{P}^{(0)}(d_{S^*}(\boldsymbol{x}_{S^*}) > 0) \\
&\le \mathrm{P}^{(0)}(d_{S^*}(\boldsymbol{x}_{S^*}) > -\epsilon) + \sup_{\substack{S:S\supseteq S^*\\|S|\le D}} \mathrm{P}^{(0)}(|d_S(\boldsymbol{x}_S) - \hat{d}_S(\boldsymbol{x}_S)| > \epsilon|D_{tr}) \\
&\quad - \mathrm{P}^{(0)}(d_{S^*}(\boldsymbol{x}_{S^*}) > 0) \\
&\le \int_{-\epsilon}^{0} h^{(0)}(z)dz + \sup_{\substack{S:S\supseteq S^*\\|S|\le D}} \mathrm{P}^{(0)}(|d_S(\boldsymbol{x}_S) - \hat{d}_S(\boldsymbol{x}_S)| > \epsilon|D_{tr}) \\
&\le \epsilon u_c + \sup_{\substack{S:S\supseteq S^*\\|S|\le D}} \mathrm{P}^{(0)}(|d_S(\boldsymbol{x}_S) - \hat{d}_S(\boldsymbol{x}_S)| > \epsilon|D_{tr}),
\end{aligned} \tag{52}
$$

for any $\epsilon \in (0, c)$. Denote $\boldsymbol{a}_S = (\Sigma^{(0)}_{S,S})^{\frac{1}{2}}[\Omega^{(1)}_{S,S}(\boldsymbol{\mu}^{(1)}_S - \hat{\boldsymbol{\mu}}^{(1)}_S) + (\Omega^{(1)}_{S,S} - \hat{\Omega}^{(1)}_{S,S})(\hat{\boldsymbol{\mu}}^{(1)}_S - \hat{\boldsymbol{\mu}}^{(0)}_S) + \hat{\Omega}^{(0)}_{S,S}(\hat{\boldsymbol{\mu}}^{(0)}_S - \boldsymbol{\mu}^{(0)}_S)]$. Notice that

$$
\begin{aligned}
d_S(\boldsymbol{x}_S) - \hat{d}_S(\boldsymbol{x}_S) = &\log\left(\frac{\pi_1}{\pi_0}\right) - \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_0}\right) + \frac{1}{2}(\boldsymbol{x}_S - \boldsymbol{\mu}^{(0)}_S)^T(\hat{\Omega}_{S,S} - \Omega_{S,S})(\boldsymbol{x}_S - \boldsymbol{\mu}^{(0)}_S) \\
&+ \boldsymbol{a}_S^T(\Sigma^{(0)}_{S,S})^{-\frac{1}{2}}(\boldsymbol{x}_S - \boldsymbol{\mu}^{(0)}_S) + \frac{1}{2}(\boldsymbol{\mu}^{(1)}_S - \boldsymbol{\mu}^{(0)}_S)^T\Omega^{(1)}_{S,S}(\hat{\boldsymbol{\mu}}^{(1)}_S - \boldsymbol{\mu}^{(1)}_S) \\
&- \frac{1}{2}(\hat{\boldsymbol{\mu}}^{(0)}_S - \boldsymbol{\mu}^{(0)}_S)^T\Omega^{(0)}_{S,S}(\hat{\boldsymbol{\mu}}^{(0)}_S - \boldsymbol{\mu}^{(0)}_S).
\end{aligned}
$$

Further, denote $\boldsymbol{z}_S = (\Sigma^{(0)}_{S,S})^{-\frac{1}{2}}(\boldsymbol{x}_S - \boldsymbol{\mu}^{(0)}_S) \sim N(\boldsymbol{0}_{|S|}, I_{|S|})$. Then it follows that

$$
\begin{aligned}
|d_S(\boldsymbol{x}_S) - \hat{d}_S(\boldsymbol{x}_S)| &\le C|\hat{\pi}_1 - \pi_1| + \frac{1}{2}\|\hat{\Omega}_{S,S} - \Omega_{S,S}\|_2 \cdot \|\Sigma^{(0)}_{S,S}\|_2 \cdot \|\boldsymbol{z}_S\|_2^2 + |\boldsymbol{a}_S^T\boldsymbol{z}_S| \\
&\quad + \frac{1}{2}\|\boldsymbol{\mu}^{(1)}_S - \boldsymbol{\mu}^{(0)}_S\|_2 \cdot \|\Omega^{(1)}_{S,S}\|_2 \cdot \|\hat{\boldsymbol{\mu}}^{(1)}_S - \boldsymbol{\mu}^{(1)}_S\|_2 \\
&\quad + \frac{1}{2}\|\hat{\boldsymbol{\mu}}^{(0)}_S - \boldsymbol{\mu}^{(0)}_S\|_2^2 \cdot \|\Omega^{(0)}_{S,S}\|_2 \\
&\le C|\hat{\pi}_1 - \pi_1| + \frac{M}{2}\|\hat{\Omega}_{S,S} - \Omega_{S,S}\|_2 \cdot \|\boldsymbol{z}_S\|_2^2 + |\boldsymbol{a}_S^T\boldsymbol{z}_S|
\end{aligned}
$$

$$+ M'Dm^{-1} \cdot \|\hat{\boldsymbol{\mu}}_S^{(1)} - \boldsymbol{\mu}_S^{(1)}\|_\infty + Dm^{-1} \cdot \|\hat{\boldsymbol{\mu}}_S^{(0)} - \boldsymbol{\mu}_S^{(0)}\|_\infty^2,$$

where

$$\begin{aligned}\|\boldsymbol{a}_S\|_2 &\lesssim \|\hat{\boldsymbol{\mu}}_S^{(1)} - \boldsymbol{\mu}_S^{(1)}\|_2 + \|\Omega_{S,S}^{(1)} - \hat{\Omega}_{S,S}^{(1)}\|_2 \cdot \|\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)}\|_2 + \|\hat{\boldsymbol{\mu}}_S^{(0)} - \boldsymbol{\mu}_S^{(0)}\|_2 \\ &\lesssim D^{\frac{1}{2}}(\|\hat{\boldsymbol{\mu}}_S^{(1)} - \boldsymbol{\mu}_S^{(1)}\|_\infty + \|\hat{\boldsymbol{\mu}}_S^{(0)} - \boldsymbol{\mu}_S^{(0)}\|_\infty) + \|\Omega_{S,S}^{(1)} - \hat{\Omega}_{S,S}^{(1)}\|_2 \cdot \|\hat{\boldsymbol{\mu}}_S^{(1)} - \hat{\boldsymbol{\mu}}_S^{(0)}\|_2.\end{aligned}$$

For any $t, t' > 0$, denote event $\mathcal{B} = \{C|\hat{\pi}_1 - \pi_1| \le \epsilon/4, \sup_{S:|S|\le D} \|\hat{\Omega}_{S,S} - \Omega_{S,S}\|_2 \le t, M'Dm^{-1}\|\hat{\boldsymbol{\mu}}_S^{(1)} - \boldsymbol{\mu}_S^{(1)}\|_\infty \le \epsilon/8, Dm^{-1}\|\hat{\boldsymbol{\mu}}_S^{(0)} - \boldsymbol{\mu}_S^{(0)}\|_\infty \le \epsilon/8, \|\boldsymbol{a}_S\|_2 \le t'\}$. When $\|\boldsymbol{a}_S\|_2 \le t'$, $\boldsymbol{a}_S^T\boldsymbol{z}_S$ is a t'-subGaussian. This yields that

$$\begin{aligned}&\mathrm{P}\left(\sup_{\substack{S:S\supseteq S^* \\ |S|\le D}} |d_S(\boldsymbol{x}_S) - \hat{d}_S(\boldsymbol{x}_S)| > \epsilon \middle| D_{tr} \in \mathcal{B}\right) \le \mathrm{P}(t\|\boldsymbol{z}_S\|_2^2 > \epsilon/4) + \mathrm{P}(|\boldsymbol{a}_S^T\boldsymbol{z}_S| > \epsilon/4 | \|\boldsymbol{a}_S\|_2 \le t') \\ &\qquad \lesssim \exp\left\{-\frac{1}{D}\left(\frac{C\epsilon}{t} - D\right)^2\right\} + \exp\left\{-C\left(\frac{\epsilon}{t'}\right)^2\right\}, \end{aligned} \tag{53}$$

where ϵ satisfies $\frac{C\epsilon}{t} > D$ and the first term comes from the tail bound of χ_1^2-distribution

$$\mathrm{P}\left(\|\boldsymbol{z}_S\|_2^2 > \frac{C\epsilon}{t}\right) = \mathrm{P}\left(\|\boldsymbol{z}_S\|_2^2 > D + \left(\frac{C\epsilon}{t} - D\right)\right) \le \exp\left\{-\frac{1}{D}\left(\frac{C\epsilon}{t} - D\right)^2\right\}.$$

Taking the expectation for the training data in (52), we have

$$\begin{aligned}&\mathbb{E} \sup_{\substack{S:S\supseteq S^* \\ |S|\le D}} [R(C_n^S) - R(C_{Bayes})] \\ &\le \epsilon u_c + \mathbb{E} \sup_{\substack{S:S\supseteq S^* \\ |S|\le D}} \mathrm{P}(|d_S(\boldsymbol{x}_S) - \hat{d}_S(\boldsymbol{x}_S)| > \epsilon | D_{tr} \in \mathcal{B}) + \mathbb{P}(\mathcal{B}^c) \\ &\lesssim \epsilon u_c + \exp\left\{-\frac{1}{D}\left(\frac{C\epsilon}{t} - D\right)^2\right\} + \exp\left\{-C\left(\frac{\epsilon}{t'}\right)^2\right\} + \exp\{-Cn\epsilon^2\} \\ &\quad + p^2\exp\left\{-Cn\left(\frac{t}{D}\right)^2\right\} + p^2\exp\left\{-Cn\left(\frac{t'}{D^{3/2}}\right)^2\right\} + p\exp\left\{-Cn\left(\frac{\epsilon}{D}\right)^2\right\}. \end{aligned} \tag{54}$$

Let $\epsilon = \frac{D^2}{C}\left(\frac{\log p}{n}\right)^{\frac{1-2\varpi}{2}}, t = D\left(\frac{\log p}{n}\right)^{\frac{1-\varpi}{2}}, t' = D^{\frac{3}{2}}\left(\frac{\log p}{n}\right)^{\frac{1-\varpi}{2}}$ for an arbitrary $\varpi \in (0, 1/2)$ and plug them into (53), we obtain that

$$\begin{aligned}&\exp\left\{-\frac{1}{D}\left(\frac{C\epsilon}{t} - D\right)^2\right\} + \exp\left\{-C\left(\frac{\epsilon}{t'}\right)^2\right\} + \exp\{-Cn\epsilon^2\} + p^2\exp\left\{-Cn\left(\frac{t}{D}\right)^2\right\} \\ &\quad + p^2\exp\left\{-Cn\left(\frac{t'}{D^{3/2}}\right)^2\right\} + p\exp\left\{-Cn\left(\frac{\epsilon}{D}\right)^2\right\}\end{aligned}$$

$$\lesssim \exp\left\{-CD^3\left(\frac{n}{\log p}\right)^{\varpi}\right\} + p^2\exp\left\{-Cn^{\varpi}(\log p)^{1-\varpi}\right\} + p\exp\left\{-CDn^{2\varpi}(\log p)^{1-\varpi}\right\}$$
$$+ \exp\left\{-C\left(\frac{n}{\log p}\right)^{\varpi}\right\}$$
$$\ll D^2\left(\frac{\log p}{n}\right)^{\frac{1-2\varpi}{2}},$$

because $\log p \lesssim n^{\varpi_0}$ for some $\varpi_0 \in (0,1)$. Therefore, due to (54), we have

$$\mathbb{E}\sup_{\substack{S:S\supseteq S^*\\|S|\le D}}[R(C_n^S) - R(C_{Bayes})] \lesssim D^2\left(\frac{\log p}{n}\right)^{\frac{1-2\varpi}{2}}.$$

B.16 Proof of Theorem 25

We first prove the following helpful lemma.

Lemma 40 *For $H \sim$ Hypergeometric$(p, \bar{p}^*, d)$, if $\min(d,\bar{p}^*)\cdot\frac{\bar{p}^* d}{p} = o(1)$ and $B_2 \ll \left(\frac{p}{\bar{p}^* d}\right)^{t+1}$, where t is a positive integer no larger than d, then $(\Pr(H \le t))^{B_2} \to 1$ as $p \to \infty$.*

Proof [Proof of Lemma 40] The cumulative distribution function of H satisfies

$$\Pr(H \le t) = 1 - \frac{\binom{d}{t+1}\binom{p-d}{\bar{p}^*-t-1}}{\binom{p}{\bar{p}^*}}\,{}_3F_2\left[\begin{array}{c} 1, t+1-\bar{p}^*, t+1-n \\ t+2, p+t+2-\bar{p}^*-d \end{array};1\right],$$

where ${}_3F_2$ is the generalized hypergeometric function (Abadir, 1999). And we have

$$\begin{aligned}
&{}_3F_2\left[\begin{array}{c} 1, t+1-\bar{p}^*, t+1-n \\ t+2, p+t+2-\bar{p}^*-d \end{array};1\right] \\
&= \sum_{n=0}^{\infty}\frac{1_n(t+1-\bar{p}^*)_n(t+1-d)_n}{(t+2)_n(p+t+2-\bar{p}^*-d)_n}\cdot\frac{1}{n!} \\
&= \sum_{n=0}^{\infty}\frac{(t+1-\bar{p}^*)_n(t+1-d)_n}{(t+2)_n(p+t+2-\bar{p}^*-d)_n} \\
&= 1 + \sum_{n=1}^{\min(d-t,\bar{p}^*-t)}\frac{(\bar{p}^*-t-1)\cdots(\bar{p}^*-t-n)(d-t-1)\cdots(d-t-n)}{(t+2)_n(p+t+2-\bar{p}^*-d)\cdots(p+t+1-\bar{p}^*-d+n)} \\
&\le 1 + \sum_{n=1}^{\min(d-t,\bar{p}^*-t)}\left(\frac{\bar{p}^* d}{p}\right)^n \\
&\le 1 + \min(d,\bar{p}^*)\cdot O\left(\frac{\bar{p}^* d}{p}\right) \\
&\le 1 + o(1),
\end{aligned}$$

where $a_n \coloneqq a(a+1)\cdots(a+n-1), a_0 \coloneqq 1$ for any real number a and positive integer n. On the other hand, we can also see that

$${}_3F_2\left[\begin{array}{c} 1, t+1-\bar{p}^*, t+1-n \\ t+2, p+t+2-\bar{p}^*-d \end{array};1\right] \ge 1.$$

For convenience, denote $\tilde{F} = {}_3F_2\left[\begin{array}{c} 1, t+1-\bar{p}^*, t+1-n \\ t+2, p+t+2-\bar{p}^*-d \end{array}; 1\right]$, then by Taylor expansion, it holds that

$$B_2 \log\left[1 - \frac{\binom{d}{t+1}\binom{p-d}{\bar{p}^*-t-1}}{\binom{p}{\bar{p}^*}}\tilde{F}\right] = -B_2 \cdot \frac{\binom{d}{t+1}\binom{p-d}{\bar{p}^*-t-1}}{\binom{p}{\bar{p}^*}}\tilde{F} + O\left(B_2\left(\frac{\binom{d}{t+1}\binom{p-d}{\bar{p}^*-t-1}}{\binom{p}{\bar{p}^*}}\right)^2\right).$$

In addition, we have

$$\frac{\binom{d}{t+1}\binom{p-d}{\bar{p}^*-t-1}}{\binom{p}{\bar{p}^*}} \le \frac{d^{t+1}\bar{p}^*(\bar{p}^*-1)\cdots(\bar{p}^*-t)}{(t+1)!(p-\bar{p}^*+t+1)\cdots(p-\bar{p}^*+1)} \le \left(\frac{\bar{p}^* d}{p}\right)^{t+1} \cdot \frac{1}{(t+1)!}.$$

Therefore

$$B_2 \cdot \frac{\binom{d}{t+1}\binom{p-d}{\bar{p}^*-t-1}}{\binom{p}{\bar{p}^*}} = o(1), B_2 \cdot \left(\frac{\binom{d}{t+1}\binom{p-d}{\bar{p}^*-t-1}}{\binom{p}{\bar{p}^*}}\right)^2 = o(1),$$

leading to $B_2 \log\left[1 - \frac{\binom{d}{t+1}\binom{p-d}{\bar{p}^*-t-1}}{\binom{p}{\bar{p}^*}}\tilde{F}\right] = o(1)$, which implies the conclusion. ■

Next let's prove the original theorem. Without loss of generality, assume there is a positive constant C such that at the first step of Algorithm 2, B_2 satisfies

$$\frac{CDp^{\bar{p}^*}}{D-\bar{p}^*+1} \le B_2 \ll \left(\frac{p}{\bar{p}^* D}\right)^{\bar{p}^*+1},$$

and at the following steps it follows

$$C \cdot \frac{(D+C_0)^{D+1} p^{\bar{p}^*} (\log p)^{p^*}}{C_0^D (D-p^*+1)} \le B_2 \ll \left(\frac{p}{\bar{p}^* D}\right)^{\bar{p}^*+1}.$$

Notice that condition (i) in Assumption 6 implies that

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{S:|S|\le D} |\mathrm{Cr}_n(S) - \mathrm{Cr}(S)| > M_n \nu(n,p,D)\right) = 0.$$

We will prove the original theorem in three steps. Denote $\eta_j^{(t)} = \mathrm{P}(j \in S_{1*}^{(t)}), \hat{\eta}_j^{(t)} = \frac{1}{B_1}\sum_{i=1}^{B_1} \mathbb{1}(j \in S_{i*}^{(t)}), \tilde{S}^{(t)} = \{j : \hat{\eta}_j^{(t)} > 1/\log p\}, \tilde{S}_*^{(t)} = \tilde{S}^{(t)} \cap S^*, \tilde{p}_t^* = |\tilde{S}_*^{(t)}|$.

(i) Step 1: When $t = 0$, due to the stepwise detectable condition, there exists a subset $S_*^{(0)} \subseteq S^*$ with cardinality $\bar{p}^*$ that satisfies the conditions. On the other hand,

$$\begin{aligned} \mathbf{P}\left(\bigcup_{k=1}^{B_2}\{S_{1k} \supseteq S_*^{(0)}\}\right) &= 1 - \left[1 - \mathbf{P}\left(S_{1k} \supseteq S_*^{(0)}\right)\right]^{B_2} \\ &= 1 - \left[1 - \frac{1}{D}\sum_{\bar{p}^* \le d \le D} \frac{\binom{p-\bar{p}^*}{d-\bar{p}^*}}{\binom{p}{d}}\right]^{B_2} \end{aligned}$$

$$\geq 1 - \left[1 - \frac{D - p^* + 1}{D} \cdot \frac{1}{p^{\bar{p}^*}}\right]^{B_2},$$

where

$$B_2 \log\left(1 - \frac{D - p^* + 1}{D} \cdot \frac{1}{p^{\bar{p}^*}}\right) \leq -B_2 \cdot \frac{D - p^* + 1}{D} \cdot \frac{1}{p^{\bar{p}^*}} \leq -C,$$

yielding that

$$\mathbf{P}\left(\bigcup_{k=1}^{B_2} \{S_{1k}^{(0)} \supseteq S_*^{(0)}\}\right) \geq 1 - e^{-C}.$$

Then we have

$$\begin{aligned}
&\mathrm{P}\left(S_{1*}^{(0)} \supseteq S_*^{(0)}\right) \\
&\geq \mathrm{P}\left(\bigcup_{k=1}^{B_2} \{S_{1k}^{(0)} \supseteq S_*^{(0)}\}, \inf_{\substack{S:|S\cap S^*|\leq \bar{p}^* \\ |S|\leq D}} \mathrm{Cr}_n(S) - \sup_{S:S\cap S^*=S_*^{(0)}} \mathrm{Cr}_n(S) > M_n\nu(n,p,D),\right. \\
&\quad \left.\bigcap_{k=1}^{B_2} \left\{|S_{1k}^{(0)} \cap S^*| \leq \bar{p}^*\right\}\right) \\
&\geq \mathrm{P}\left(\bigcup_{k=1}^{B_2} \{S_{1k}^{(0)} \supseteq S_*^{(0)}\}, \sup_{S:|S|\leq D} |\mathrm{Cr}_n(S) - \mathrm{Cr}(S)| \leq \frac{1}{2} M_n\nu(n,p,D), \bigcap_{k=1}^{B_2} \left\{|S_{1k}^{(0)} \cap S^*| \leq \bar{p}^*\right\}\right) \\
&\geq \mathbf{P}\left(\bigcup_{k=1}^{B_2} \{S_{1k}^{(0)} \supseteq S_*^{(0)}\}\right) - \mathbb{P}\left(\sup_{S:|S|\leq D} |\mathrm{Cr}_n(S) - \mathrm{Cr}(S)| > \frac{1}{2} M_n\nu(n,p,D)\right) \\
&\quad - \mathbf{P}\left(\bigcup_{k=1}^{B_2} \left\{|S_{1k}^{(0)} \cap S^*| \geq \bar{p}^* + 1\right\}\right) \\
&\geq 1 - \frac{3}{2} e^{-C}, \qquad (55)
\end{aligned}$$

as n is sufficiently large. The last inequality holds because there exists a variable $H \sim \text{Hypergeometric}(p, \bar{p}^*, D)$, such that

$$\begin{aligned}
\mathbf{P}\left(\bigcup_{k=1}^{B_2} \left\{|S_{1k}^{(0)} \cap S^*| \geq \bar{p}^* + 1\right\}\right) &= 1 - \left[1 - \mathbf{P}\left(|S_{1k}^{(0)} \cap S^*| \geq \bar{p}^* + 1\right)\right]^{B_2} \\
&= 1 - \left[1 - \frac{1}{D} \sum_{1\leq d\leq D} \mathbf{P}\left(|S_{1k}^{(0)} \cap S^*| \geq \bar{p}^* + 1 || S_{1k}^{(0)}| = d\right)\right]^{B_2} \\
&\leq 1 - \left[1 - \frac{1}{D} \sum_{1\leq d\leq D} \mathbf{P}\left(|S_{1k}^{(0)} \cap S^*| \geq \bar{p}^* + 1 || S_{1k}^{(0)}| = D\right)\right]^{B_2}
\end{aligned}$$

$$= \mathbf{P}\left(\bigcup_{k=1}^{B_2}\left\{|S_{1k}^{(0)} \cap S^*| \geq \bar{p}^* + 1\right\} \Big| \bigcap_{k=1}^{B_2}\left\{|S_{1k}^{(0)}| = D\right\}\right)$$

$$= 1 - \mathrm{P}(H \leq \bar{p}^*)$$
$$\to 0$$

due to Lemma 40.

Then for any $j \in S_*^{(0)}$, it holds that

$$\mathrm{P}\left(j \in S_{1*}^{(0)}\right) \geq \mathrm{P}\left(S_{1*}^{(0)} \supseteq S_*^{(0)}\right) \geq 1 - \frac{3}{2}e^{-C}.$$

And by Hoeffding's inequality (Petrov, 2012):

$$\mathrm{P}\left(\hat{\eta}_j^{(1)} \geq 1 - 2e^{-C}\right) \geq \mathrm{P}\left(\hat{\eta}_j^{(1)} - \eta_j^{(1)} > -\frac{1}{2}e^{-C}\right) \geq 1 - \exp\left\{-2B_1 \cdot \frac{1}{4}e^{-2C}\right\}.$$

Following by union bounds, we can obtain

$$\mathrm{P}\left(\bigcap_{j \in S_*^{(0)}}\left\{\hat{\eta}_j^{(1)} \geq 1 - 2e^{-C}\right\}\right) \geq 1 - p^* \exp\left\{-\frac{1}{2}B_1 e^{-2C}\right\}.$$

(ii) Step 2: When $t = 1$, let's first condition on some specific $\tilde{S}^{(0)}$ defined before as $\{j : \hat{\eta}_j^{(0)} > C_0/\log p\}$ satisfying $|\tilde{S}_*^{(0)}| = \bar{p}^*$. Later we will take the expectation to get the inequality without conditioning. To simplify and distinguish the notations, we omit the condition mentioned above and denote the new corresponding conditional probabilities as $\mathrm{P}_c, \mathbf{P}_c, \mathbb{P}_c$.

For any specific $\tilde{S}_*^{(0)}$, due to the stepwise detectable condition again, there exists a subset $S_*^{(1)} \subseteq S^*$ with cardinality $\bar{p}^*$ satisfies the conditions. Similar to step 1, we have

$$\mathbf{P}_c\left(\bigcup_{k=1}^{B_2}\left\{S_{1k}^{(1)} \supseteq \tilde{S}_*^{(0)} \cup S_*^{(1)}\right\}\right) \tag{56}$$
$$= 1 - \left[1 - \frac{1}{D}\sum_{\tilde{p}_1^* + \bar{p}^* \leq d \leq D} \mathbf{P}_c\left(S_{1k}^{(1)} \supseteq \tilde{S}_*^{(0)} \cup S_*^{(1)} \big| |S_{1k}^{(1)}| = d\right)\right]^{B_2},$$

where

$$\mathbf{P}_c\left(S_{1k}^{(1)} \supseteq \tilde{S}_*^{(0)} \cup S_*^{(1)} \big| |S_{1k}^{(1)}| = d\right)$$
$$\geq \mathbf{P}_c\left(S_{1k}^{(1)} \supseteq S_*^{(1)} \big| |S_{1k}^{(1)}| = d, S_{1k}^{(1)} \supseteq \tilde{S}_*^{(0)}, S_{1k}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset\right)$$
$$\times \mathbf{P}_c\left(S_{1k}^{(1)} \supseteq \tilde{S}_*^{(0)}, S_{1k}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset \big| |S_{1k}^{(1)}| = d\right). \tag{57}$$

For convenience, let's consider a series $(j_1, \ldots, j_d)$ sampled from $\{1, \ldots, p\}$ without replacement with sampling weight $(\hat{\eta}_1^{(1)}, \ldots, \hat{\eta}_p^{(1)})$ in this order. We use $\mathrm{P}_c(j_1, \ldots, j_d || S_{1k}^{(1)}| = d)$ to represent the corresponding probability and use $\mathrm{P}_c(j_i | j_1, \ldots, j_{i-1}, |S_{1k}^{(1)}| = d)$ to represent the conditional probability for sampling j_i given $j_1, \ldots, j_{i-1}$.

Based on the notations defined above, there holds

$$\begin{aligned}
&\mathbf{P}_c\left(S_{1k}^{(1)} \supseteq \tilde{S}_*^{(0)}, S_{1k}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset \big| |S_{1k}^{(1)}| = d\right) \\
&= \sum_{\substack{(j_1,\ldots,j_d) \supseteq \tilde{S}_*^{(0)} \\ (j_1,\ldots,j_d) \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset}} \mathbf{P}_c\left(j_1, \ldots, j_d \big| |S_{1k}^{(1)}| = d\right) \\
&= \sum_{\substack{(j_1,\ldots,j_d) \supseteq \tilde{S}_*^{(0)} \\ (j_1,\ldots,j_d) \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset}} \mathbf{P}_c\left(j_1 \big| |S_{1k}^{(1)}| = d\right) \mathbf{P}_c\left(j_2 | j_1, |S_{1k}^{(1)}| = d\right) \\
&\quad \cdots \mathbf{P}_c(j_d | j_1, \ldots, j_{d-1}, |S_{1k}^{(1)}| = d). \qquad (58)
\end{aligned}$$

Since for $j_i \in \tilde{S}_*^{(0)}$, it holds

$$\begin{aligned}
\mathbf{P}_c(j_i | j_1, \ldots, j_{i-1}, |S_{1k}^{(1)}| = d) &\geq \mathbf{P}_c(j_i || S_{1k}^{(1)}| = d) \\
&\geq \frac{\hat{\eta}_{j_i}^{(0)}}{\sum_{j \in \tilde{S}_*^{(0)}} \hat{\eta}_j^{(0)} + \sum_{j \notin \tilde{S}_*^{(0)}} \frac{C_0}{p}} \\
&\geq \frac{C_0}{(D + C_0) \log p}.
\end{aligned}$$

And for $j_i \in \{1, \ldots, p\} \backslash \tilde{S}^{(0)}$, it holds

$$\begin{aligned}
\mathbf{P}_c(j_i | j_1, \ldots, j_{i-1}, |S_{1k}^{(1)}| = d) &\geq \mathbf{P}_c(j_i || S_{1k}^{(1)}| = d) \\
&\geq \frac{\hat{\eta}_{j_i}^{(0)}}{\sum_{j \in \tilde{S}_*^{(0)}} \hat{\eta}_j^{(0)} + \sum_{j \notin \tilde{S}_*^{(0)}} \frac{C_0}{p}} \\
&\geq \frac{C_0}{(D + C_0) p}.
\end{aligned}$$

Therefore by plugging these inequalities into (58), it yields that when $d \geq \tilde{p}_*^{(0)}$,

$$\begin{aligned}
&\mathbf{P}_c\left(S_{1k}^{(1)} \supseteq \tilde{S}_*^{(0)}, S_{1k}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset \big| |S_{1k}^{(1)}| = d\right) \\
&\geq \binom{p - |\tilde{S}^{(0)}|}{d - \tilde{p}_*^{(0)}} \binom{d}{\tilde{p}_*^{(0)}} \tilde{p}_*^{(0)}! (d - \tilde{p}_*^{(0)})! \cdot \left(\frac{C_0}{(D + C_0) \log p}\right)^{\tilde{p}_*^{(0)}} \cdot \left(\frac{C_0}{(D + C_0) p}\right)^{d - \tilde{p}_*^{(0)}} \\
&= \left(1 - \frac{|\tilde{S}^{(0)}|}{p}\right) \cdots \left(1 - \frac{|\tilde{S}^{(0)}| - \tilde{p}_*^{(0)} + d - 1}{p}\right) \cdot d(d-1) \cdots (d - \tilde{p}_*^{(0)} + 1) \\
&\quad \cdot \frac{C_0^d}{(D + C_0)^d (\log p)^{\tilde{p}_*^{(0)}}}
\end{aligned}$$

$$\geq \left(1 - \frac{D \log p}{C_0 p}\right)^{d - \tilde{p}_*^{(0)}} \cdot \frac{\tilde{p}_*^{(0)}!}{(\log p)^{\tilde{p}_*^{(0)}}} \cdot \left(\frac{C_0}{D + C_0}\right)^d$$
$$\geq \frac{1}{(\log p)^{p^*}} \cdot \left(\frac{C_0}{D + C_0}\right)^D, \tag{59}$$

when n is sufficiently large. In the last second inequality, we used the fact that $|\tilde{S}^{(0)}| \leq \frac{D \log p}{C_0}$. On the other hand, given $|S_{1k}^{(1)}| = d, S_{1k}^{(1)} \supseteq \tilde{S}_*^{(0)}, S_{1k}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset$, the indicators of whether the remaining features are sampled out or not follow the restricted multinomial distribution with parameters $(p - |\tilde{S}^{(0)}|, d - \tilde{p}_*^{(0)}, \mathbf{1})$, which means that the remaining variables have the same sampling weights. Then for $d \geq \tilde{p}_*^{(0)} - \bar{p}^*$,

$$\begin{aligned}
&\mathbf{P}_c\left(S_{1k}^{(1)} \supseteq S_*^{(1)} \big| |S_{1k}^{(1)}| = d, S_{1k}^{(1)} \supseteq \tilde{S}_*^{(0)}, S_{1k}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset\right) \\
&= \frac{\binom{p - |\tilde{S}^{(0)}| - \bar{p}^*}{d - \tilde{p}_*^{(0)} - \bar{p}^*}}{\binom{p - |\tilde{S}^{(0)}|}{d - \tilde{p}_*^{(0)}}} \\
&\geq \left(\frac{\max(d - \tilde{p}_*^{(0)} - \bar{p}^*, 1)}{p - |\tilde{S}^{(0)}|}\right)^{\bar{p}^*} \\
&\geq \frac{1}{p^{\bar{p}^*}},
\end{aligned}$$

which combined with (56), (57) and (59) leads to

$$\begin{aligned}
&\mathbf{P}_c\left(\bigcup_{k=1}^{B_2} \left\{S_{1k}^{(1)} \supseteq \tilde{S}_*^{(0)} \cup S_*^{(1)}\right\}\right) \\
&\geq 1 - \left[1 - \frac{1}{D} \sum_{\tilde{p}_1^* + \bar{p}^* \leq d \leq D} \frac{1}{(\log p)^{p^*}} \cdot \left(\frac{C_0}{D + C_0}\right)^D \cdot \frac{1}{p^{\bar{p}^*}}\right]^{B_2} \\
&\geq 1 - \frac{1}{2} e^{-C},
\end{aligned}$$

since $B_2 \geq C \cdot \frac{(D + C_0)^{D+1} p^{\bar{p}^*} (\log p)^{p^*}}{C_0^D (D - p^* + 1)}$.

Furthermore,

$$\mathbf{P}_c\left(\bigcup_{k=1}^{B_2} \left\{|S_{1k}^{(1)} \cap (S^* \backslash \tilde{S}_*^{(0)})| \geq \bar{p}^* + 1\right\}\right) = 1 - \left[1 - \mathbf{P}_c\left(|S_{11}^{(1)} \cap (S^* \backslash \tilde{S}_*^{(0)})| \geq \bar{p}^* + 1\right)\right]^{B_2},$$

where

$$\begin{aligned}
&\mathbf{P}_c\left(|S_{11}^{(1)} \cap (S^* \backslash \tilde{S}_*^{(0)})| \geq \bar{p}^* + 1\right) \\
&= \mathbf{P}_c\left(|S_{11}^{(1)} \cap (S^* \backslash \tilde{S}_*^{(0)})| \geq \bar{p}^* + 1 \big| S_{11}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset\right) \mathbf{P}_c\left(S_{11}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset\right)
\end{aligned}$$

$$+ \mathrm{P}_c\left(|S_{11}^{(1)} \cap (S^* \backslash \tilde{S}_*^{(0)})| \geq \bar{p}^* + 1 \middle| S_{11}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) \neq \emptyset\right) \mathrm{P}_c\left(S_{11}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) \neq \emptyset\right)$$

$$\leq \mathrm{P}_c\left(|S_{11}^{(1)} \cap (S^* \backslash \tilde{S}_*^{(0)})| \geq \bar{p}^* + 1 \middle| S_{11}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset\right)$$
$$= \frac{1}{D} \sum_{1 \leq d \leq D} \mathrm{P}_c\left(|S_{11}^{(1)} \cap (S^* \backslash \tilde{S}_*^{(0)})| \geq \bar{p}^* + 1 \middle| S_{11}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset, |S_{11}^{(1)}| = d\right)$$
$$\leq \frac{1}{D} \sum_{1 \leq d \leq D} \mathrm{P}\left(|S_{11}^{(1)} \cap (S^* \backslash \tilde{S}_*^{(0)})| \geq \bar{p}^* + 1 \middle| S_{11}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset, |S_{11}^{(1)}| = D\right)$$
$$= \mathrm{P}_c\left(|S_{11}^{(1)} \cap (S^* \backslash \tilde{S}_*^{(0)})| \geq \bar{p}^* + 1 \middle| S_{11}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset, |S_{11}^{(1)}| = D\right).$$

Similar to step 1, because of Lemma 40, it follows that

$$\mathrm{P}_c\left(\bigcup_{k=1}^{B_2} \left\{|S_{1k}^{(1)} \cap (S^* \backslash \tilde{S}_*^{(0)})| \geq \bar{p}^* + 1\right\}\right)$$
$$\geq 1 - \left[1 - \mathrm{P}_c\left(|S_{11}^{(1)} \cap (S^* \backslash \tilde{S}_*^{(0)})| \geq \bar{p}^* + 1 \middle| S_{11}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset, |S_{11}^{(1)}| = D\right)\right]^{B_2}$$
$$= \mathrm{P}_c\left(\bigcup_{k=1}^{B_2} \left\{|S_{1k}^{(1)} \cap (S^* \backslash \tilde{S}_*^{(0)})| \geq \bar{p}^* + 1\right\} \middle| \bigcap_{k=1}^{B_2} \left\{S_{1k}^{(1)} \cap (\tilde{S}^{(0)} \backslash \tilde{S}_*^{(0)}) = \emptyset, |S_{1k}^{(1)}| = D\right\}\right)$$

$$= o(1),$$

uniformly for any $\tilde{S}^{(0)}$. Then similar to (55), we have for $j \in \tilde{S}_*^{(0)} \cup S_*^{(1)}$:

$$\eta_j^{(1)} = \mathrm{P}_c\left(j \in S_{1*}^{(1)}\right) \geq \mathrm{P}_c\left(S_{1*}^{(1)} \supseteq \tilde{S}_*^{(0)} \cup S_*^{(1)}\right) \geq 1 - \frac{3}{2}e^{-C}.$$

And by Hoeffding's inequality (Petrov, 2012):

$$\mathrm{P}\left(\hat{\eta}_j^{(1)} \geq 1 - 2e^{-C} \middle| \tilde{S}^{(0)}\right) \geq \mathrm{P}\left(\hat{\eta}_j^{(1)} - \eta_j^{(1)} > -\frac{1}{2}e^{-C}\right) \geq 1 - \exp\left\{-2B_1 \cdot \frac{1}{4}e^{-2C}\right\}.$$

By union bounds, it holds that

$$\mathrm{P}\left(\bigcap_{j \in \tilde{S}_*^{(0)} \cup S_*^{(1)}} \left\{\hat{\eta}_j^{(1)} \geq 1 - 2e^{-C}\right\} \middle| \tilde{S}^{(0)}\right) \geq 1 - p^* \exp\left\{-\frac{1}{2}B_1 e^{-2C}\right\}.$$

Since the above conclusions hold for any $\tilde{S}^{(0)}$ and $|\tilde{S}_*^{(0)}| \geq |S_*^{(0)}| = \bar{p}^*$, we can conclude that

$$\mathrm{P}\left(\sum_{j=1}^{p} \mathbb{1}(\hat{\eta}_j^{(1)} \geq 1 - 2e^{-C}) \geq 2\bar{p}^*\right)$$

$$\geq \mathrm{E}_{\tilde{S}^{(0)}}\left[\mathrm{P}\left(\bigcap_{j\in\tilde{S}_*^{(0)}\cup S_*^{(1)}}\left\{\hat{\eta}_j^{(1)}\geq 1-2e^{-C}\right\}\middle|\tilde{S}^{(0)}\right)\middle||S_*^{(0)}|=\bar{p}^*\right]$$
$$\geq \left(1-p^*\exp\left\{-\frac{1}{2}B_1e^{-2C}\right\}\right)\left(1-p^*\exp\left\{-\frac{1}{2}B_1e^{-2C}\right\}\right)$$
$$\geq 1-2p^*\exp\left\{-\frac{1}{2}B_1e^{-2C}\right\}.$$

After the the second iteration step, with probability $1-2p^*\exp\left\{-\frac{1}{2}B_1e^{-2C}\right\}$, there will be at least $2\bar{p}^*$ features of S^* covered in $\tilde{S}^{(1)}$ and having $\hat{\eta}_j > 1-2e^{-C}$. Similarly, after the the t-th iteration step, there will be at least $(t+1)\bar{p}^*$ features of S^* covered in $\tilde{S}^{(t)}$ and having $\hat{\eta}_j > 1-2e^{-C}$.

(iii) Step 3: By step 2, after at most $t' = \lceil\frac{p^*}{\bar{p}^*}\rceil - 1$ iterations, $\tilde{S}^{(t')}$ will cover S^*. Without loss of generality, let's assume the smallest t' satisfying $\tilde{S}^{(t')}\supseteq S^*$ equal to $\lceil\frac{p^*}{\bar{p}^*}\rceil$. Then when the iteration number $t=\lceil\frac{p^*}{\bar{p}^*}\rceil$, we have

$$\mathrm{P}\left(\bigcap_{j\in S^*}\left\{\hat{\eta}_j^{(t-1)}\geq 1-2e^{-C}\right\}\right)\geq 1-tp^*\exp\left\{-\frac{1}{2}B_1e^{-2C}\right\}.$$

Using the same notations defined at the beginning of step 2 (notice that here we only need to condition on $\bigcap_{j\in S^*}\{\hat{\eta}_j^{(t-1)}\geq 1-2e^{-C}\}$), we have

$$\begin{aligned}&\mathbf{P}_c\left(S_{1k}^{(t)}\supseteq S^*\right)\\ &\geq \frac{1}{D}\sum_{p^*\leq d\leq D}\sum_{(j_1,\ldots,j_d)\supseteq S^*}\mathbf{P}_c(j_1,\ldots,j_d||S_{1k}^{(t)}|=d)\\ &=\frac{1}{D}\sum_{p^*\leq d\leq D}\sum_{(j_1,\ldots,j_d)\supseteq S^*}\mathbf{P}_c\left(j_1\middle||S_{1k}^{(t)}|=d\right)\cdots\mathbf{P}_c\left(j_d|j_1,\ldots,j_{d-1},|S_{1k}^{(t)}|=d\right).\end{aligned}$$

Similar to step 2, for $j_i\in S^*$, it holds

$$\mathbf{P}_c(j_i|j_1,\ldots,j_{i-1},|S_{1k}^{(t)}|=d)\geq\mathbf{P}_c(j_i||S_{1k}^{(t)}|=d)\geq\frac{\hat{\eta}_{j_i}^{(t-1)}}{\sum_{j\in S^*}\hat{\eta}_j^{(t-1)}+\sum_{j\notin S^*}\frac{C_0}{p}}\geq\frac{1-2e^{-C}}{D+C_0}.$$

And for $j_i\in\{1,\ldots,p\}\backslash S^*$, it holds

$$\mathbf{P}_c(j_i|j_1,\ldots,j_{i-1},|S_{1k}^{(t)}|=d)\geq\mathbf{P}_c(j_i||S_{1k}^{(t)}|=d)\geq\frac{\hat{\eta}_{j_i}^{(1)}}{\sum_{j\in S^*}\hat{\eta}_j^{(t)}+\sum_{j\notin S^*}\frac{C_0}{p}}\geq\frac{C_0}{(D+C_0)p}.$$

Thus, we have

$$\mathbf{P}_c\left(S_{1k}^{(t)}\supseteq S^*\right)\geq\frac{1}{D}\sum_{p^*\leq d\leq D}\sum_{(j_1,\ldots,j_d)\supseteq S^*}\left(\frac{1-2e^{-C}}{D+C_0}\right)^{p^*}\cdot\left(\frac{C_0}{(D+C_0)p}\right)^{d-p^*}$$

$$
\begin{aligned}
&\geq \frac{1}{D}\sum_{p^*\leq d\leq D}\binom{p-p^*}{d-p^*}d!\left(\frac{1-2e^{-C}}{D+C_0}\right)^{p^*}\cdot\left(\frac{C_0}{(D+C_0)p}\right)^{d-p^*}\\
&\geq \frac{D-p^*+1}{D}\cdot\left(\frac{1-2e^{-C}}{D+C_0}\right)^{p^*}\cdot\left(\frac{C_0}{D+C_0}\right)^{D-p^*}\cdot\left(1-\frac{D-1}{p}\right)^{D-p^*}\\
&= \frac{(D-p^*+1)(1-2e^{-C})^{p^*}C_0^{D-p^*}}{D(D+C_0)^D}\cdot\left(1-\frac{D-1}{p}\right)^{D-p^*},
\end{aligned}
$$

leading to

$$
\begin{aligned}
&\mathbf{P}_c\left(\bigcup_{k=1}^{B_2}\left\{S_{1k}^{(t)}\supseteq S^*\right\}\right)\\
&=1-\left[1-\mathbf{P}_c\left(S_{1k}^{(t)}\supseteq S^*\right)\right]^{B_2}\\
&\geq 1-\exp\left\{-B_2\cdot\frac{(D-p^*+1)(1-2e^{-C})^{p^*}C_0^{D-p^*}}{D(D+C_0)^D}\cdot\left(1-\frac{D-1}{p}\right)^{D-p^*}\right\}\\
&\geq 1-\exp\left\{-C\cdot\frac{(D+C_0)p^{\bar{p}^*}}{DC_0^{p^*}}\cdot\left(1-\frac{D-1}{p}\right)^{D-p^*}\right\}.
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
&\mathbf{P}\left(\bigcup_{k=1}^{B_2}\left\{S_{1k}^{(t)}\supseteq S^*\right\}\right)\\
&\geq \mathbf{P}_c\left(S_{1k}^{(t)}\supseteq S^*\right)\mathrm{P}\left(\bigcap_{j\in S^*}\left\{\hat{\eta}_j^{(t-1)}\geq 1-2e^{-C}\right\}\right)\\
&\geq 1-\exp\left\{-C\cdot\frac{(D+C_0)^{p^*+1}p^{\bar{p}^*}}{DC_0^{p^*}}\cdot\left(1-\frac{D-1}{p}\right)^{D-p^*}\right\}-p^*\left\lceil\frac{p^*}{\bar{p}^*}\right\rceil\exp\left\{-\frac{1}{2}B_1e^{-2C}\right\}.
\end{aligned}
$$

Then similar to (55), there holds

$$
\begin{aligned}
\mathrm{P}(S_{1*}^{(t)}\supseteq S^*)&\geq \mathrm{P}\left(\bigcup_{k=1}^{B_2}\{S_{1k}\supseteq S_*^{(t)}\},\inf_{\substack{S:S\not\supseteq S^*\\|S|\leq D}}\mathrm{Cr}_n(S)-\sup_{S:S\supseteq S^*}\mathrm{Cr}_n(S)>M_n\nu(n,p,D)\right)\\
&\geq \mathrm{P}\left(\bigcup_{k=1}^{B_2}\{S_{1k}\supseteq S_*^{(t)}\},\sup_{S:|S|\leq D}|\mathrm{Cr}_n(S)-\mathrm{Cr}(S)|\leq\frac{1}{2}M_n\nu(n,p,D)\right)\\
&\geq 1-\mathbf{P}\left(\bigcup_{k=1}^{B_2}\left\{S_{1k}^{(t)}\supseteq S^*\right\}\right)-\mathbb{P}\left(\sup_{S:|S|\leq D}|\mathrm{Cr}_n(S)-\mathrm{Cr}(S)|>\frac{1}{2}M_n\nu(n,p,D)\right).
\end{aligned}
$$

Combined with all the conclusions above, as $n, B_1, B_2 \to \infty$, we get

$$\begin{aligned}\mathrm{P}(S_{1*}^{(t)} \not\supseteq S^*) \leq & \exp\left\{-C \cdot \frac{(D+C_0)^{p^*+1}p^{\bar{p}^*}}{DC_0^{p^*}} \cdot \left(1-\frac{D-1}{p}\right)^{D-p^*}\right\} \\ & + p^* \left\lceil \frac{p^*}{\bar{p}^*} \right\rceil \exp\left\{-\frac{1}{2}B_1 e^{-2C}\right\} \\ & + \mathbb{P}\left(\sup_{S:|S|\leq D} |\mathrm{Cr}_n(S) - \mathrm{Cr}(S)| > \frac{1}{2}M_n \nu(n,p,D)\right) \\ \to & \ 0.\end{aligned}$$

And for $t > \lceil \frac{p^*}{\bar{p}^*} \rceil$, the same conclusion can be obtained by following the same procedure, which completes our proof.

References

K. M. Abadir. An introduction to hypergeometric functions for economists. *Econometric Reviews*, 18(3):287–330, 1999.

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1948.

H. Ahn, H. Moon, M. J. Fazzari, N. Lim, J. J. Chen, and R. L. Kodell. Classification by ensembles from random partitions of high-dimensional data. *Computational Statistics & Data Analysis*, 51(12):6166–6179, 2007.

H. Akaike. Information theory and an extension of the maximum likelihood principle. Proceeding of IEEE international symposium on information theory, 1973.

S. D. Bay. Combining nearest neighbor classifiers through multiple feature subsets. In *ICML*, volume 98, pages 37–45. Citeseer, 1998.

T. B. Berrett and R. J. Samworth. Efficient two-sample functional estimation and the super-oracle phenomenon. *arXiv preprint arXiv:1904.09347*, 2019.

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

P. J. Bickel, E. Levina, et al. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

R. Blaser and P. Fryzlewicz. Random rotation ensembles. *The Journal of Machine Learning Research*, 17(1):126–151, 2016.

R. Blaser and P. Fryzlewicz. Regularizing axis-aligned ensembles via data rotations that favor simpler learners, 2019.

T. Boot and D. Nibbering. Subspace methods. In *Macroeconomic Forecasting in the Era of Big Data*, pages 267–291. Springer, 2020.

L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

L. Breiman. Pasting small votes for classification in large databases and on-line. *Machine learning*, 36(1-2):85–103, 1999.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

M. Brookes. The matrix reference manual. *Imperial College London*, 3, 2005.

R. Bryll, R. Gutierrez-Osuna, and F. Quek. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern recognition*, 36(6):1291–1302, 2003.

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

K. P. Burnham and D. R. Anderson. Practical use of the information-theoretic approach. In *Model selection and inference*, pages 75–117. Springer, 1998.

T. I. Cannings and R. J. Samworth. Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):959–1035, 2017.

J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

J. Chen and Z. Chen. Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, pages 555–574, 2012.

L. Devroye and T. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

D. Dua and C. Graff. Uci machine learning repository. school of information and computer science, university of california, irvine, ca, 2019.

R. J. Durrant and A. Kabán. Random projections as regularizers: learning a linear discriminant from fewer observations than dimensions. *Machine Learning*, 99(2):257–286, 2015.

B. Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, 1975.

C.-G. Esseen. A moment inequality with an application to the central limit theorem. *Scandinavian Actuarial Journal*, 1956(2):160–170, 1956.

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

J. Fan, Y. Feng, J. Jiang, and X. Tong. Feature augmentation via nonparametrics and selection (fans) in high-dimensional classification. *Journal of the American Statistical Association*, 111(513):275–287, 2016.

Y. Fan and C. Y. Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552, 2013.

Y. Fan, Y. Kong, D. Li, Z. Zheng, et al. Innovated interaction screening for high-dimensional nonlinear classification. *The Annals of Statistics*, 43(3):1243–1272, 2015.

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

E. Fix. *Discriminatory analysis: nonparametric discrimination, consistency properties*. USAF school of Aviation Medicine, 1951.

Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

S. Ganguly, J. Ryu, Y.-H. Kim, Y.-K. Noh, and D. D. Lee. Nearest neighbor density functional estimation based on inverse laplace transform. *arXiv preprint arXiv:1805.08342*, 2018.

N. García-Pedrajas and D. Ortiz-Boyer. Boosting random subspace method. *Neural Networks*, 21(9):1344–1362, 2008.

I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the nips 2003 feature selection challenge. In *Advances in neural information processing systems*, pages 545–552, 2005.

N. Hao, Y. Feng, and H. H. Zhang. Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, 113(522):615–625, 2018.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

C. Higuera, K. J. Gardiner, and K. J. Cios. Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. *PloS one*, 10(6), 2015.

T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

D. Hsu, S. Kakade, T. Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

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

R. Kohavi, G. H. John, et al. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

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

Y. Li and J. S. Liu. Robust variable and interaction selection for logistic regression and general index models. *Journal of the American Statistical Association*, 114(525):271–286, 2019.

M. E. Lopes. Estimating a sharp convergence bound for randomized ensembles. *Journal of Statistical Planning and Inference*, 204:35–44, 2020.

M. E. Lopes et al. Estimating the algorithmic variance of randomized ensembles via the bootstrap. *The Annals of Statistics*, 47(2):1088–1112, 2019.

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

Q. Mai, Y. Yang, and H. Zou. Multiclass sparse discriminant analysis. *arXiv preprint arXiv:1504.05845*, 2015.

G. J. McLachlan. Mahalanobis distance. *Resonance*, 4(6):20–26, 1999.

M. Mukhopadhyay and D. B. Dunson. Targeted random projection for prediction from high-dimensional features. *Journal of the American Statistical Association*, pages 1–13, 2019.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, nov 2012. *URL http://www2. imm. dtu. dk/pubdb/p. php*, 3274:14, 2012.

V. V. Petrov. *Sums of independent random variables*, volume 82. Springer Science & Business Media, 2012.

R. Rao and Y. Wu. A strongly consistent procedure for model selection in a regression problem. *Biometrika*, 76(2):369–374, 1989.

L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

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

M. Skurichina and R. P. Duin. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications*, 5(2):121–135, 2002.

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

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

H. Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, 2009.

Q. Wang, S. R. Kulkarni, and S. Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

Q. Zhang and H. Wang. On bic's selection consistency for discriminant analysis. *Statistica Sinica*, pages 731–740, 2011.