Electronic Journal of Statistics
Vol. 16 (2022) 3343–3384
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2022

The robust nearest shrunken centroids classifier for high-dimensional heavy-tailed data*

Shaokang Ren and Qing Mai

Department of Statistics, Florida State University, Tallahassee, Florida 32306, U.S.A.
e-mail: sr17k@fsu.edu; qmai@fsu.edu

Abstract: The nearest shrunken centroids classifier (NSC) is a popular high-dimensional classifier. However, it is prone to inaccurate classification when the data is heavy-tailed. In this paper, we develop a robust generalization of NSC (RNSC) which remains effective under such circumstances. By incorporating the Huber loss both in the estimation and the calculation of the score function, we reduce the impacts of heavy tails. We rigorously show the variable selection, estimation, and prediction consistency in high dimensions under weak moment conditions. Empirically, our proposal greatly outperforms NSC and many other successful classifiers when data is heavy-tailed while remaining comparable to NSC in the absence of heavy tails. The favorable performance of RNSC is also demonstrated in a real data example.

MSC2020 subject classifications: Primary 62H30; secondary 62J07.
Keywords and phrases: Heavy-tailed data, high-dimensional classification, Huber loss, robust estimator, nearest shrunken centroids classifier.

Received April 2021.

Contents

* This project was supported in part by the grant CCF-1908969 from the U.S. National Science Foundation.

1. Introduction

The nearest shrunken centroids classifier (NSC, [49]) is a popular high-dimensional classifier [50, 24, e.g]. It assigns each observation to the class with the nearest centroid, i.e, the within-class mean. In high dimensions, NSC performs variable selection by shrinking the centroid estimates. Thanks to its simplicity, NSC is extremely interpretable and computationally efficient. Moreover, it is observed to achieve remarkable accuracy on many benchmark datasets.

However, heavy-tailed data have been attracting much attention in recent years, as such data frequently arise in many fields, such as biometry, ecological systems, finance, and sociology [43, 39, 47, 53]. NSC is vulnerable in this situation. For one thing, NSC estimates the centroid of each class with the sample mean, but the sample mean is easily impacted by a few outliers, and tends to be inaccurate for heavy-tailed data. In theory, the concentration property for the sample mean is usually proved by assuming the predictors to be sub-Gaussian or sub-exponential, but such assumptions are not appropriate for heavy-tailed data. For the other, NSC measures the distance between the observation and each centroid with the squared Euclidean distance. The quadratic form of this distance amplifies the influence of a few extreme values as well.

To tackle this challenge, we propose the robust nearest shrunken centroids method (RNSC) that is suitable for potentially heavy-tailed data. RNSC gains robustness by combining the Huber loss and Huber estimators with the original NSC method. The impacts of outliers are mitigated, while the resulting classifier continues to be sparse and interpretable. In theory, RNSC is consistent when only the fourth moment exists, which ensures the applicability of RNSC on a wide range of data. In numerical studies, RNSC exhibits better performance than many existing methods when heavy tails or outliers are present, and is comparable to NSC when data are Gaussian.

RNSC is a unique addition to the robust high-dimensional classification literature, even though there exist a few pioneering proposals for this topic. For example, [23] proposed the component-wise median estimators in a high-dimensional scheme. The properties of median-based classifier were also discussed by [21] and [28]. Although the median is robust and can be applied to heavy-tailed data, it is likely to suffer from efficiency loss. In contrast, RNSC balances between the mean and the median estimates depending on how heavy-tailed the data is. Other examples for robust classifiers include [6] and [20], but it is unclear if similar frameworks can be established for more robust versions of NSC.

RNSC is motivated by recent advancements in robust statistics, especially those employing the Huber loss. For example, [31, 19] combined the Huber loss with adaptive LASSO penalty to obtain robust estimation. [48] gave a sharp phase transition for Huber estimators of regression parameters. [2] further looked into the properties and advantages of the Huber estimator when estimating covariance and precision matrices in the high-dimensional case. [58] studied the theoretical properties and the application of the Huber estimator for the high-dimensional multiple testing problem. [34] showed a general result on the theoretical properties of robust M-estimators in the high-dimensional scenario where the data is disturbed by heavy-tailed distributions or outliers. Our proposal of RNSC further generalizes these works to classification problems.

We also note that there exist many high-dimensional classification methods. However, most of them do not consider heavy-tailed data. For example, there are extensions of NSC [52, 46, 51, 12], logistic regression [24, 40, e.g], linear discriminant analysis [15, 13, 56, 17, 38, 36, e.g], support vector machine [4, 5, 57, e.g], among others. Most of these methods with theoretical justifications require tail conditions such as sub-Gaussian or sub-exponential, and thus may not be appropriate for heavy-tailed data.

The rest of this article is organized as follows. We propose RNSC in Section 2. The theoretical results of RNSC are discussed in Section 3. In Section 4, we present the simulation studies in different heavy-tailed or outlier scenarios to further reveal the advantages of our proposal. In Section 5, we give a real-data example for RNSC. The proof of theorems is in the Appendix.

2. Methodology

2.1. The nearest shrunken centroids method

We first briefly review the nearest shrunken centroids (NSC) classifier. In a classification problem, we have a pair of random variables, $(Y, \mathbf{X})$, where the predictors $\mathbf{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ and the response $Y \in \{1, \ldots, K\}$, with K being a positive integer. Let $\Pr(Y = k) = \pi_k$ be the prior probability that an observation belongs to Class k. We want a prediction on Y based on the information from $\mathbf{X}$. The classical nearest centroids method assigns an observation to the class with the closest centroid with respect to ℓ_2 distance. In high dimensions, the nearest shrunken centroids (NSC) method further enforces variable selection to facilitate interpretation and accurate prediction.

Suppose that we observe the dataset $\{Y_i, \mathbf{X}_i\}_{i=1}^n$, where $(Y_i, \mathbf{X}_i)$ are independent and identically distributed copies of $(Y, \mathbf{X})$, and n is the sample size. In high dimensions, we have that p is much larger than n. We further let $\mathcal{C}_k$ be the set of indices of the n_k samples in Class k.

In NSC, we first find initial estimates for the centroids of the k-th class and the variability of each X_j:

$$\bar{X}_{\cdot jk} = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} X_{ij}, \quad S_j^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} (X_{ij} - \bar{X}_{\cdot jk})^2. \tag{1}$$

We further compute the grand mean for X_j as $\bar{X}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$, and the sample proportion for Class k as $\hat{\pi}_k = \frac{n_k}{n}$, which is also the estimator for π_k. Then we shrink $\bar{X}_{\cdot jk}$ as follows. Let

$$d_{jk} = \frac{\bar{X}_{\cdot jk} - \bar{X}_{\cdot j}}{m_k S_j}, \tag{2}$$

where $m_k = \sqrt{1/n_k - 1/n}$ so that $m_k S_j$ equals the estimated standard error of the numerator in d_{jk}. For a user-speficied parameter $\Lambda \geq 0$, we soft-threshold d_{jk} to be $d'_{jk} = \text{sign}(d_{jk})(|d_{jk}| - \Lambda)_+$. Then the shrunken estimator for μ_{jk} is:

$$\bar{X}'_{\cdot jk} = \bar{X}_{\cdot j} + m_k S_j d'_{jk}, \tag{3}$$

and the estimated classifier is given by:

$$\hat{\delta}_{nsc}(\mathbf{X}_*) = \arg\min_k \sum_{j=1}^{p} \frac{(X_{*,j} - \bar{X}'_{\cdot jk})^2}{S_j^2} - 2\log\hat{\pi}_k. \tag{4}$$

It is easy to see that, for sufficiently large Λ, for many j we have $\bar{X}'_{\cdot j1} = \cdots = \bar{X}'_{\cdot jK}$. These variables do not have any effect on the final classification.

2.2. Our proposal of the robust nearest shrunken centroids method

The nearest shrunken centroids method is a powerful method that has been proved to have excellent performance [50, 24]. However, it is sensitive to heavy-tailed distributions and outliers for at least two reasons. First, it is apparent that $\bar{X}_{\cdot j}$, $\hat{X}'_{\cdot jk}$ and S_j attempt to estimate the (conditional) mean and variance of X_j. They are sample estimates with some shrinkage when applicable. However, sample estimates are known to be unstable in the presence of heavy tails. If data have several observations far away from the truth, the sample estimates tend to be inaccurate. Consequently, variable selection and prediction could be negatively impacted. Second, even if we knew the true (conditional) means and variance, NSC uses the ℓ_2 loss to measure the distances between $\mathbf{X}$ and the centroids. If $\mathbf{X}$ is drawn from a heavy-tailed distribution, it is likely that $\mathbf{X}$ could have a few elements that drive the prediction by coincidence and thus the classification is prone to errors.

To resolve these issues, we propose the robust NSC by replacing both the estimates and the ℓ_2 loss with robust counterparts. We achieve both goals through the Huber loss [25]. The Huber loss is defined as:

$$f_r(Z) = \begin{cases} Z^2, & \text{for } |Z| \leq r, \\ 2r(|Z| - \frac{1}{2}r), & \text{for } |Z| > r, \end{cases} \tag{5}$$

where $r > 0$ is a parameter that controls the level of robustness for the resulting estimate. It is well-known that, by clipping the ℓ_1 and the ℓ_2 loss at r, the Huber

loss effectively reduces the contribution of extreme values. The Huber loss has recently been regenerating many interests in high dimensions but has not been combined with classification methods to the best of our knowledge.

To apply the Huber loss, we first note some simple facts in NSC. It is easy to see that the estimates $\bar{X}_{\cdot j}$ and $\bar{X}_{\cdot jk}$ defined in (1) are minimizers to the ℓ_2 loss:

$$\bar{X}_{\cdot j} = \arg\min_{\alpha\in\mathbb{R}} \sum_{i=1}^{n} (X_{ij} - \alpha)^2, \quad \bar{X}_{\cdot jk} = \arg\min_{\alpha\in\mathbb{R}} \sum_{i\in\mathcal{C}_k} (X_{ij} - \alpha)^2, \tag{6}$$

while S_j^2 is also related to the ℓ_2 loss, as $S_j^2 = \sum_{k=1}^{K} \hat{\pi}_k (Q_{jk} - \bar{X}_{\cdot jk}^2)$, where

$$Q_{jk} = \arg\min_{\alpha\in\mathbb{R}} \sum_{i\in\mathcal{C}_k} (X_{ij}^2 - \alpha)^2. \tag{7}$$

To make these estimates more robust, we replace these estimates with suitable Huber estimates. We let

$$\tilde{X}_{\cdot jk} = \arg\min_{\alpha\in\mathbb{R}} \{\sum_{i\in\mathcal{C}_k} f_H(X_{ij} - \alpha)\}, \quad \tilde{Q}_{jk} = \arg\min_{\alpha\in\mathbb{R}} \{\sum_{i\in\mathcal{C}_k} f_H(X_{ij}^2 - \alpha)\}, \tag{8}$$

where $H > 0$ is a tuning parameter. We use $\tilde{X}_{\cdot jk}$ as an initial robust estimator for the conditional mean of X_j within Class k, and $\tilde{Q}_{jk}$ as a robust estimator for the second moment. Then we estimate the grand mean and the conditional variance of X_j as follows:

$$\tilde{X}_{\cdot j} = \sum_{k=1}^{K} \hat{\pi}_k \tilde{X}_{\cdot jk}, \quad \tilde{S}_j^2 = \sum_{k=1}^{K} \hat{\pi}_k (\tilde{Q}_{jk} - \tilde{X}_{\cdot jk}^2). \tag{9}$$

With these Huber estimates, we soft-threshold the centroid estimates in a similar way to (2)–(3). Namely, for a tuning parameter $\Lambda \geq 0$, we let

$$\tilde{d}_{jk} = \frac{\tilde{X}_{\cdot jk} - \tilde{X}_{\cdot j}}{m_k \tilde{S}_j}, \quad \tilde{d}'_{jk} = \mathrm{sign}(\tilde{d}_{jk})(|\tilde{d}_{jk}| - \Lambda)_+. \tag{10}$$

The shrunken Huber estimator for the centroid of X_j within Class k is

$$\hat{X}'_{\cdot jk} = \tilde{X}_{\cdot j} + m_k \tilde{S}_j \tilde{d}'_{jk}. \tag{11}$$

Moreover, we also use the Huber loss when we calculate the discriminant score. We assign a new observation $\mathbf{X}_*$ to the class

$$\hat{\delta}_h(\mathbf{X}_*) = \arg\min_k \sum_{j=1}^{p} f_h\Big(\frac{X_{*,j} - \hat{X}'_{\cdot jk}}{\tilde{S}_j}\Big) - 2\log\hat{\pi}_k, \tag{12}$$

where $h > 0$ is a tuning parameter. Compared with (4), we use the Huber loss instead of the ℓ_2 loss to measure the distance between $\mathbf{X}_*$ with each centroid

so that the effects of extreme values are limited. In practice, the parameters H and h are tuned separately to obtain higher accuracy. Numerically, there is a small probability that a variance estimator is a negative number close to zero. We exclude the jth predictor from the classifier if $\tilde{S}_j < 0$.

We refer to this modified NSC as the robust nearest shrunken centroids method (RNSC). As NSC, when Λ is reasonably large, most $\hat{X}'_{\cdot jk}$ equal $\tilde{X}_{\cdot j}$ by shrinkage, and the corresponding X_j is excluded from final classification in RNSC. The remaining variables give us a sparse classifier. RNSC has a similar interpretation as NSC, as it assigns an observation to its closest centroid and the selected variables can still be viewed as those differentially distributed across classes.

However, RNSC is more widely applicable than NSC, especially when the data are potentially heavy-tailed. In the next section, we will rigorously show in theory that RNSC is consistent under much weaker tail conditions than NSC. The robustness comes at a very low price, as we will later demonstrate in numerical studies that, when the heavy tail is not an issue, RNSC still works almost as well as NSC.

We want to remark that robust estimators other than the Huber estimator could be used to robustify the sample mean and variance. For example, the median-of-means estimator has similar robust properties to the Huber estimator [42, 31, 7, 27, 2]. We choose to use the Huber estimator in RNSC because, in addition to the estimation of the parameters, we also need a robust way to calculate the discriminant score. RNSC uses the Huber loss to replace the squared loss in the discriminant score for this purpose. The median-of-means estimator is not associated with a robust loss function that can be used in the discriminant score.

Finally, RNSC implicitly assumes that the marginal within-class variances are constant across classes so that the pooled estimator $\tilde{S}_j^2$ in (8) can be used. As suggested by the referee, we note that RNSC can be generalized to the heterogeneous case where the within-class variance is not constant. We explore this direction by developing the Heterogeneous Robust Nearest Shrunken Centroids method (HRNSC). Relevant results along this line are presented in Appendix A.

2.3. Choice of tuning parameters

RNSC has three tuning parameters, H, h, and Λ. We propose to tune them by two-step cross-validation. In the first step, we find the optimal H by tuning the pair of parameters (H, Λ) with the ℓ_2 loss classifier by cross-validation. In the second step, we fix H as chosen in the first step, and find the optimal (h, Λ) by cross-validation. Note that Λ is tuned twice because in both steps we need a reasonably large Λ to construct a sparse classifier that minimizes the cross-validation error. But only the second choice of Λ is used in the final classifier. This tuning procedure is used in RNSC in our reported numerical studies with the number of folds set to 10.

We admit that such a tuning procedure may have some loss of robustness, as H is chosen according to the ℓ_2 loss instead of a robust loss. If we could

tune (H, h, Λ) jointly and use the Huber loss throughout the cross-validation, the results could be better. However, we choose to tune H and h separately for computational concerns. Also, we observe that such a tuning procedure leads to fairly good results in numerical studies to be presented. In the future, it will be interesting to develop a robust tuning procedure. A related work is [9], in which the authors considered robust cross-validation in a low-dimensional non-parametric problem.

As suggested by the referee, an alternative approach for fitting RNSC is to apply the recent tuning-free principle by [55]. It is shown therein that we can solve an equation system to simultaneously find a suitable Huber loss parameter and the corresponding estimator. We describe such a method, referred to as the tuning-free RNSC (TF-RNSC), in Section 2.4. We want to remark though that TF-RNSC only avoids tuning on H, but still resorts to cross-validation on h and Λ.

2.4. The tuning-free RNSC

We describe the tuning-free RNSC (TF-RNSC) as a variant of RNSC that alleviates the computation cost of cross-validation when we construct the Huber estimates. TF-RNSC is based on the tuning-free principle in [55]. We first give a brief review of this principle.

Consider m independent and identically distributed (i.i.d.) random variables $Z_1, \ldots, Z_m$ with mean ξ. For a given r, we can find the Huber estimator for ξ as

$$\hat{\xi}_r = \arg\min_{\alpha} \sum_{i=1}^{m} f_r(Z_i - \alpha). \tag{13}$$

In order to automatically choose r, [55] propose to solve the following equation system:

$$\begin{cases} \sum_{i=1}^{m} \psi_r(Z_i - \alpha) = 0, \\ \frac{1}{m} \sum_{i=1}^{m} \min\{(Z_i - \alpha)^2, r^2\} - \frac{z}{m} = 0, \end{cases} \tag{14}$$

where $\psi_r(x) = \text{sign}(x) \min(|x|, r) = \frac{1}{2} f_r'(x)$ is $\frac{1}{2}$ of the derivative of the Huber loss and z refers to a user-specified parameter that controls the confidence level. The authors suggest $z = \log(m)$ in their paper, which we adopt. As a result, we obtain a choice of $\hat{r}$ and the corresponding $\hat{\xi}_{\hat{r}}$ without tuning.

In TF-RNSC, we do not construct estimates according to (8) that requires user-specified Huber loss parameters. Instead, when we estimate the Class k centroid for the j-th variable, we solve (14) by replacing $\{Z_i\}_{i=1}^m$ with $\{X_{i,jk}\}_{i \in \mathcal{C}_k}$ to obtain $\tilde{X}_{\cdot jk}^{TF}$. Similarly, we estimate the Class k second moment of the j-th variable by $\tilde{Q}_{jk}^{TF}$ that solves (8) with $\{Z_i\}_{i=1}^m$ replaced by $\{X_{i,jk}^2\}_{i \in \mathcal{C}_k}$. Then TF-RNSC proceeds much the same as RNSC, with the only exception that $\tilde{X}_{\cdot jk}^{TF}$ and $\tilde{Q}_{jk}^{TF}$ are used in the place of $\tilde{X}_{\cdot jk}$ and $\tilde{Q}_{jk}$.

With these tuning-free estimates, we further apply cross-validation to choose (h, Λ) in the final classifier. The tuning-free principle does not apply in this part,

as our ultimate goal is prediction instead of estimation. In this sense, we cannot completely avoid tuning in RNSC even with the assistance of the tuning-free principle. But we abuse the terminology and refer to the resulting classifier as tuning-free RNSC (TF-RNSC) to distinguish it from our proposal of RNSC.

3. Theoretical results

In this section, we study the properties and benefits of RNSC under the high-dimensional setting where p could be much larger than n. For variable selection, We show that RNSC consistently obtains the set of all important variables. For the classifier, we first show that with high probability the estimated classifier gives the same result as the true one. Both the results for variable selection and prediction require a mild condition that includes many heavy-tailed distributions. We then give a special case that the prediction error of RNSC converges to the Bayes error in the Gaussian scenario.

We only present the theoretical results for RNSC, but we conjecture that TF-RNSC has similar properties that can be proved with some simple modifications to our proofs for RNSC. The theoretical properties of RNSC are a result of the convergence of $\tilde{X}_{jk}$ and $\tilde{Q}_{jk}$ in Propositions 2 & 3 (Appendix B.1). For TF-RNSC, $\tilde{X}_{jk}^{TF}$ and $\tilde{Q}_{jk}^{TF}$ have similar properties according to Theorems 2.1 & 2.2 in [55].

Similar to NSC, RNSC can be used as a heuristic classifier without model assumptions. However, for the sake of theoretical studies, we introduce an intuitive model that allows us to define the set of important variables. We assume that the number of classes K is fixed and $\Pr(Y = k) = \pi_k \in (0, 1)$. We further assume the following model for $\mathbf{X}$ given Y,

$$\mathbf{X} = \boldsymbol{\mu}_k + \boldsymbol{\varepsilon}, \quad \text{if } Y = k, \tag{15}$$

where $\boldsymbol{\mu}_k = (\mu_{1k}, \ldots, \mu_{pk})^T \in \mathbb{R}^p$ is the conditional mean for Class k and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^T \in \mathbb{R}^p$ is the noise term. For any $j \in \{1, \ldots, p\}$, ε_j follows a distribution g_j with mean 0 and variance σ_j^2. This model is also studied in [35], but the main focus in that paper is variable transformation. In comparison, we work with the original data without transformation.

Throughout the rest of this section, let C be a generic positive constant that can vary in different places. The following assumptions are needed to obtain the results of variable selection and classifier convergence. Recall that g_j is the probability density function for ε_j.

(A1) There exists some constant $\zeta > 0$ such that $\max_k\{\|\boldsymbol{\mu_k}\|_\infty\} \le \zeta$.
(A2) There exist constants $u > 0$ and $U > 0$ such that $u \le \sigma_j \le U$ for all j.
(A3) For the prior π_k, there are constants $0 < c_1 < 1$ and $0 < c_2 < 1$ such that $c_1 < \pi_k < c_2$ for all k.
(A4) There exists some constant $V > 0$ such that $g_j(\varepsilon_j) \le V$ for all j.
(A5) There exists some constant $\kappa > 0$ such that $\mathrm{E}\varepsilon_j^4 \le \kappa^2$ for all j.

All these assumptions are very mild. Assumptions (A1) and (A2) guarantee that the mean and variance do not go to infinity as the dimension increases. These two technical conditions simplify our calculation, but, if needed, we can also allow ζ, u, U to diverge with n, p at the price of more tedious proofs. Assumption (A3) bounds π_k away from 0 and 1, which ensures that we have a decent sample size for each Class k. Assumptions (A4) and (A5) are regularity conditions on the distribution of ε_j. We only assume that g_j's are bounded and ε_j's have uniformly bounded fourth moments. These assumptions are much weaker than the popular sub-Gaussian assumption in the high-dimensional statistics literature, and allow the application of RNSC to heavy-tailed data.

We first present the theoretical result on variable selection. We start with defining the set of important variables. For any $k_1, k_2 \in \{1, \ldots, K\}$ such that $k_1 \neq k_2$, we have $f_h(\frac{x-\mu_{jk_1}}{\sigma_j}) = f_h(\frac{x-\mu_{jk_2}}{\sigma_j})$ for all $x \in \mathbb{R}$ if and only if $\mu_{jk_1} = \mu_{jk_2}$. Thus, the jth predictor is an important variable if and only if there exist $k_1 \neq k_2$ such that $\mu_{jk_1} \neq \mu_{jk_2}$. We define the set of important variables as

$$\mathcal{D} = \{j : \text{there exist } k_1, k_2 \in \{1, \ldots, K\} \quad \text{s.t.} \quad \mu_{jk_1} \neq \mu_{jk_2}\}. \tag{16}$$

We also denote the number of important predictors as $q = |\mathcal{D}|$. The estimator for $\mathcal{D}$ is

$$\hat{\mathcal{D}} = \{j : \text{there exist } k_1, k_2 \in \{1, \ldots, K\} \quad \text{s.t.} \quad \hat{X}'_{jk_1} \neq \hat{X}'_{jk_2}\}. \tag{17}$$

Further define

$$N_0 = \min_{k_1, k_2, j} \{|\mu_{jk_1} - \mu_{jk_2}| : |\mu_{jk_1} - \mu_{jk_2}| > 0\} \tag{18}$$

as the smallest nonzero mean difference. We have the following theorem.

Theorem 1. *Assume that (A1)-(A5) hold. Let N_0 be defined as in (18). Then we have the following conclusions:*

1. *Let $H = \frac{v^2}{\epsilon}$, where $v \geq \max\{U, \kappa\}$ and $\epsilon > 0$ s.t. $\epsilon \leq \min\{\frac{1}{2}, \frac{\sqrt{2}}{4}v, N_0/16\}$. Then with the choice of Λ such that $\frac{8\epsilon}{m_0 w} \leq \Lambda \leq \frac{N_0 - 8\epsilon}{C m_0 w}$, where $m_0 = \sqrt{C/n}$, $0 < w \leq u - C\epsilon$, we have*

$$\Pr(\hat{\mathcal{D}} = \mathcal{D}) \geq 1 - Cp \exp\{-\frac{Cn\epsilon^2}{v^2}\}. \tag{19}$$

2. *Furthermore, assume that $n \to \infty$ and $\frac{\log p}{n} \to 0$. Let $C\sqrt{\log p} \ll \Lambda \ll CN_0\sqrt{n} - C\sqrt{\log p}$. Then under the condition that $H \to \infty$ and $H \ll C\sqrt{\frac{n}{\log p}}$, we have*

$$\Pr(\hat{\mathcal{D}} = \mathcal{D}) \to 1. \tag{20}$$

Theorem 1 shows that, with a proper choice of tuning parameters H and Λ, RNSC exactly recovers the set of important variables with a probability tending to 1 even when $\log p = o(n)$, which is usually the best we can do in

high dimensions. Also, recall that this is achieved without the sub-Gaussian or sub-exponential assumption. NSC is unlikely to have the same property, as the sample mean is used and we cannot derive similar concentration inequalities without sub-Gaussian/sub-exponential assumptions.

Next, we show that RNSC consistently estimates the classifier as well. Apparently, the classifier defined in (12) intends to approximate

$$\delta_h(\mathbf{X}_*) = \arg\min_k \sum_{j=1}^{p} f_h\left(\frac{X_{*,j} - \mu_{jk}}{\sigma_j}\right) - 2\log \pi_k. \tag{21}$$

Let the training data set be $(Y_{tr}, \mathbf{X}_{tr})$. The following theorem shows that this approximation is very accurate.

Theorem 2. *Assume that (A1)-(A5) hold. Let N_0 be defined as in (18).*

1. *Let $H = \frac{v^2}{\epsilon}$, where $v \geq \max\{U, \kappa\}$ and $\epsilon > 0$ s.t. $\epsilon \leq \min\{\frac{1}{2}, \frac{\sqrt{2}}{4}v, Cu, \frac{N_0}{16}\}$. Then with the choice of $\Lambda = \frac{8\epsilon}{m_0 w}$, where $m_0 = \sqrt{C/n}$ and $0 < w \leq u - C\epsilon$, we have*
$$\Pr\left(\hat{\delta}_h(\mathbf{X}_*) \neq \delta_h(\mathbf{X}_*) \middle| (Y_{tr}, \mathbf{X}_{tr})\right) \leq qhC\epsilon^{4/5} \tag{22}$$
with probability greater than $1 - (q+p)C\exp\{-\frac{Cn\epsilon^2}{v^2}\}$.
2. *Assume that $n \to \infty$ and $\frac{q^{5/2}h^{5/2}\log p}{n} \to 0$ for any $h > 0$. With $H \to \infty$, $H \ll \sqrt{\frac{v^2 n}{\log p}}$, $C\frac{\Lambda}{\sqrt{n}} \to 0$ and $\Lambda \gg C\sqrt{\log p}$, we have*
$$\Pr\left(\hat{\delta}_h(\mathbf{X}_*) \neq \delta_h(\mathbf{X}_*) \middle| (Y_{tr}, \mathbf{X}_{tr})\right) \to 0. \tag{23}$$

Theorem 2 implies that the estimated classifier converges to the truth when only up to the fourth moment exists for $\mathbf{X}$. This moment condition includes many heavy-tailed distributions that are frequently encountered. Therefore, The results in Theorem 2 guarantee that the classification problem with heavy-tailed data or data with outliers can be handled by RNSC. We note though, for any fixed h, we need $\frac{q^{5/2}\log p}{n} \to 0$. This requirement is stronger than classifiers with the normality assumption [8], which can often handle $\frac{q\log p}{n} \to 0$. RNSC needs the model to be sparser (i.e, q to be smaller) because $\mathbf{X}_*$ can be heavy-tailed. Even though we can consistently identify $\mathcal{D}$ and estimate $\boldsymbol{\mu}_k$ and σ_j^2 accurately, $\mathbf{X}_*$ could inflate the estimation error in the discriminant score. However, in what follows we will show that, if the data is not heavy-tailed, there is no need for the stronger requirement on q for RNSC to be consistent in prediction.

Consider the special case that $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Delta})$ with $\boldsymbol{\Delta}$ being diagonal. Our model in (15) reduces to the diagonal linear discriminant analysis (LDA) model. Consequently, NSC can be viewed as an estimate for the so-called Bayes rule:

$$\delta_{bayes}(\mathbf{X}_*) = \arg\min_k \sum_{j=1}^{p} (X_{*,j} - \mu_{jk})^2/\sigma_j^2 - 2\log \pi_k, \tag{24}$$

which is the optimal classifier [24]. We show that RNSC also consistently estimates the Bayes rule under the normal assumption. In other words, we gain improved robustness with little loss of efficiency when data is not heavy-tailed. We denote ϕ as the probability density function for a standard normal random variable.

Theorem 3. *Assume that (A1)-(A3) hold. Further assume that* $\varepsilon_j \sim N(0, \sigma_j^2)$ *independently. Then we have following conclusions:*

1. *Let* $H = \frac{v^2}{\epsilon}$, *where* $v \geq \max\{U, \kappa\}$ *and* $\epsilon > 0$ *s.t.* $\epsilon \leq \min\{\frac{1}{2}, \frac{\sqrt{2}}{4}v, Cu, \frac{N_0}{16}\}$. *Then with the choice of* $\Lambda = \frac{8\epsilon}{m_0 w}$, *where* $m_0 = \sqrt{C/n}$ *and* $0 < w \leq u - C\epsilon$, *we have*

$$\Pr\left(\hat{\delta}_h(\mathbf{X}_*) \neq \delta_{bayes}(\mathbf{X}_*)\middle|(Y_{tr}, \mathbf{X}_{tr})\right) \leq C(q\epsilon^2)^{\frac{1}{4}} + q\left(\frac{\phi(Ch - C)}{Ch - C}\right) \quad (25)$$

   with probability greater than $1 - (q + p)C\exp\{-\frac{Cn\epsilon^2}{v^2}\}$.
2. *Furthermore, assume that* $n \to \infty$ *and* $\frac{q\log p}{n} \to 0$. *With* $H \to \infty$, $H \ll \sqrt{\frac{v^2 n}{\log p}}$, $C\frac{\Lambda}{\sqrt{n}} \to 0$ *and* $\Lambda \gg C\sqrt{\log p}$ *and* $h \to \infty$ *and* $h \gg \sqrt{\log q}$, *we have*

$$\Pr\left(\hat{\delta}_h(\mathbf{X}_*) \neq \delta_{bayes}(\mathbf{X}_*)\middle|(Y_{tr}, \mathbf{X}_{tr})\right) \to 0. \quad (26)$$

Theorem 3 shows that RNSC gives a similar prediction as the Bayes classifier under the normality and independence conditions. The dimensionality is allowed to diverge at the rate of $\frac{q\log p}{n} = o(1)$. This rate is identical to that of LDA methods, which explicitly assumes that the data are Gaussian. Hence, RNSC is expected to work as well as the less robust methods when there is no heavy tail.

As suggested by the referee, we further compare our theoretical results with sparse linear discriminant analysis (LDA) methods [8, 13, 56, 17, 38, 36, e.g]. To start, we note that our model in (15) includes the LDA model as a special case. The LDA model indicates that $\mathbf{X}$ is normal given Y [24]. If we assume that $\boldsymbol{\varepsilon}$ follows the multivariate normal distribution, (15) reduces to the LDA model. The normality assumption is critical for the theoretical study of sparse LDA methods, as most of them rely on the sub-Gaussian property to show the consistency in ultra-high dimensions. However, in our Theorems 1 & 2, we do not make the normality assumption and thus obtain theoretical results under weaker assumptions than the LDA model.

Another difference between LDA and RNSC is the way they handle correlations. RNSC (and its predecessor, NSC) is a distance-based classifier that calculates the distance in a coordinate-wise manner. It makes no attempt to model the correlation among $\boldsymbol{\varepsilon}$. The classifier and the active set $\mathcal{D}$ are fully determined by within-class means and variances. Theorems 1 & 2 do not make assumptions on the correlation structure, either, as RNSC converges to its population counterpart regardless of the correlation structure. On the other hand, LDA explicitly exploits the correlation among variables. For example, if $K = 2$, LDA aims to estimate $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$, where $\boldsymbol{\Sigma}$ is the covariance matrix for $\mathbf{X}$ within Class k, and we need to select the set $\mathcal{S} = \{j : \beta_j \neq 0\}$ [8, 17, 38].

These existing works also show that $\mathcal{D}$ can be quite different from $\mathcal{S}$. In this sense, when some special correlation exists and the normality assumption holds, RNSC could be sub-optimal to LDA even on the population level. But this sub-optimality is the price we pay for the extra flexibility in analyzing heavy-tailed data that does not satisfy the LDA model.

However, there is one special case when RNSC and sparse LDA are closely related, which is studied in Theorem 3. If $\varepsilon_j \sim N(0, \sigma_j^2)$ independently, the model in (15) becomes a diagonal LDA model (i.e, the LDA model with a diagonal covariance matrix). As such, RNSC and LDA intend to estimate the same Bayes rule. Theorem 3 shows that, under suitable conditions, RNSC indeed converges to the Bayes rule under the diagonal LDA model.

4. Simulation results

We present some simulation results to demonstrate the performance of RNSC. We consider several models to study binary-class and multi-class problems in heavy-tailed and outlier scenarios. For all models, we set the dimension $p = 2000$ and sample size $n = 100K$, where K is the number of classes. We assume the model in (15). We consider two sets of parameters in Models A and B, combined with four different distributions for ε_j in Models X.1–X.4, where X=A or B. The parameters are given as follows.

1. Model A (Binary): $K = 2$,
   $\boldsymbol{\mu}_1 = \mathbf{0}$,
   $\boldsymbol{\mu}_2 = 0.5 \times (-1.9, -1.8, -1.7, -1.6, -1.5, 2.5, 2.6, 2.7, 2.8, 2.9, 0_{p-10})$.
2. Model B (Multiclass): $K = 4$,
   $\boldsymbol{\mu}_1 = \mathbf{0}$,
   $\boldsymbol{\mu}_2 = (1.9, 1.3, -1.6, 1.4, -0.7, 0_{p-5})$,
   $\boldsymbol{\mu}_3 = (0_5, -2, 1.4, -1.9, -1.5, 0.8, 0_{p-10})$,
   $\boldsymbol{\mu}_4 = (0_{10}, -1.8, -2, -1.7, -2, -0.7, 0_{p-15})$.

For each selection of parameters, we consider four settings.

1. Model X.1 (Normal): $\varepsilon_j \sim N(0, 1)$ for all j.
2. Model X.2 (Heavy-tailed): $\varepsilon_j \sim t_3$ for all j.
3. Model X.3 (Outliers): $\varepsilon_j \sim 0.99N(0, 1) + 0.01N(10, 1)$.
4. Model X.4 (Heavy-tailed with outliers): $\varepsilon_j \overset{d}{=} Z_1 - 9Z_2$, where $Z_1 \sim t_4$ and $Z_2 \sim \text{Binomial}(0.01)$.

We consider our proposal of RNSC and TF-RNSC defined in Section 2.4 in the numerical studies. Other competitors include the original NSC method by [49], the median-based classifier (Med; [23]), sparse optimal scoring (SOS; [13]), support vector machine with SCAD penalty (SVM-s; [57]) and ℓ_1 penalized logistic regression (Logistic; [24]). NSC is implemented by `R` package `pamr`. SVM-s is implemented by the `R` package `penalizedSVM` and SVM-s in Model B is performed by one-vs-one method. The ℓ_1 penalized logistic regression is implemented by the `R` package `glmnet`. In multi-class problems, SOS is implemented

by the `R` package `sparseLDA`, while in binary case, SOS is implemented by the `R` package `TULIP` [45] by exploiting an equivalence between SOS and direct sparse discriminant analysis [38, 37]. We consider 200 replicates for each model and report the results in Table 1.

TABLE 1

Simulation results. The means of prediction error and variable selection of 200 replications are reported. The prediction errors as PE *are reported in percentage, while the number of correctly selected variables of each method is denoted as* C *(The truth is 10 for Model A and 15 for Model B) and the number of incorrectly selected ones is* IC*. The standard error of each result is given in the parentheses. The standard error smaller than* 0.1 *are rounded to* 0*.*

Model		TF-RNSC	RNSC	NSC	Med	SOS	Logistic	SVM-s
	PE(%)	4.4(0.11)	5(0.14)	4.5(0.12)	4.5(0.11)	5.6(0.13)	6.3(0.13)	6.4(0.17)
A.1	C	10(0)	10(0)	10(0)	10(0)	9(0.1)	10(0)	9(0.1)
	IC	50(4.8)	140(11.7)	16(2.4)	21(3.3)	22(1.9)	50(0.7)	26(2.3)
	PE(%)	7(0.08)	7.5(0.12)	7.2(0.1)	7.2(0.09)	13.9(0.11)	9.5(0.11)	9(0.11)
B.1	C	15(0.1)	15(0)	14(0.1)	14(0.1)	14(0)	14(0)	14(0.1)
	IC	72(8.4)	181(20.6)	31(6)	37(6.2)	15(0.2)	150(1.4)	58(3.5)
	PE(%)	10.9(0.18)	11(0.19)	15.1(0.21)	15.4(0.21)	16.2(0.23)	17.2(0.24)	16.2(0.24)
A.2	C	9(0.1)	10(0)	8(0.1)	8(0.1)	8(0.1)	8(0.1)	9(0.1)
	IC	23(2.1)	108(6.5)	13(1.5)	12(2)	20(1.5)	40(1.5)	51(3.4)
	PE(%)	16.4(0.12)	16.5(0.16)	21.3(0.15)	21.7(0.15)	26.8(0.16)	24.6(0.17)	23(0.17)
B.2	C	14(0.1)	15(0)	13(0.1)	13(0.1)	13(0.1)	14(0.1)	14(0.1)
	IC	17(2.5)	102(9.8)	6(1.1)	9(1.4)	24(0.1)	109(3.2)	144(6.1)
	PE(%)	6.1(0.15)	6.6(0.16)	8.9(0.14)	8.9(0.15)	9.8(0.18)	10.8(0.18)	9.7(0.2)
A.3	C	10(0.1)	10(0)	9(0.1)	9(0.1)	8(0.1)	9(0.1)	9(0.1)
	IC	34(3.1)	129(9)	10(1.3)	8(1.2)	18(1.7)	36(1.2)	31(3.1)
	PE(%)	8.9(0.1)	9(0.12)	11.6(0.12)	12.1(0.13)	18.7(0.13)	15.5(0.15)	12.5(0.12)
B.3	C	14(0.1)	15(0)	14(0.1)	14(0.1)	14(0.1)	14(0.1)	14(0.1)
	IC	30(3.6)	121(11.6)	6(1.7)	5(0.9)	21(0.2)	124(3)	75(4.4)
	PE(%)	10.8(0.19)	9.9(0.18)	14.9(0.19)	14.6(0.19)	15.6(0.22)	17(0.25)	15.8(0.2)
A.4	C	9(0.1)	10(0)	8(0.1)	8(0.1)	8(0.1)	9(0.1)	9(0.1)
	IC	21(1.8)	111(7.6)	10(1.1)	9(1.3)	14(1.4)	41(1.5)	42(3.1)
	PE(%)	15.6(0.13)	15.5(0.15)	21.4(0.14)	21(0.14)	26.3(0.15)	24(0.19)	22.6(0.16)
B.4	C	14(0.1)	15(0)	13(0.1)	13(0.1)	14(0.1)	14(0.1)	14(0.1)
	IC	18(2.6)	106(11.4)	5(0.8)	5(0.9)	23(0.2)	103(3.4)	118(5.4)

For the prediction error in Table 1, we can see that in Models A.1 & B.1, our proposal has slightly worse performance than NSC. This is because the noise term is normal, and NSC is a more efficient estimate of the Bayes rule. RNSC has a little loss of efficiency because it uses robust estimates. But RNSC has noticeable advantages in all the other settings, where the data have heavy tails or extreme outliers. Moreover, RNSC successfully selects all important variables in most replicates in all models. Other competitors are likely to miss at least 1 or 2 important variables in most replicates. Such results support the application of RNSC in practice, where normality assumptions could be too stringent. They also support our theoretical studies that RNSC is consistent even when sub-Gaussian assumptions are not met.

On the other hand, TF-RNSC consistently gives a competitive performance in all the simulation settings. Even in Models A.1 & B.1 where the normality assumption holds, TF-RNSC is comparable to NSC. In all the models except for Model A.4, TF-RNSC is either comparable to or better than RNSC. Such a fact demonstrates the potential of TF-RNSC.

5. Application to a real dataset

In this section, we further demonstrate the performance of RNSC on the lung cancer data ([22]). This dataset contains 181 patient samples with 12533 gene expression levels as predictors. The response is of two classes, either mesothelioma (MPM) or adenocarcinoma (ADCA). These two classes refer to two types of lung-cancer-related issues, where MPM is relatively rare and ADCA is much more common([22]).

In the analysis, the dataset is randomly split into two parts, with the training set containing 30% of the samples (54 patients) and the testing set containing the rest. We fit classifiers by RNSC and all the competitors included in Section 4. For RNSC, we choose the tuning parameter by 10-fold cross-validation in the same way as in Section 4. The classification errors are evaluated on the testing set. We repeat this procedure for 100 times. The results are reported in Table 2. It can be seen that RNSC has better accuracy on prediction than all the other methods. TF-RNSC is the second best, but the one-sided paired t-test suggests that the difference between RNSC and TF-RNSC is borderline insignificant (p-value=0.054). However, RNSC is significantly better than all the other methods at the 0.05 level, while TF-RNSC is not significantly better than NSC (p-value=0.2513) or Med (p-value=0.1569). In what follows, we focus on the results of RNSC.

TABLE 2

The means of prediction errors (PE(%)), number of selected genes (SG), and numbers of frequently selected genes (# of FSG). The standard errors are given in the parentheses.

Method	TF-RNSC	RNSC	NSC	Med	SOS	Logistic	SVM-s
PE(%)	2.3(0.31)	1.7(0.18)	2.7(0.32)	2.6(0.24)	3.7(0.24)	3.8 (0.24)	17.4(0.17)
SG	707(107)	477(150)	214(70)	209(80)	20(0.66)	10(0.35)	8(0.94)
# of FSG	70	23	12	2	4	0	8

The favorable performance of RNSC can be partially explained by the heavy tails or the presence of outliers in the data. For simplicity, we say that a gene is "frequently selected" (FSG for short) if it is selected at least in 70 out of the 100 replicates. Eleven genes are uniquely frequently selected by RNSC, but not by other methods. For the sake of space, we plot the empirical probability density functions of 4 of them in Figure 1. It can be seen that there are outliers far away from the bell shape in these 4 gene expressions when the patients are detected as ADCA. Gene expressions of 39409_at and 39795_at tend to be heavy-tailed when the patients are detected as MPM. Hence, it is more difficult for methods such as NSC to detect them.

We further check the empirical excess kurtosis (EEK) of selected genes by RNSC and NSC. The EEK is defined as empirical kurtosis minus 3, where 3 is the kurtosis of normal distribution. Large EEK is indicative of heavy tails or the presence of outliers. The EEKs of genes for each class are presented in Figure 2. These genes are either selected by NSC or uniquely selected by RNSC. Within the class of ADCA, the EEKs of the expression of selected genes are generally large. Thus, the robust method of RNSC is more appropriate on this dataset

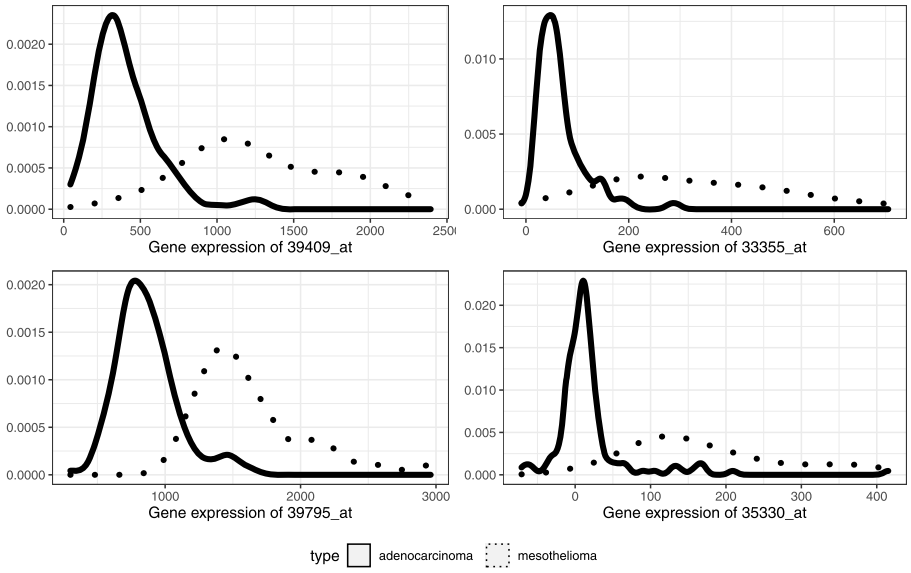

FIGURE 1. *Empirical probability density curves within each class for 4 out of 11 genes that are uniquely selected by RNSC are presented here.*

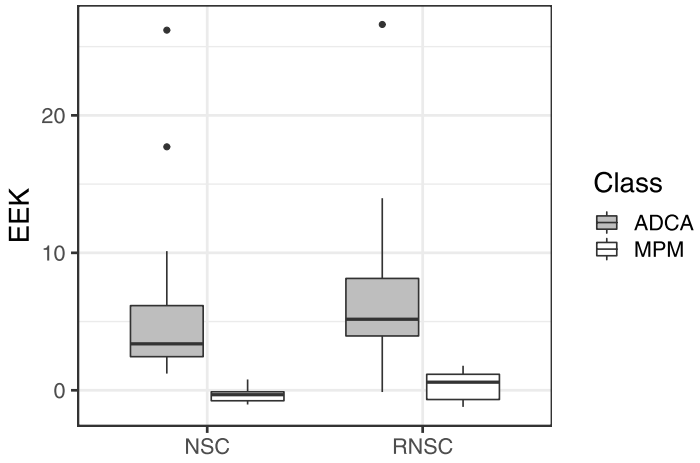

FIGURE 2. *The distribution of empirical excess kurtosis(EEK) of genes selected by NSC or uniquely selected by RNSC is presented for the given class. When the patient is detected as ADCA, the expressions of selected genes have large kurtosis.*

and produces more accurate classification results.

Finally, there is evidence that the genes uniquely selected by RNSC are associated with MPM. We list several genes out of them and the publications that point out their association with the disease. WT1 products are found to be a

possible marker for MPM by [1]. [30] observes an elevated expression level of PEA-15 in MPM cells. [29] finds the over-expression of SEMA3C in MPM cells. It is observed by [14] that there are significant gene expression differences of GFPT2 between ADCA and MPM. PTGIS is deregulated with statistical significance in at least one cell line in [41]. These studies support that the variables identified by RNSC are biologically meaningful as well.

6. Discussion

We propose RNSC as a robust high-dimensional classifier. Numerical and theoretical results suggest that RNSC is a competitive classifier on a wide range of classification problems. The improved robustness in RNSC is a result of employing the Huber loss. We use the Huber loss because it is naturally related to the ℓ_2 loss in the original NSC method, and strikes a delicate balance between efficiency and robustness. We note though that there are other approaches for high-dimensional robust statistics. For example, estimation for conditional median regression was proposed based on different regression models [3, 16, 54, e.g]. [10] developed a truncation-type estimator. It is out of the scope of this paper to conduct an exhaustive study of robust extensions of NSC with these works, but such an investigation is expected to be an interesting future research topic.

Appendix A: The heterogeneous robust nearest shrunken centroids method

In NSC and RNSC, the pooled sample variances, S_j^2 and $\tilde{S}_j^2$, are used in standardization, which leads to the assumption that the true variances σ_j^2 for each j stays constant across classes in Section 3. However, if the true variances are not constant, the shrinkage procedure in (10) and the estimation for (16) may need to be modified. Hence, we consider a generalization of RNSC to heterogeneous problems, referred to as the heterogeneous robust nearest shrunken centroids classifier (HRNSC).

A.1. Methodology

Similar to the idea of RNSC, we aim to develop the HRNSC classifier of the form

$$\hat{\delta}_{\text{HRNSC}}(\mathbf{X}_*) = \arg\min_k \sum_{j=1}^{p} f_h\left(\frac{X_{*,j} - \mathring{X}'_{\cdot jk}}{\mathring{S}'_{jk}}\right) + 2\log \mathring{S}'_{jk} - 2\log \hat{\pi}_k, \tag{27}$$

where $\mathbf{X}_*$ is a new observation, and $\mathring{X}'_{\cdot jk}, \mathring{S}'_{jk}$ are robust estimates of the within-class mean and variance that will be formally defined later. Note that HRNSC differs from RNSC in that we have a class-specific $\mathring{S}'_{jk}$ and the additional term

of $2\log \mathring{S}'_{jk}$ in the classifier. The class-specific $\mathring{S}'_{jk}$ models the potential heterogeneity in the data. The term $2\log \mathring{S}'_{jk}$ makes the HRNSC classifier identical to the quadratic discriminant analysis (QDA) when we make the normality assumption and let $h = \infty$. For some recent developments of sparse QDA, see [44, 32, 18, 26, e.g].

To construct $\mathring{X}'_{\cdot jk}, \mathring{S}'_{jk}$, we start with finding some robust initial estimates. Let $\tilde{X}_{\cdot j}$, $\tilde{X}_{\cdot jk}$, $\tilde{Q}_{jk}$ and $\tilde{S}_j^2$ be defined as in (8) and (9), and denote

$$\tilde{S}_{jk}^2 = \tilde{Q}_{jk} - \tilde{X}_{\cdot jk}^2 \tag{28}$$

as the initial estimator for the within class variance. These estimates involve minimizers of the Huber loss and are thus robust. But they do not induce sparsity in the final estimator. To enforce sparsity, we further consider shrunken versions of them.

We note that, a predictor X_j is not important in (27) if and only if $\mathring{X}'_{\cdot jk}, \mathring{S}'_{jk}$ are both constant across k. Consequently, we need to shrink both $\tilde{X}_{\cdot jk}$ and $\tilde{S}_{jk}^2 = \tilde{Q}_{jk} - \tilde{X}_{\cdot jk}^2$ towards the pooled mean and variance, respectively. In other words, we need to shrink $\tilde{X}_{\cdot jk} - \tilde{X}_{\cdot j}$ and $\tilde{S}_{jk}^2 - \tilde{S}_j^2$ to zero. Recall that, in RNSC, the shrinkage is scaled by a robust estimate of the standard error of each estimate. We use the same technique in HRNSC. Exact formulas for the standard errors of $\tilde{X}_{\cdot jk}$ and $\tilde{S}_{jk}^2$ are difficult to derive, but we can use some surrogates as follows.

Recall that $\bar{X}_{\cdot jk}$ is the sample mean within Class k and $\bar{X}_{\cdot j}$ is the pooled sample mean for the jth predictor. We also define $S_{jk}^2 = \frac{1}{n_k - 1}\sum_{i \in \mathcal{C}_k}(X_{ij} - \bar{X}_{\cdot jk})^2$ as the sample variance within Class k and $\bar{S}_j^2 = \sum_{k=1}^K \hat{\pi}_k S_{jk}^2$ as their average. These sample estimates are different from our Huber estimates, but their standard errors are easy to obtain, and we use robust estimates of their standard errors to scale our shrinkage of the Huber estimates. This approach is only an approximation, but yield reasonably good performance in our numerical studies. To this end, we have the following lemma concerning the variance of $\bar{X}_{\cdot jk} - \bar{X}_{\cdot j}$ and $S_{jk}^2 - \bar{S}_j^2$. The proof of this lemma is in Section A.3.

Lemma 1. *For X_{ij}, assume that X_{ij} has mean μ_{jk} and variance σ_{jk}^2 given $Y_i = k$. Further denote its b-th moment within Class k as $M_{bjk} = \mathbb{E}[X_{ij}^b \mid Y_i = k]$ for $b \in \{2, 3, 4\}$. Then for the random variables $\bar{X}_{\cdot jk} - \bar{X}_{\cdot j}$ and $S_{jk}^2 - \bar{S}_j^2$ defined as above, we have*

$$\mathrm{Var}[\bar{X}_{\cdot jk} - \bar{X}_{\cdot j} \mid n_1, \ldots, n_K] = (\frac{1}{n_k} - \frac{2}{n})\sigma_{jk}^2 + \frac{1}{n}\sum_{k=1}^K \hat{\pi}_k \sigma_{jk}^2, \tag{29}$$

and

$$\begin{aligned}&\mathrm{Var}[S_{jk}^2-\bar{S}_j^2\mid n_1,\ldots,n_K]\\=&\frac{(n-n_k)^2}{n^2}\left(\frac{M_{4jk}-4\mu_{jk}M_{3jk}+6\mu_{jk}^2M_{2jk}-3\mu_{jk}^4}{n_k}-\frac{\sigma_{jk}^4(n_k-3)}{n_k(n_k-1)}\right)\\+&\sum_{k'\neq k}\frac{n_{k'}^2}{n^2}\left(\frac{M_{4jk'}-4\mu_{jk'}M_{3jk'}+6\mu_{jk'}^2M_{2jk'}-3\mu_{jk'}^4}{n_{k'}}-\frac{\sigma_{jk'}^4(n_{k'}-3)}{n_{k'}(n_{k'}-1)}\right).\end{aligned}\tag{30}$$

To estimate the right-hand side of (29) & (30) for heavy-tailed data, we again resort to Huber estimates. According to (8), (28), and (9), we have $\tilde{X}_{jk}$ as the estimate for μ_{jk}, $\tilde{Q}_{jk}$ for M_{2jk}, $\tilde{S}_{jk}^2$ for σ_{jk}^2, and $\tilde{S}_j^2$ for $\sum_{k=1}^K\hat{\pi}_k\sigma_{jk}^2$, respectively. We further define the Huber estimators for third and fourth moments as

$$\tilde{M}_{bjk}=\arg\min_{\alpha\in\mathbb{R}}\{\sum_{i\in\mathcal{C}_k}f_H(X_{ij}^3-\alpha)\}\tag{31}$$

for M_{bjk} with $b\in\{3,4\}$.

Then we have

$$\begin{aligned}\tilde{T}_{jk}^4=&\frac{(n-n_k)^2}{n^2}\left[\frac{\tilde{M}_{4jkH}-4\tilde{X}_{jk}\tilde{M}_{3jkH}+6\tilde{X}_{jk}^2\tilde{Q}_{jk}-3\tilde{X}_{jk}^4}{n_k}-\frac{\tilde{S}_{jk}^4(n_k-3)}{n_k(n_k-1)}\right]\\+&\sum_{k'\neq k}\frac{n_{k'}^2}{n^2}\left[\frac{\tilde{M}_{4jk'H}-4\tilde{X}_{jk'}\tilde{M}_{3jk'H}+6\tilde{X}_{jk'}^2\tilde{Q}_{jk'}-3\tilde{X}_{jk'}^4}{n_{k'}}-\frac{\tilde{S}_{jk'}^4(n_{k'}-3)}{n_{k'}(n_{k'}-1)}\right]\end{aligned}\tag{32}$$

as the robust estimator for $\mathrm{Var}[S_{jk}^2-\bar{S}_j^2\mid n_1,\ldots,n_K]$.

Based on this conclusion, we consider the following shrunken estimator. Let

$$\tilde{v}_{jk}^2=\frac{\tilde{S}_{jk}^2-\tilde{S}_j^2}{\tilde{T}_{jk}^2},\quad \tilde{v}_{jk}'^2=\mathrm{sign}(\tilde{v}_{jk}^2)(|\tilde{v}_{jk}^2|-\Lambda_2)_+,\tag{33}$$

where $\Lambda_2\geq 0$ is a tuning parameter. With $\tilde{v}_{jk}^2$ shrunken by Λ_2, the shrunken Huber estimator for the variance σ_{jk}^2 within Class k is

$$\mathring{S}_{jk}'^2=\tilde{S}_j^2+\tilde{T}_{jk}^2\tilde{v}_{jk}'^2.\tag{34}$$

To scale $\tilde{X}_{\cdot jk}-\tilde{X}_{\cdot j}$, we again propose a robust estimator for the variance of $\bar{X}_{\cdot jk}-\bar{X}_{\cdot j}$ as a surrogate. As the robust shrunken estimator for the within class variance is obtained in (34), we use $\mathring{S}_{jk}'^2$ as the robust estimator for σ_{jk}^2 now. Then $(\frac{1}{n_k}-\frac{2}{n})\mathring{S}_{jk}'^2+\frac{1}{n}\tilde{S}_j^2$ is a surrogate for the variance of $\tilde{X}_{\cdot jk}-\tilde{X}_{\cdot j}$ according to Lemma 1.

Then following the same shrinking procedure of $\tilde{X}_{\cdot jk}$ in Section 2.2, we have

$$\mathring{d}_{jk} = \frac{\tilde{X}_{\cdot jk} - \tilde{X}_{\cdot j}}{\sqrt{(\frac{1}{n_k} - \frac{2}{n})\mathring{S}'^2_{jk} + \frac{1}{n}\tilde{S}^2_j}}, \quad \mathring{d}'_{jk} = \text{sign}(\mathring{d}_{jk})(|\mathring{d}_{jk}| - \Lambda_1)_+. \tag{35}$$

The shrunken Huber estimator for the centroid of X_j within Class k is

$$\mathring{X}'_{\cdot jk} = \tilde{X}_{\cdot j} + \sqrt{(\frac{1}{n_k} - \frac{2}{n})\mathring{S}'^2_{jk} + \frac{1}{n}\tilde{S}^2_j}\mathring{d}'_{jk}. \tag{36}$$

In HRNSC, we plug (34) & (36) into (27). Compared to RNSC, HRNSC adapts to non-constant within-class variance.

A.2. Numerical analysis

We have three sets of numerical results for HRNSC. We first apply HRNSC on Models A.X and B.X in Section 4 to compare it with the methods in Table 1. Then we modify model A.X by changing the variance of five predictors following the important variables to be C.X. To be specific, we let the variance of ϵ_{jk} in model $A.X$ to be k times of its original variance for any $10 < j \leq 15$. We last check its performance on the real dataset previously discussed in Section 5.

HRNSC has many additional tuning parameters compared to existing methods. For the sake of time, we employ tuning-free Huber estimators in HRNSC and use grid-search to determine the parameters Λ_1, Λ_2 and h with 10-fold cross-validation in HRNSC. The results of 200 replicates for Models A.X and B.X, and Models C.X are given in Tables 3 and 4, respectively.

TABLE 3

The means of prediction errors, number of correctly and incorrectly selected variables for HRNSC for Models A.X and B.X. The standard errors are in the parentheses. The true value of correctly selected variables is 10 for Model A.X and 15 for Model B.X.

HRNSC	A.1	B.1	A.2	B.2
PE(%)	4.6(0.11)	7.4(0.09)	14.5(0.2)	21.2(0.14)
C	10(0)	15(0.1)	9(0.1)	14(0.1)
IC	94(6)	145(11)	28(1.9)	30(2.7)
	A.3	B.3	A.4	B.4
PE(%)	7.4(0.15)	10.3(0.11)	14.5(0.21)	20.6(0.15)
C	10(0.1)	14(0.1)	9(0.1)	14(0.1)
IC	40(2.8)	41(3.8)	24(1.8)	27(2.1)

Moreover, we apply HRNSC on the lung cancer dataset introduced in Section 5 for 100 replicates. On average, HRNSC has the prediction error of 2.0% (0.72), 149(45.0) selected genes, and 17 frequently selected genes (in the parentheses are corresponding standard errors.).

It can be seen that both in the simulation and real data study that HRNSC outperforms NSC, but does not have a clear advantage over RNSC. One potential reason is that, given the high dimensions and the low sample sizes, estimating too many parameters may bring in excessive variability. Another possibility

TABLE 4

The means of prediction errors, number of correctly and incorrectly selected variables are for Models C.X. The standard errors are in the parentheses. The true value of correctly selected variables is 15 for Model C.X.

Model	C.1			C.2		
Method	HRNSC	RNSC	NSC	HRNSC	RNSC	NSC
PE(%)	4.4(0.11)	5.1(0.13)	4.5(0.11)	15(0.21)	10.8(0.18)	14.9(0.2)
C	11(0.1)	11(0.1)	10(0.1)	9(0.1)	10(0.1)	8(0.1)
IC	150(13.3)	153(12.3)	22(3.5)	31(2.1)	108(6.3)	11(1.6)
Model	C.3			C.4		
Method	HRNSC	RNSC	NSC	HRNSC	RNSC	NSC
PE(%)	7.5(0.16)	6.7(0.17)	8.9(0.14)	14.6(0.22)	10.2(0.17)	14.7(0.19)
C	10(0.1)	11(0.1)	9(0.1)	9(0.1)	10(0.1)	8(0.1)
IC	45(3)	132(8.5)	15(2.1)	27(1.8)	101(6.6)	9(1.9)

is that HRNSC may be intrinsically less robust than RNSC because it needs to estimate the third and fourth moments. Therefore, we believe that the full development of HRNSC is worth further investigation, which is out of the scope of the current paper.

A.3. Proof of Lemma 1

We prove Lemma 1 in this section. The following proposition is necessary in the proof.

Proposition 1. *(Corollary 1 of [11] with $N = \infty$) Suppose $\{Z_1, \ldots, Z_m\}$ are m i.i.d copies of a random variables Z with mean ξ and variance ϖ^2. Denote its sample variance as $s^2 = \frac{\sum_{i=1}^m (Z_i - \bar{Z})^2}{m-1}$, where $\bar{Z}$ is the sample mean. Then we have following property:*

$$\mathrm{Var}[s^2] = \frac{\mathbb{E}[(Z-\xi)^4]}{m} - \frac{\varpi^4(m-3)}{m(m-1)}. \tag{37}$$

Proof of Lemma 1. For $\bar{X}_{\cdot jk} - \bar{X}_{\cdot j}$, as X_{ij} are independent for all i, we have that

$$\begin{aligned}
&\mathrm{Var}[\bar{X}_{\cdot jk} - \bar{X}_{\cdot j} \mid n_1, \ldots, n_K] \\
=&\mathrm{Var}[\frac{1}{n_k}\sum_{i\in\mathcal{C}_k} X_{ij} - \frac{1}{n}\sum_{i\in\mathcal{C}_k} X_{ij} \mid n_1, \ldots, n_K] + \mathrm{Var}[\frac{1}{n}\sum_{i\notin\mathcal{C}_k} X_{ij} \mid n_1, \ldots, n_K] \\
=&(\frac{1}{n_k} - \frac{1}{n})^2 n_k \sigma_{jk}^2 + \frac{1}{n^2}\sum_{k'\neq k} n_{k'}\sigma_{jk'}^2 \\
=&(\frac{1}{n_k} - \frac{2}{n})\sigma_{jk}^2 + \frac{1}{n}\sum_{k=1}^K \hat{\pi}_k \sigma_{jk}^2.
\end{aligned}$$

For $S_{jk}^2 - \bar{S}_j^2$, we know that

$$S_{jk}^2 - \bar{S}_j^2 = -(\sum_{k'\neq k} \frac{n_{k'}}{n} S_{jk'}^2 + \frac{n-n_k}{n} S_{jk}^2).$$

Since all $S_{jk'}^2$ and S_{jk}^2 are independent, we have

$$\operatorname{Var}[S_{jk}^2 - \bar{S}_j^2 \mid n_1, \ldots, n_K] = \sum_{k' \neq k} \frac{n_{k'}^2}{n^2} \operatorname{Var}[S_{jk'}^2 \mid n_{k'}] + \frac{(n-n_k)^2}{n^2} \operatorname{Var}[S_{jk}^2 \mid n_k].$$

By Proposition 1, we have

$$\operatorname{Var}[S_{jk}^2 \mid n_k] = \frac{\mathbb{E}[(X_{ij} - \mu_{jk})^4 \mid Y_i = k]}{n_k} - \frac{\sigma_{jk}^4 (n_k - 3)}{n_k - 1}$$

for all k. Therefore, we have

$$\begin{aligned}
&\operatorname{Var}[S_{jk}^2 - \bar{S}_j^2 \mid n_1, \ldots, n_K] \\
=&\frac{(n-n_k)^2}{n^2} \left(\frac{\mathbb{E}[(X_{ij} - \mu_{jk})^4 \mid Y_i = k]}{n_k} - \frac{\sigma_{jk}^4 (n_k - 3)}{n_k (n_k - 1)} \right) \\
+&\sum_{k' \neq k} \frac{n_{k'}^2}{n^2} \left(\frac{\mathbb{E}[(X_{ij} - \mu_{jk'})^4 \mid Y_i = k']}{n_{k'}} - \frac{\sigma_{jk'}^4 (n_{k'} - 3)}{n_{k'} (n_{k'} - 1)} \right).
\end{aligned}$$

As $\mathbb{E}[(X_{ij} - \mu_{jk})^4 \mid Y_i = k] = M_{4jk} - 4\mu_{jk} M_{3jk} + 6\mu_{jk}^2 M_{2jk} - 3\mu_{jk}^4$ by straightforward calculation, we obtain (30) in Lemma 1. □

Appendix B: Proof of theoretical results

To prove Theorems 1, 2 and 3, we first show that the Huber estimator $\tilde{X}_{\cdot jk}$ converges to the true mean μ_{jk} and $\tilde{S}_j$ converges to the true standard error σ_j in Section B.1. Then, the properties of variable selection and convergence of classifier will be justified.

In this part, to give details in the proof and simplify the notation, we define

$$N_j = \min_{k_1, k_2} \{ |\mu_{jk_1} - \mu_{jk_2}| : |\mu_{jk_1} - \mu_{jk_2}| > 0 \}, \tag{38}$$

$$\tau(\epsilon) = 4 \sum_{k=1}^{K} [\exp\{-\frac{n\pi_k \epsilon^2}{2v^2}\} + \exp\{-\frac{n\pi_k^2}{4}\} + \exp\{-\frac{n\pi_k \epsilon^2}{3}\}], \tag{39}$$

and

$$\vartheta(\epsilon) = \frac{C_1 \epsilon^{4/5} + (2\zeta C_1 + 4w + 16wC_2)\epsilon + (4 + 16C_2) C_1 \epsilon^2}{uw}, \tag{40}$$

where $C_1 = K[(8\zeta + 20) + U^2]/u$, $C_2 = \sqrt{1 + 2/\min_k\{\pi_k\}}$ and $w = u - C_1 \epsilon$.

The terms that contain constants such as C_1, C_2, m_0, u, U, V and ζ will be eventually merged into the constant C.

B.1. Necessary Lemmas

Theorem 5 in [19] gives a general convergence result of the Huber estimator.

Proposition 2 (Theorem 5 in [19]). *Let $\{Z_1, \ldots, Z_m\}$ be m i.i.d random variables with mean ξ and variance ϖ^2. If $\tilde{Z} = \{\alpha \mid \min_\alpha\{\sum_{i=1}^m f_r(Z_i - \alpha)\}\}$ is the Huber estimator for ξ, then for any ϵ such that $0 < \epsilon \leq \frac{\sqrt{2}}{4}v$, where $v \geq \varpi$, by letting $r = \frac{v^2}{\epsilon}$, we have*

$$\Pr(|\tilde{Z} - \xi| \geq 4\epsilon) \leq 2\exp\{-\frac{m\epsilon^2}{v^2}\}. \tag{41}$$

To simplify the expression in our discussion, we replace δ in the original paper with $\epsilon = v\sqrt{\frac{\log(1/\delta)}{m}}$ in Proposition 2. Same expression will be applied in following propositions.

The result in Proposition 2 can be generalized to the Huber estimators for variance. As mentioned after Theorem 5 in [19], we have the following proposition:

Proposition 3 (The discussion following Theorem 5 in [19]). *Let $\{Z_1, \ldots, Z_m\}$ be m i.i.d random variables with mean ξ and variance ϖ^2. If the Huber estimator for $Q = \mathrm{E}Z_i^2$ is $\tilde{Q} = \{\alpha \mid \min_\alpha\{\sum_{i=1}^m f_r(Z_i^2 - \alpha)\}\}$, then for any ϵ such that $0 < \epsilon \leq \frac{\sqrt{2}}{4}v$, where $v \geq \sqrt{\mathrm{Var}(Z_i^2)}$, by letting $r = \frac{v^2}{\epsilon}$, we have*

$$\Pr(|\tilde{Q} - Q| \geq 4\epsilon) \leq 2\exp\{-\frac{m\epsilon^2}{v^2}\}. \tag{42}$$

For any fixed $j \in \{1, \ldots, p\}$, any class k and $i \in \mathcal{C}_k$, $\{X_{ij}\}$ are independent and identically distributed. We can then generalize the conclusions in Propositions 2 & 3 to obtain the concentration inequalities for $\tilde{X}_{\cdot jk}$ and $\tilde{S}_j$. Considering the fact that $\tilde{X}_{\cdot jk}$ and $\tilde{S}_j$ are related to n_k and $\hat{\pi}_k$, it is necessary to introduce the following proposition to bound them.

Proposition 4 (Hoeffding's inequality). *Let n_k be the sum of n independent and identically distributed Bernoulli random variables with probability π_k and $\hat{\pi}_k = \frac{n_k}{n}$ be the estimator for π_k. Then we have*

$$\Pr(n_k \leq n\pi_k/2) \leq \exp\{-n\pi_k^2/4\}, \tag{43}$$

and

$$\Pr(|\hat{\pi}_k - \pi_k| \geq \pi_k\epsilon) \leq 2\exp\{-\frac{n\pi_k\epsilon^2}{3}\}, \tag{44}$$

for any k and $\epsilon > 0$.

Combining Propositions 2 & 4, we have the following result for $\tilde{X}_{\cdot jk}$:

Lemma 2. *For any $j \in \{1,\dots,p\}$, any class k and $i \in \mathcal{C}_k$, let $\tilde{X}_{\cdot jk}$ defined in (8) be the Huber estimator for the class mean μ_{jk}. Then by letting $H = \frac{v^2}{\epsilon}$, where $v \geq \sigma_j$ and $0 < \epsilon \leq \frac{\sqrt{2}}{4}v$, we have*

$$\Pr(|\tilde{X}_{\cdot jk} - \mu_{jk}| \geq 4\epsilon) \leq 2\exp\{-\frac{n\pi_k\epsilon^2}{2v^2}\} + 2\exp\{-\frac{n\pi_k^2}{4}\}. \tag{45}$$

Proof of Lemma 2. By Proposition 2, we have the following concentration inequality for $\tilde{X}_{\cdot jk}$ conditional on Y:

$$\Pr(|\tilde{X}_{\cdot jk} - \mu_{jk}| \geq 4\epsilon \mid Y) \leq 2\exp\{-\frac{n_k\epsilon^2}{v^2}\}.$$

Define $\mathcal{A} = \{n_k\colon n_k \geq n\pi_k/2\}$. Then for the marginal probability, we have

$$\begin{aligned}
\Pr(|\tilde{X}_{\cdot jk} - \mu_{jk}| \geq 4\epsilon) &= \mathrm{E}[\Pr(|\tilde{X}_{\cdot jk} - \mu_{jk}| \geq 4\epsilon \mid Y)] \\
&\leq \mathrm{E}[2\exp\{-\frac{n_k\epsilon^2}{v^2}\}] \\
&= \mathrm{E}[2\exp\{-\frac{n_k\epsilon^2}{v^2}\}I_{\mathcal{A}}] + \mathrm{E}[2\exp\{-\frac{n_k\epsilon^2}{v^2}\}I_{\mathcal{A}^c}] \\
&\leq 2\exp\{-\frac{n\pi_k\epsilon^2}{2v^2}\} + 2\mathrm{E}[I_{\mathcal{A}^c}]
\end{aligned}$$

By (43), $\mathrm{E}[I_{\mathcal{A}^c}] \leq \Pr(n_k \leq n\pi_k/2) \leq \exp\{-\frac{n\pi_k^2}{4}\}$. Thus,

$$\Pr(|\tilde{X}_{\cdot jk} - \mu_{jk}| \geq 4\epsilon) \leq 2\exp\{-\frac{n\pi_k\epsilon^2}{2v^2}\} + 2\exp\{-\frac{n\pi_k^2}{4}\}. \qquad \square$$

By applying the same strategy on the Huber estimator for $Q_{jk} = \mathrm{E}X_{ij}^2$, we obtain the following result.

Lemma 3. *For any $j \in \{1,\dots,p\}$, any class k and $i \in \mathcal{C}_k$, let $\tilde{Q}_{jk}$ defined in (8) be the Huber estimator for the second moment $Q_{jk} = \mathrm{E}X_{ij}^2$. Then by letting $H = \frac{v^2}{\epsilon}$, where $v \geq \max\left\{\sigma_j, \sqrt{\mathrm{Var}(X_{ij}^2)}\right\}$ and $0 < \epsilon \leq \frac{\sqrt{2}}{4}v$, we have*

$$\Pr(|\tilde{Q}_{jk} - Q_{jk}| \geq 4\epsilon) \leq 2\exp\{-\frac{n\pi_k\epsilon^2}{2v^2}\} + 2\exp\{-\frac{n\pi_k^2}{4}\}. \tag{46}$$

Proof of Lemma 3. By Proposition 3, we have the following concentration inequality for $\tilde{Q}_{jk}$ conditional on Y:

$$\Pr(|\tilde{Q}_{jk} - Q_{jk}| \geq 4\epsilon \mid Y) \leq 2\exp\{-\frac{n_k\epsilon^2}{v^2}\}.$$

Then following the same method in the proof of Lemma 2, we have the result

$$\Pr(|\tilde{Q}_{jk} - Q_{jk}| \geq 4\epsilon) \leq 2\exp\{-\frac{n\pi_k\epsilon^2}{2v^2}\} + 2\exp\{-\frac{n\pi_k^2}{4}\}. \qquad \square$$

Combining the results in Lemmas 2 & 3, we now can obtain the concentration inequality for $\tilde{S}_j$.

Lemma 4. *Assume that (A1)-(A3) and (A5) hold. Define $\tau(\epsilon)$ as in (39) and $\tilde{S}_j$ as in (9).*

Then, by letting $H = \frac{v^2}{\epsilon}$, where $v \geq \max\{\sigma_j, \kappa\}$ and $0 < \epsilon \leq \min\{\frac{1}{2}, \frac{\sqrt{2}}{4}v\}$, we have

$$\Pr\left(\bigcup_{k=1}^{K}\left\{\{|\tilde{X}_{\cdot jk} - \mu_{jk}| \geq 4\epsilon\} \cup \{|\hat{\pi}_k - \pi_k| \geq \epsilon\}\right\} \cup \{|\tilde{S}_j - \sigma_j| \geq C_1\epsilon\}\right) \leq \tau(\epsilon), \tag{47}$$

where C_1 is defined in (40).

Proof of Lemma 4. We start from the event $\{|\tilde{S}_j - \sigma_j| \geq C_1\epsilon\}$. Since $u \leq \sigma_j$ for all j when (A5) holds, $u \leq \tilde{S}_j + \sigma_j$ as well. Then we have

$$\begin{aligned}
&\{|\tilde{S}_j - \sigma_j| \geq C_1\epsilon\} \\
\subset &\{|\tilde{S}_j - \sigma_j|(\tilde{S}_j + \sigma_j) \geq uC_1\epsilon\} \\
\subset &\{|\tilde{S}_j^2 - \sigma_j^2| \geq K[(8\zeta + 20) + U^2]\epsilon\}.
\end{aligned}$$

By the definition of $\tilde{S}_j^2$, we know

$$\begin{aligned}
&\{|\tilde{S}_j^2 - \sigma_j^2| \geq K[(8\zeta + 20) + U^2]\epsilon\} \\
\subset &\bigcup_{k=1}^{K}\{|\hat{\pi}_k(\tilde{Q}_{jk} - \tilde{X}_{\cdot jk}^2) - \pi_k\sigma_j^2| \geq [(8\zeta + 20) + U^2]\epsilon\} \\
\subset &\bigcup_{k=1}^{K}\left\{\{|\hat{\pi}_k(\tilde{Q}_{jk} - \tilde{X}_{\cdot jk}^2) - \hat{\pi}_k\sigma_j^2| \geq (8\zeta + 20)\epsilon\} \cup \{|\hat{\pi}_k\sigma_j^2 - \pi_k\sigma_j^2| \geq U^2\epsilon\}\right\} \\
= &\bigcup_{k=1}^{K}\left(\mathcal{A}_{1,k} \cup \mathcal{A}_{2,k}\right),
\end{aligned}$$

Where $\mathcal{A}_{1,k} = \{|\hat{\pi}_k(\tilde{Q}_{jk} - \tilde{X}_{\cdot jk}^2) - \hat{\pi}_k\sigma_j^2| \geq (8\zeta + 20)\epsilon\}$ and $\mathcal{A}_{2,k} = \{|\hat{\pi}_k\sigma_j^2 - \pi_k\sigma_j^2| \geq U^2\epsilon\}$. We consider $\{\mathcal{A}_{1,k}\}$ first. Since $\hat{\pi}_k \leq 1$,

$$\mathcal{A}_{1,k} \subset \{|(\tilde{Q}_{jk} - \tilde{X}_{\cdot jk}^2) - \sigma_j^2| \geq (8\zeta + 20)\epsilon\}.$$

Since $\sigma_j^2 = Q_{jk} - \mu_{jk}^2$ for any k,

$$\begin{aligned}
&\{|(\tilde{Q}_{jk} - \tilde{X}_{\cdot jk}^2) - \sigma_j^2| \geq (8\zeta + 20)\epsilon\} \\
\subset &\{|\tilde{Q}_{jk} - Q_{jk}| \geq 4\epsilon\} \cup \{|\tilde{X}_{\cdot jk}^2 - \mu_{jk}^2| \geq 8(\zeta + 2)\epsilon\}.
\end{aligned}$$

We know $\Pr(|\tilde{Q}_{jk} - Q_{jk}| \geq 4\epsilon) \leq 2\exp\{-\frac{n\pi_k\epsilon^2}{2v^2}\} + 2\exp\{-\frac{n\pi_k^2}{4}\}$ by Lemma 3. For the term $\{|\tilde{X}_{\cdot jk}^2 - \mu_{jk}^2| \geq 8(\zeta + 2)\epsilon\}$, we have the following argument.

When the event $\{|\tilde{X}_{\cdot jk} - \mu_{jk}| < 4\epsilon\}$ holds, we can show that

$$\begin{aligned}|\tilde{X}^2_{\cdot jk} - \mu^2_{jk}| &= |\tilde{X}_{\cdot jk} - \mu_{jk}||\tilde{X}_{\cdot jk} + \mu_{jk}| \\ &\le |\tilde{X}_{\cdot jk} - \mu_{jk}|(|\tilde{X}_{\cdot jk} - \mu_{jk}| + 2|\mu_{jk}|) \\ &< 16\epsilon^2 + 8|\mu_{jk}|\epsilon.\end{aligned}$$

Given $\epsilon \le \frac{1}{2}$, we have $\epsilon^2 < \epsilon$ and hence $16\epsilon^2 + 8|\mu_{jk}|\epsilon < 8(\zeta + 2)\epsilon$ when (A1) holds. Then we have

$$\{|\tilde{X}_{\cdot jk} - \mu_{jk}| < 4\epsilon\} \subset \{|\tilde{X}^2_{\cdot jk} - \mu^2_{jk}| < 8(\zeta + 2)\epsilon\}.$$

Inversely,

$$\{|\tilde{X}^2_{\cdot jk} - \mu^2_{jk}| \ge 8(\zeta + 2)\epsilon\} \subset \{|\tilde{X}_{\cdot jk} - \mu_{jk}| \ge 4\epsilon\}.$$

Therefore,

$$\begin{aligned}&\{|\hat{\pi}_k(\tilde{Q}_{jk} - \tilde{X}^2_{\cdot jk}) - \hat{\pi}_k\sigma_j^2| \ge (8\zeta + 20)\epsilon\} \\ \subset &\{|\tilde{Q}_{jk} - Q_{jk}| \ge 4\epsilon\} \cup \{|\tilde{X}_{\cdot jk} - \mu_{jk}| \ge 4\epsilon\}.\end{aligned}$$

On the other hand, since $\sigma_j \le U$ for all j when (A2) holds, for the term $\mathcal{A}_{2,k}$, we have

$$\mathcal{A}_{2,k} \subset \{|\hat{\pi}_k - \pi_k| \ge \epsilon\} \subset \{|\hat{\pi}_k - \pi_k| \ge \pi_k\epsilon\}.$$

Combining these three parts, we obtain

$$\begin{aligned}&\{|\tilde{S}_j - \sigma_j| \ge C_1\epsilon\} \\ \subset &\bigcup_{k=1}^{K} \left\{\{|\tilde{Q}_{jk} - Q_{jk}| \ge 4\epsilon\} \cup \{|\tilde{X}_{\cdot jk} - \mu_{jk}| \ge 4\epsilon\} \cup \{|\hat{\pi}_k - \pi_k| \ge \pi_k\epsilon\}\right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}&\bigcup_{k=1}^{K} \left\{\{|\tilde{X}_{\cdot jk} - \mu_{jk}| \ge 4\epsilon\} \cup \{|\hat{\pi}_k - \pi_k| \ge \epsilon\}\right\} \cup \{|\tilde{S}_j - \sigma_j| \ge C_1\epsilon\} \\ \subset &\bigcup_{k=1}^{K} \left\{\{|\tilde{Q}_{jk} - Q_{jk}| \ge 4\epsilon\} \cup \{|\tilde{X}_{\cdot jk} - \mu_{jk}| \ge 4\epsilon\} \cup \{|\hat{\pi}_k - \pi_k| \ge \pi_k\epsilon\}\right\}.\end{aligned}$$

By the results in Proposition 4 and Lemmas 2 & 3, we can show

$$\begin{aligned}&\Pr\left(\bigcup_{k=1}^{K} \left\{\{|\tilde{X}_{\cdot jk} - \mu_{jk}| \ge 4\epsilon\} \cup \{|\hat{\pi}_k - \pi_k| \ge \epsilon\}\right\} \cup \{|\tilde{S}_j - \sigma_j| \ge C_1\epsilon\}\right) \\ \le &\sum_{k=1}^{K} \left(\Pr(|\tilde{Q}_{jk} - Q_{jk}| \ge 4\epsilon) + \Pr(|\tilde{X}_{\cdot jk} - \mu_{jk}| \ge 4\epsilon) + \Pr(|\hat{\pi}_k - \pi_k| \ge \pi_k\epsilon)\right) \\ \le &\quad 4\sum_{k=1}^{K}[\exp\{-\frac{n\pi_k\epsilon^2}{2v^2}\} + \exp\{-\frac{n\pi_k^2}{4}\} + \exp\{-\frac{n\pi_k\epsilon^2}{3}\}] = \tau(\epsilon). \quad \square\end{aligned}$$

This lemma helps us merge several events that will be simultaneously used in later proofs into one. We can show the convergence of $\tilde{S}_j$ by using it as well.

Lemma 5. *Assume that (A1)-(A3) and (A5) hold. Define* $\tau(\epsilon)$ *as in* (39) *and* $\tilde{S}_j$ *as in* (9).

Then, by letting $H = \frac{v^2}{\epsilon}$, *where* $v \geq \max\{\sigma_j, \kappa\}$ *and* $0 < \epsilon \leq \min\{\frac{1}{2}, \frac{\sqrt{2}}{4}v\}$, *we have*

$$\Pr(|\tilde{S}_j - \sigma_j| \geq C_1\epsilon) \leq \tau(\epsilon), \tag{48}$$

Proof of Lemma 5. This is a direct consequence of Lemma 4. □

B.2. Proof of Theorem 1

With the setup above, we can now give the variable selection result. To prove $\hat{\mathcal{D}} \to \mathcal{D}$, we first show that for all $j \in \mathcal{D}^c$, all $\tilde{d}'_{jk}$ shrink to 0 with a proper choice of Λ, which indicates $\hat{X}'_{\cdot jk} = \hat{X}'_{\cdot j}$ for all k. Then we show that, on the other hand, when $j \in \mathcal{D}$, at least two of $\hat{X}'_{\cdot jk}$ are different for some k_1 and k_2 with probability converging to 1.

We have the following femma for the case where $j \in \mathcal{D}^c$.

Lemma 6. *Let* $j \in \mathcal{D}^c$. *Assume that (A1)-(A5) hold. Let* $H = \frac{v^2}{\epsilon}$, *where* $0 < \epsilon \leq \min\{\frac{1}{2}, \frac{\sqrt{2}}{4}v\}$ *and* $v \geq \max\{\sigma_j, \kappa\}$. *Then with the choice of* $\Lambda \geq \frac{8\epsilon}{m_0 w}$, *where* $m_0 = \sqrt{C/n}$ *and* $w = u - C_1\epsilon$, *we have*

$$\Pr(\tilde{d}'_{jk} \neq 0 \text{ for some } k) \leq C\exp\{-\frac{Cn\epsilon^2}{v^2}\}. \tag{49}$$

Proof of Lemma 6. We can show

$$\Pr(\tilde{d}'_{jk} \neq 0 \text{ for some } k) = \Pr\left(\bigcup_{k=1}^{K}\{\tilde{d}'_{jk} \neq 0\}\right).$$

For a given class k, we have

$$\Pr(\tilde{d}'_{jk} \neq 0) = \Pr(|\tilde{X}_{\cdot jk} - \tilde{X}_{\cdot j}| > m_k\tilde{S}_j\Lambda).$$

According to (A3), all π_k are bounded away from 0 and 1. Thus, given a proper constant C for $m_0 = \sqrt{C/n}$, we have $m_k \geq m_0$ for all k. Then with the choice of $\Lambda \geq \frac{8\epsilon}{m_0 w}$, where $8\epsilon < m_k\tilde{S}_j\Lambda$ when $|\tilde{S}_j - \sigma_j| \leq C_1\epsilon$ holds. Thus,

$$\Pr(|\tilde{X}_{\cdot jk} - \tilde{X}_{\cdot j}| > m_k\tilde{S}_j\Lambda) \leq \Pr(\{|\tilde{X}_{\cdot jk} - \tilde{X}_{\cdot j}| \geq 8\epsilon\} \cup \{|\tilde{S}_j - \sigma_j| \geq C_1\epsilon\})$$

for $j \in \mathcal{D}^c$, $\mu_{j1} = \ldots = \mu_{jK} = \mu_j$. Then we have

$$\begin{aligned}\Pr(|\tilde{X}_{\cdot jk} - \tilde{X}_{\cdot j}| \geq 8\epsilon) &\leq \Pr(|\tilde{X}_{\cdot jk} - \mu_j| + |\tilde{X}_{\cdot j} - \mu_j| \geq 8\epsilon) \\ &\leq \Pr(\{|\tilde{X}_{\cdot jk} - \mu_j| \geq 4\epsilon\} \cup \{|\tilde{X}_{\cdot j} - \mu_j| \geq 4\epsilon\}).\end{aligned}$$

By the definition of $\tilde{X}_{\cdot j}$, we have

$$\{|\tilde{X}_{\cdot j}-\mu_j|\geq 4\epsilon\}=\left\{\sum_{k=1}^{K}\hat{\pi}_k(|\tilde{X}_{\cdot jk}-\mu_j)|\geq 4\epsilon\right\}\subset\bigcup_{k=1}^{K}\left\{|\tilde{X}_{\cdot jk}-\mu_j|\geq 4\epsilon\right\}.$$

Therefore,

$$\Pr(\{|\tilde{X}_{\cdot jk}-\mu_j|\geq 4\epsilon\}\cup\{|\tilde{X}_{\cdot j}-\mu_j|\geq 4\epsilon\})\leq\Pr\left(\bigcup_{k=1}^{K}\{|\tilde{X}_{\cdot jk}-\mu_j|\geq 4\epsilon\}\right),$$

and the LHS of (49) becomes

$$\begin{aligned}&\Pr(\tilde{d}'_{jk}\neq 0\text{ for some }k)=\Pr\left(\bigcup_{k=1}^{K}\{\tilde{d}'_{jk}\neq 0\}\right)\\&\leq\Pr\left(\bigcup_{k=1}^{K}\{|\tilde{X}_{\cdot jk}-\tilde{X}_{\cdot j}|\geq 8\epsilon\}\cup\{|\tilde{S}_j-\sigma_j|\geq C_1\epsilon\}\right)\\&\leq\Pr\left(\bigcup_{k=1}^{K}\{|\tilde{X}_{\cdot jk}-\mu_j|\geq 4\epsilon\}\cup\{|\tilde{S}_j-\sigma_j|\geq C_1\epsilon\}\right).\end{aligned}$$

According to Lemma 4,

$$\Pr\left(\bigcup_{k=1}^{K}\{|\tilde{X}_{\cdot jk}-\mu_j|\geq 4\epsilon\}\cup\{|\tilde{S}_j-\sigma_j|\geq C_1\epsilon\}\right)\leq\tau(\epsilon).$$

Combining the results above, we finally obtain

$$\Pr(\tilde{d}'_{jk}\neq 0\text{ for some }k)=\Pr\left(\bigcup_{k=1}^{K}\{\tilde{d}'_{jk}\neq 0\}\right)\leq\tau(\epsilon)\leq C\exp\{-\frac{Cn\epsilon^2}{v^2}\}.\quad\square$$

The following lemma proves the case where $j\in\mathcal{D}$.

Lemma 7. *Let $j\in\mathcal{D}$. Assume that (A1)-(A5) hold. Define N_j as in (38). Let $H=\frac{v^2}{\epsilon}$, where $v\geq\max\{\sigma_j,\kappa\}$ and $0<\epsilon\leq\min\{\frac{1}{2},\frac{\sqrt{2}}{4}v,N_j/16\}$. Then with the choice of $\Lambda\leq\frac{N_j-8\epsilon}{Cm_0w}$, where $m_0=\sqrt{C/n}$ and $w=u-C_1\epsilon$, we have*

$$\Pr(\hat{X}'_{jk_1}=\hat{X}'_{jk_2}\text{ for any }k_1,\,k_2)\leq C\exp\{-\frac{Cn\epsilon^2}{v^2}\}.\tag{50}$$

Proof of Lemma 7. Among all Huber mean estimators, $\{\tilde{X}_{\cdot jk},\ldots,\tilde{X}_{\cdot jK}\}$ for any $j\in\mathcal{D}$, we pick the smallest one, denoted as $\tilde{X}_{jk_a}$, and the largest one, denoted as $\tilde{X}_{jk_b}$. Since $\tilde{X}_{\cdot j}=\sum_{k=1}^{K}\hat{\pi}_k\tilde{X}_{jk}$, we have $\tilde{X}_{jk_a}\leq\tilde{X}_{\cdot j}\leq\tilde{X}_{jk_b}$.

Therefore,

$$\hat{X}'_{jk_b}-\hat{X}'_{jk_a}=(\tilde{X}_{jk_b}-\tilde{X}_{\cdot j}-m_{k_b}\tilde{S}_j\Lambda)_+ +(\tilde{X}_{\cdot j}-\tilde{X}_{jk_a}-m_{k_a}\tilde{S}_j\Lambda)_+.$$

Since the threshold function $(x)_+$ is convex and satisfies the property $(ax)_+ = a(x)_+$ for any real number a, we have

$$\begin{aligned}&\frac{1}{2}(\tilde{X}_{jk_b} - \tilde{X}_{\cdot j} - m_{k_b}\tilde{S}_j\Lambda)_+ + \frac{1}{2}(\tilde{X}_{\cdot j} - \tilde{X}_{jk_a} - m_{k_a}\tilde{S}_j\Lambda)_+\\ \geq&(\frac{1}{2}(\tilde{X}_{jk_b} - \tilde{X}_{\cdot j} - m_{k_b}\tilde{S}_j\Lambda + \tilde{X}_{\cdot j} - \tilde{X}_{jk_a} - m_{k_a}\tilde{S}_j\Lambda))_+,\end{aligned}$$

which is equivalent to

$$\begin{aligned}&(\tilde{X}_{jk_b} - \tilde{X}_{\cdot j} - m_{k_b}\tilde{S}_j\Lambda)_+ + (\tilde{X}_{\cdot j} - \tilde{X}_{jk_a} - m_{k_a}\tilde{S}_j\Lambda)_+\\ \geq&(\tilde{X}_{jk_b} - \tilde{X}_{jk_a} - m_{k_b}\tilde{S}_j\Lambda - m_{k_a}\tilde{S}_j\Lambda)_+.\end{aligned}$$

For any $j \in \mathcal{D}$, if the event $\{|\tilde{X}_{\cdot jk} - \mu_{jk}| < 4\epsilon\}$ holds for all k, event $\{|\tilde{S}_j - \sigma_j| < C_1\epsilon\}$ and Assumptions (A1) and (A2) hold, we have $Cm_0w\Lambda \geq (m_{k_a}\tilde{S}_j\Lambda + m_{k_b}\tilde{S}_j\Lambda)$ for some constant $C \geq 2\max_{l_1,l_2}\left\{\sqrt{\frac{1/\pi_{l_1}-1}{1/\pi_{l_2}-1}}\right\}\frac{U+C_1\epsilon}{u-C_1\epsilon}$. Thus

$$\begin{aligned}&(\tilde{X}_{jk_b} - \tilde{X}_{jk_a} - m_{k_b}\tilde{S}_j\Lambda - m_{k_a}\tilde{S}_j\Lambda)_+\\ \geq&(\tilde{X}_{jk_b} - \tilde{X}_{jk_a} - Cm_0w\Lambda)_+\\ =&((\tilde{X}_{jk_b} - \mu_{jk_b}) - (\tilde{X}_{jk_a} - \mu_{jk_a}) + (\mu_{jk_b} - \mu_{jk_a}) - Cm_0w\Lambda)_+.\end{aligned}$$

We now show that $(\mu_{jk_b} - \mu_{jk_a}) > 0$. The fact that $\tilde{X}_{jk_b} - \tilde{X}_{jk_a} \geq 0$ indicates that

$$\tilde{X}_{jk_b} - \tilde{X}_{jk_a} = (\tilde{X}_{jk_b} - \mu_{jk_b}) - (\tilde{X}_{jk_a} - \mu_{jk_a}) + (\mu_{jk_b} - \mu_{jk_a}) \geq 0.$$

for any ϵ. If $\mu_{jk_b} - \mu_{jk_a} < 0$, then $(\mu_{jk_b} - \mu_{jk_a}) \leq -N_j$ by definition. Thus, we have

$$(\tilde{X}_{jk_b} - \mu_{jk_b}) - (\tilde{X}_{jk_a} - \mu_{jk_a}) + (\mu_{jk_b} - \mu_{jk_a}) \leq 8\epsilon - N_j.$$

Under the condition given in the lemma that $\epsilon \leq \frac{N_j}{16}$, we have $\tilde{X}_{jk_b} - \tilde{X}_{jk_a} < 0$, which contradicts to our choices of $\tilde{X}_{jk_b}$ and $\tilde{X}_{jk_a}$. On the other hand, if $\mu_{jk_b} = \mu_{jk_a}$, then since $j \in \mathcal{D}$, there is another μ_{jk_c} such that $\mu_{jk_c} \neq \mu_{jk_b}$. If $\mu_{jk_c} > \mu_{jk_b}$, $\mu_{jk_c} - \mu_{jk_a} \geq N_j$ by definition. Thus, we have

$$\tilde{X}_{jk_c} - \tilde{X}_{jk_b} = (\tilde{X}_{jk_c} - \mu_{jk_c}) - (\tilde{X}_{jk_b} - \mu_{jk_b}) + (\mu_{jk_c} - \mu_{jk_b}) \geq N_j - 8\epsilon > 0$$

when $\epsilon \leq \frac{N_j}{16}$. In this case, $\tilde{X}_{jk_c} > \tilde{X}_{jk_b}$ contradicting to our choice of $\tilde{X}_{jk_b}$. On the other hand, if $\mu_{jk_c} < \mu_{jk_b}$, $\mu_{jk_c} - \mu_{jk_a} \leq N_j$. In this case, we have

$$\tilde{X}_{jk_c} - \tilde{X}_{jk_a} = (\tilde{X}_{jk_c} - \mu_{jk_c}) - (\tilde{X}_{jk_a} - \mu_{jk_a}) + (\mu_{jk_c} - \mu_{jk_a}) \leq 8\epsilon - N_j < 0$$

when $\epsilon \leq \frac{N_j}{16}$. Thus, $\tilde{X}_{jk_c} < \tilde{X}_{jk_a}$, contradicting to our choice of $\tilde{X}_{jk_a}$.

Therefore, $(\mu_{jk_b} - \mu_{jk_a}) > 0$ and we obtain

$$(\tilde{X}_{jk_b} - \tilde{X}_{jk_a} - m_{k_b}\tilde{S}_j\Lambda - m_{k_a}\tilde{S}_j\Lambda)_+ \geq (N_j - 8\epsilon - Cm_0w\Lambda)_+.$$

When $\epsilon \le N_j/16$, with the choice of $0 < \Lambda \le \frac{N_j - 8\epsilon}{Cm_0 w}$, $\hat{X}'_{jk_b} - \hat{X}'_{jk_a} \ge (N_j - 8\epsilon - Cm_0 w\Lambda)_+ > 0$.

Combining the results above with Lemma 4, we reach the result

$$\begin{aligned}
&\Pr(\hat{X}'_{jk_1} = \hat{X}'_{jk_2} \text{ for any } k_1, k_2)\\
\le &\Pr(\{|\tilde{X}_{\cdot jk} - \mu_{jk}| \ge 4\epsilon \text{ for some } k\} \cup \{|\tilde{S}_j - \sigma_j| \ge C_1\epsilon\})\\
\le &\Pr\left(\bigcup_{k=1}^{K}\{|\tilde{X}_{\cdot jk} - \mu_j| \ge 4\epsilon\} \cup \{|\tilde{S}_j - \sigma_j| \ge C_1\epsilon\}\right)\\
\le &\tau(\epsilon) \le C\exp\{-\frac{Cn\epsilon^2}{v^2}\}.
\end{aligned}$$

□

Lemma 7 aims to show that the important variables can be selected as long as all Huber estimators, $\tilde{X}_{\cdot jk}$ and $\tilde{S}_j$, are sufficiently close to the truth. Combining this result with the conclusion in Lemma 6, we finally can prove the result of variable selection.

Proof of Theorem 1. Under the conditions stated in Theorem 1, by the result of Lemma 6 & 7, we have

$$\begin{aligned}
&\Pr(\hat{\mathcal{D}} \ne \mathcal{D})\\
\le &\Pr\left(\bigcup_{j\in\mathcal{D}}\{\hat{X}'_{jk_1} = \hat{X}'_{jk_2} \text{ for any } k_1, k_2\}\right) + \Pr\left(\bigcup_{j\in\mathcal{D}^c}\{\tilde{d}'_{jk} \ne 0 \text{ for some } k\}\right)\\
\le &Cp\exp\{-\frac{Cn\epsilon^2}{v^2}\}.
\end{aligned}$$

Therefore,

$$\Pr(\hat{\mathcal{D}} = \mathcal{D}) \ge 1 - Cp\exp\{-\frac{Cn\epsilon^2}{v^2}\}.$$

Furthermore, assume that $n \to \infty$ and $\frac{\log p}{n} \to 0$. Then when $H \to \infty$, $H \ll \sqrt{\frac{v^2 n}{\log p}}$ and $C\sqrt{\log p} \ll \Lambda \ll CN_0\sqrt{n} - C\sqrt{\log p}$, we let $\epsilon \to 0$ and $\epsilon \gg \sqrt{\frac{\log p}{n}}$. This indicates that $Cp\exp\{-\frac{Cn\epsilon^2}{v^2}\} \to 0$ and $\Pr(\hat{\mathcal{D}} = \mathcal{D}) \to 1$. □

B.3. Proof of Theorems 2 and 3

For the classifier in the nearest shrunken centroids method, we denote

$$\rho_k^{nsc}(\mathbf{X}) = \sum_{j=1}^{p}\frac{(X_j - \mu_{jk})^2}{\sigma_j^2} - 2\log\pi_k \tag{51}$$

and its estimator as

$$\hat{\rho}_k^{nsc}(\mathbf{X}) = \sum_{j=1}^{p}\frac{(X_j - \bar{X}'_{\cdot jk})^2}{S_j^2} - 2\log\hat{\pi}_k. \tag{52}$$

For our proposal, denote

$$\rho_{h,k}(\mathbf{X}) = \sum_{j=1}^{p} f_h(\frac{X_j - \mu_{jk}}{\sigma_j}) - 2\log\pi_k \tag{53}$$

and its estimator as

$$\hat{\rho}_{h,k}(\mathbf{X}) = \sum_{j=1}^{p} f_h(\frac{X_j - \hat{X}'_{\cdot jk}}{\tilde{S}_j}) - 2\log\hat{\pi}_k. \tag{54}$$

For the non-gaussian scenario, the Bayes misclassification rate does not have an explicit form. Therefore, we consider the difference between the classifier and its estimator. Intuitively, $\hat{\delta}_h(\mathbf{X}_*) = \delta_h(\mathbf{X}_*)$ if all the estimators for Huber loss functions of important variables converge to the truth. To give a cleaner expression of the bound, we first show that $\hat{\rho}_{h,k} \to \rho_{h,k}$ for any k, then prove $\Pr\left(\hat{\delta}_h(\mathbf{X}_*) \neq \delta_h(\mathbf{X}_*)\middle|(Y_{tr}, \mathbf{X}_{tr})\right) \to 0$ when $n \to \infty$.

We start from the Huber loss function part.

Proposition 5 (Markov inequality)**.** *Let $X_{*,j}$ be a random variable with mean μ_{jY_*} and fourth moment $\mathrm{E}(\varepsilon_j^*)^4 \leq \kappa^2 < \infty$. Then for any $\epsilon > 0$, we have*

$$\Pr\left(|X_{*,j} - \mu_{jY_*}| \geq \frac{1}{\epsilon^{1/5}}\right) \leq \kappa^2\epsilon^{4/5}.$$

Then we have the following lemma.

Lemma 8. *Assume that (A1)-(A5) hold. Define $\tau(\epsilon)$ as in (39) and $\vartheta(\epsilon)$ as in (40). Let $H = \frac{v^2}{\epsilon}$, where $v \geq \max\{\sigma_j, \kappa\}$ and $0 < \epsilon \leq \min\{\frac{1}{2}, \frac{\sqrt{2}}{4}v, \frac{u}{3C_1}\}$. Then with the choice of $\Lambda = \frac{8\epsilon}{m_0 w}$, where $m_0 = \sqrt{C/n}$ and $w = u - C_1\epsilon$, we have*

$$\Pr\left(\bigcup_{k=1}^{K}\left\{|f_h(\frac{X_{*,j} - \hat{X}'_{\cdot jk}}{\tilde{S}_j}) - f_h(\frac{X_{*,j} - \mu_{jk}}{\sigma_j})| \geq h\vartheta(\epsilon)\right\}\middle|(Y_{tr}, \mathbf{X}_{tr})\right) \leq \kappa^2\epsilon^{4/5} \tag{55}$$

with probability greater than $1 - \tau(\epsilon)$.

Proof of Lemma 8. Since the derivative of the Huber loss $|\frac{df_h}{dx}| \leq h$, we can show that $\left|\frac{X_{*,j} - \hat{X}'_{\cdot jk}}{\tilde{S}_j} - \frac{X_{*,j} - \mu_{jk}}{\sigma_j}\right| \leq \vartheta(\epsilon)$ indicates $|f_h(\frac{X_{*,j} - \hat{X}'_{\cdot jk}}{\tilde{S}_j}) - f_h(\frac{X_{*,j} - \mu_{jk}}{\sigma_j})| \leq h\vartheta(\epsilon)$. Then we can show that

$$\begin{aligned}
&\Pr\left(\bigcup_{k=1}^{K}\left\{|f_h(\frac{X_{*,j} - \hat{X}'_{\cdot jk}}{\tilde{S}_j}) - f_h(\frac{X_{*,j} - \mu_{jk}}{\sigma_j})| \geq h\vartheta(\epsilon)\right\}\middle|(Y_{tr}, \mathbf{X}_{tr})\right) \\
\leq &\Pr\left(\bigcup_{k=1}^{K}\left\{\left|\frac{X_{*,j} - \hat{X}'_{\cdot jk}}{\tilde{S}_j} - \frac{X_{*,j} - \mu_{jk}}{\sigma_j}\right| \geq \vartheta(\epsilon)\right\}\middle|(Y_{tr}, \mathbf{X}_{tr})\right).
\end{aligned}$$

Now consider the event $\left\{\left|\frac{X_{*,j}-\hat{X}'_{\cdot jk}}{\tilde{S}_j}-\frac{X_{*,j}-\mu_{jk}}{\sigma_j}\right|\geq\vartheta(\epsilon)\right\}$. Given the training set $(Y_{tr},\mathbf{X}_{tr})$,

$$\left|\frac{X_{*,j}-\hat{X}'_{\cdot jk}}{\tilde{S}_j}-\frac{X_{*,j}-\mu_{jk}}{\sigma_j}\right|\leq\left|\frac{X_{*,j}(\sigma_j-\tilde{S}_j)}{\tilde{S}_j\sigma_j}\right|+\left|\frac{\mu_{jk}\tilde{S}_j-\hat{X}'_{\cdot jk}\sigma_j}{\tilde{S}_j\sigma_j}\right|.$$

If $|X_{*,j}-\mu_{jY_*}|<\frac{1}{\epsilon^{1/5}}$ and $|\tilde{S}_j-\sigma_j|<C_1\epsilon$, then when (A1) holds, the first term becomes

$$\left|\frac{X_{*,j}(\sigma_j-\tilde{S}_j)}{\tilde{S}_j\sigma_j}\right|=\left|\frac{(\mu_{jY_*}+X_{*,j}-\mu_{jY_*})(\sigma_j-\tilde{S}_j)}{\tilde{S}_j\sigma_j}\right|<\frac{\zeta C_1\epsilon+C_1\epsilon^{4/5}}{w\sigma_j}.$$

On the other hand, since $|\hat{X}'_{\cdot jk}-\mu_{jk}|\leq|\tilde{X}_{\cdot jk}-\mu_{jk}|+m_k\tilde{S}_j\Lambda$, when $|\tilde{X}_{\cdot jk}-\mu_{jk}|<4\epsilon$ and $|\tilde{S}_j-\sigma_j|<C_1\epsilon$, the second term turns to be:

$$\begin{aligned}
&\left|\frac{\mu_{jk}\tilde{S}_j-\hat{X}'_{\cdot jk}\sigma_j}{\tilde{S}_j\sigma_j}\right|\\
\leq&\left|\frac{\mu_{jk}\tilde{S}_j-\hat{X}'_{\cdot jk}\tilde{S}_j}{\tilde{S}_j\sigma_j}\right|+\left|\frac{\hat{X}'_{\cdot jk}\tilde{S}_j-\hat{X}'_{\cdot jk}\sigma_j}{\tilde{S}_j\sigma_j}\right|\\
=&\left|\frac{\mu_{jk}-\hat{X}'_{\cdot jk}}{\sigma_j}\right|+\left|\frac{[\mu_{jk}+(\hat{X}'_{\cdot jk}-\mu_{jk})](\tilde{S}_j-\sigma_j)}{\tilde{S}_j\sigma_j}\right|\\
\leq&\frac{|\tilde{X}_{\cdot jk}-\mu_{jk}|+m_k\tilde{S}_j\Lambda}{\sigma_j}+\frac{\zeta C_1\epsilon}{w\sigma_j}+\frac{(|\tilde{X}_{\cdot jk}-\mu_{jk}|+m_k\tilde{S}_j\Lambda)C_1\epsilon}{w\sigma_j}\\
<&\frac{4\epsilon+m_k\tilde{S}_j\Lambda}{\sigma_j}+\frac{\zeta C_1\epsilon}{w\sigma_j}+\frac{(4\epsilon+m_k\tilde{S}_j\Lambda)C_1\epsilon}{w\sigma_j}.
\end{aligned}$$

Follow the choice of Λ in Theorem 1 and let $\Lambda=\frac{8\epsilon}{m_0w}$, then $m_k\tilde{S}_j\Lambda=\frac{m_k}{m_0}\frac{\tilde{S}_j}{w}8\epsilon$. If $|\tilde{S}_j-\sigma_j|<C_1\epsilon$ and $\epsilon\leq\frac{u}{3C_1}\leq\frac{\sigma_j}{3C_1}$, we have $\frac{\tilde{S}_j}{w}\leq2$; If $|\hat{\pi}_k-\pi_k|<\pi_k\epsilon$ and $\epsilon\leq\frac{1}{2}$, we have $\frac{m_k}{m_0}\leq C_2$. Hence, $m_k\tilde{S}_j\Lambda<16C_2\epsilon$ under these two conditions.

When (A1) and (A2) hold, $|\mu_{jk}|<\zeta$ and $u<\sigma_j<U$. Then we have

$$\begin{aligned}
&\left|\frac{X_{*,j}-\hat{X}'_{\cdot jk}}{\tilde{S}_j}-\frac{X_{*,j}-\mu_{jk}}{\sigma_j}\right|\\
<&\frac{C_1\epsilon^{4/5}+(2\zeta C_1+4w+16wC_2)\epsilon+(4+16C_2)C_1\epsilon^2}{w\sigma_j}\\
\leq&\frac{C_1\epsilon^{4/5}+(2\zeta C_1+4w+16wC_2)\epsilon+(4+16C_2)C_1\epsilon^2}{uw}\\
=&\vartheta(\epsilon).
\end{aligned}$$

Summarizing the conditions above, we have

$$\Pr\left(\bigcup_{k=1}^{K}\left\{\left|\frac{X_{*,j}-\hat{X}'_{\cdot jk}}{\tilde{S}_j}-\frac{X_{*,j}-\mu_{jk}}{\sigma_j}\right|\geq\vartheta(\epsilon)\right\}\middle|(Y_{tr},\mathbf{X}_{tr})\right)$$

$$\leq \Pr\left(|X_{*,j} - \mu_{jY_*}| \geq \frac{1}{\epsilon^{1/5}} \middle| (Y_{tr}, \mathbf{X}_{tr})\right)$$

if $\bigcap_{k=1}^{K} \left\{\{|\tilde{X}_{\cdot jk} - \mu_{jk}| < 4\epsilon\} \cap \{|\hat{\pi}_k - \pi_k| < \epsilon\}\right\} \cap \{|\tilde{S}_j - \sigma_j| < C_1\epsilon\}$ holds.

By the results in Lemma 4 and Proposition 5, with probability greater than $1 - \tau(\epsilon)$, we have

$$\begin{aligned}
&\Pr\left(\bigcup_{k=1}^{K}\left\{|f_h(\frac{X_{*,j} - \hat{X}'_{\cdot jk}}{\tilde{S}_j}) - f_h(\frac{X_{*,j} - \mu_{jk}}{\sigma_j})| \geq h\vartheta(\epsilon)\right\} \middle| (Y_{tr}, \mathbf{X}_{tr})\right) \\
\leq &\Pr\left(\bigcup_{k=1}^{K}\left\{\left|\frac{X_{*,j} - \hat{X}'_{\cdot jk}}{\tilde{S}_j} - \frac{X_{*,j} - \mu_{jk}}{\sigma_j}\right| \geq \vartheta(\epsilon)\right\} \middle| (Y_{tr}, \mathbf{X}_{tr})\right) \\
\leq &\kappa^2\epsilon^{4/5}
\end{aligned}$$

□

With the preparations above, we can eventually obtain the convergence of the classifier.

Proof of Theorem 2. Denote the event

$$\mathcal{B}_1 = \bigcup_{k=1}^{K}\{\min_{l\neq k}\{|\rho_{h,l}(\mathbf{X}_*) - \rho_{h,k}(\mathbf{X}_*)|\} > 4K\epsilon + 2qh\vartheta(\epsilon) \mid Y_* = k\}$$

and

$$\mathcal{B}_2 = \{|\hat{\rho}_{h,k}(\mathbf{X}_*) - \rho_{h,k}(\mathbf{X}_*)| \leq 2K\epsilon + qh\vartheta(\epsilon) \text{ for all } k\},$$

then a sufficient condition for $\{\hat{\delta}_h(\mathbf{X}_*) = \delta_h(\mathbf{X}_*)\}$ is $\mathcal{B}_1 \cap \mathcal{B}_2$. Therefore, we can show that

$$\begin{aligned}
&\Pr\left(\hat{\delta}_h(\mathbf{X}_*) \neq \delta_h(\mathbf{X}_*) \middle| (Y_{tr}, \mathbf{X}_{tr})\right) \leq 1 - \Pr\left(\mathcal{B}_1 \cap \mathcal{B}_2 | (Y_{tr}, \mathbf{X}_{tr})\right) \\
\leq &\sum_{k=1}^{K} \Pr(\min_{l\neq k}\{|\rho_{h,l}(\mathbf{X}_*) - \rho_{h,k}(\mathbf{X}_*)|\} \leq 4K\epsilon + 2qh\vartheta(\epsilon) \mid Y_* = k) \\
&+ \Pr\left(|\hat{\rho}_{h,k}(\mathbf{X}_*) - \rho_{h,k}(\mathbf{X}_*)| > 2K\epsilon + qh\vartheta(\epsilon) \text{ for some } k \mid (Y_{tr}, \mathbf{X}_{tr})\right) \\
= &\sum_{k=1}^{K} \Pr(\mathcal{B}'_{1,k} \mid Y_* = k) + \Pr\left(\mathcal{B}_2^c \mid (Y_{tr}, \mathbf{X}_{tr})\right),
\end{aligned}$$

where $\mathcal{B}'_{1,k} = \{\min_{l\neq k}\{|\rho_{h,l}(\mathbf{X}_*) - \rho_{h,k}(\mathbf{X}_*)|\} \leq 4K\epsilon + 2qh\vartheta(\epsilon)\}$ and $\mathcal{B}_2^c$ is the compliment of $\mathcal{B}_2$. We start with $\sum_{k=1}^{K} \Pr(\mathcal{B}'_{1,k} \mid Y_* = k)$. We have

$$\begin{aligned}
&\Pr(\mathcal{B}'_{1,k} \mid Y_* = k) \\
\leq &\sum_{l\neq k} \Pr(|\rho_{h,l}(\mathbf{X}_*) - \rho_{h,k}(\mathbf{X}_*)| \leq 4K\epsilon + 2qh\vartheta(\epsilon) \mid Y_* = k) \\
\leq &\sum_{l\neq k}\sum_{j\in\mathcal{D}} \Pr\left(|f_h(\frac{\varepsilon_j^* + \mu_{jk} - \mu_{jl}}{\sigma_j}) - f_h(\frac{\varepsilon_j^*}{\sigma_j})| \leq \frac{4K\epsilon + 2qh\vartheta(\epsilon)}{q}\right)
\end{aligned}$$

for each k. Now consider the equation $|f_h(\frac{x+\mu_{jk}-\mu_{jl}}{\sigma_j}) - f_h(\frac{x}{\sigma_j})| = 0$. Based on the value of μ_{jk}, μ_{jl} and h, it can be simplified to a quadratic equation or linear equation, with two solutions $x = a_1(\mu_{jk}, \mu_{jl})$, $x = a_2(\mu_{jk}, \mu_{jl})$ or one solution $x = a_1(\mu_{jk}, \mu_{jl})$, where a_i are constants determined by μ_{jk}, μ_{jl}.

When (A4) holds, the density function $|g_j(\varepsilon_j^*)| \leq V$. Thus in either cases, we have

$$\begin{aligned}
&\Pr\left(|f_h(\frac{\varepsilon_j^* + \mu_{jk} - \mu_{jl}}{\sigma_j}) - f_h(\frac{\varepsilon_j^*}{\sigma_j})| \leq \frac{4K\epsilon + 2qh\vartheta(\epsilon)}{q}\right) \\
\leq &\sum_i \Pr\left(|\varepsilon_j^* - a_i| \leq \frac{\sigma_j(4K\epsilon + 2qh\vartheta(\epsilon))}{q}\right) \\
\leq &\frac{4VU(4K\epsilon + 2qh\vartheta(\epsilon))}{q}.
\end{aligned}$$

Then we can show that

$$\sum_{k=1}^{K} \Pr(\mathcal{B}'_{1,k} \mid Y_* = k) \leq 4KVU(4K\epsilon + 2qh\vartheta(\epsilon)).$$

For $\mathcal{B}_2^c$, we can show

$$\begin{aligned}
\mathcal{B}_2^c \subset \Bigg[&\bigcup_{k=1}^{K} \left\{ \sum_{j=1}^{p} |f_h(\frac{X_{*,j} - \hat{X}'_{\cdot jk}}{\tilde{S}_j}) - f_h(\frac{X_{*,j} - \mu_{jk}}{\sigma_j})| \geq qh\vartheta(\epsilon) \right\} \\
&\bigcup_{k=1}^{K} \{|\log \hat{\pi}_k - \log \pi_k| > 2\epsilon\} \Bigg].
\end{aligned}$$

We first consider the prior term. Since

$$|\log \hat{\pi}_k - \log \pi_k| = |\log(1 + \frac{\hat{\pi}_k - \pi_k}{\pi_k})| \leq |\frac{\hat{\pi}_k - \pi_k}{\pi_k}|,$$

we have $\{|2\log \hat{\pi}_k - 2\log \pi_k| \geq 2\epsilon\} \subset \{|\hat{\pi}_k - \pi_k| \geq \pi_k\epsilon\}$.

Therefore, under the conditions where $\{|\tilde{X}_{\cdot jk} - \mu_{jk}| < 4\epsilon\}$, $\{|\hat{\pi}_k - \pi_k| < \pi_k\epsilon\}$, $\{|\tilde{S}_j - \sigma_j| < C_1\epsilon\}$, and $\{\hat{\mathcal{D}} = \mathcal{D}\}$ for all j and k, we have

$$\mathcal{B}_2^c \subset \bigcup_{j\in\mathcal{D}} \bigcup_{k=1}^{K} \left\{ |f_h(\frac{X_{*,j} - \hat{X}'_{\cdot jk}}{\tilde{S}_j}) - f_h(\frac{X_{*,j} - \mu_{jk}}{\sigma_j})| \geq h\vartheta(\epsilon) \right\}.$$

According to the conclusions in Lemmas 5 & 8, Proposition 4 and Theorem 1, with probability greater than $1 - (q+p)C\exp\{-\frac{Cn\epsilon^2}{v^2}\}$, we now have

$$\begin{aligned}
&\Pr\left(\mathcal{B}_2^c \mid (Y_{tr}, \mathbf{X}_{tr})\right) \\
\leq &\sum_{j\in\mathcal{D}} \Pr\left(\bigcup_{k=1}^{K} \left\{ |f_h(\frac{X_{*,j} - \hat{X}'_{\cdot jk}}{\tilde{S}_j}) - f_h(\frac{X_{*,j} - \mu_{jk}}{\sigma_j})| \geq h\vartheta(\epsilon) \right\} \middle| (Y_{tr}, \mathbf{X}_{tr}) \right)
\end{aligned}$$

$\leq q\kappa^2\epsilon^{4/5}$.

Combining the results on $\mathcal{B}'_{1,k}$ and $\mathcal{B}_2$, with probability greater than $1-(q+p)C\exp\{-\frac{Cn\epsilon^2}{v^2}\}$, we have

$$\begin{aligned}&\Pr\left(\hat{\delta}_h(\mathbf{X}_*)\neq\delta_h(\mathbf{X}_*)\middle|(Y_{tr},\mathbf{X}_{tr})\right)\\ \leq& q\kappa^2\epsilon^{4/5}+4KVU(4K\epsilon+2qh\vartheta(\epsilon))\\ \leq& qhC\epsilon^{4/5}\end{aligned}$$

Furthermore, when $n\to\infty$ and $\frac{q^{5/2}h^{5/2}\log p}{n}\to 0$ for any $h>0$, by letting $H\to\infty$ while $H\ll\sqrt{\frac{v^2n}{\log p}}$, $C\frac{\Lambda}{\sqrt{n}}\to 0$ while $\Lambda\gg C\sqrt{\log p}$, we have $\epsilon^{4/5}\to 0$ and $\epsilon^{4/5}\gg(\frac{\log p}{n})^{2/5}$. Then $qhC\epsilon^{4/5}\to 0$, which indicates that our estimator for the classifier consistently gives the same result as the true one. □

We then prove a special case where we have the normality assumption. The following proposition is needed in the proof of Theorem 3.

Proposition 6 (Lemma 11 in [33])**.** *Let Φ and ϕ denote the distribution and density functions of a standard Gaussian random variable. Then*

$$\frac{\phi(t)}{2t}\leq 1-\Phi(t)\leq\frac{\phi(t)}{t}\ \text{if } t\geq 1. \tag{56}$$

Proof of Theorem 3. [49] points out that the NSC classifier is equivalent to the linear discriminant analysis classifier when the covariance matrix in LDA is diagonal. Given the condition that all predictors are independent, we have

$$\Pr\left(\hat{\delta}_h(\mathbf{X}_*)\neq\delta_{bayes}(\mathbf{X}_*)\middle|(Y_{tr},\mathbf{X}_{tr})\right)=\Pr\left(\hat{\delta}_h(\mathbf{X}_*)\neq\delta_{nsc}(\mathbf{X}_*)\middle|(Y_{tr},\mathbf{X}_{tr})\right).$$

Denote the event

$$\mathcal{B}=\left\{|\frac{X_{*,j}-\hat{X}'_{\cdot jk}}{\tilde{S}_j}|\leq h\text{ for any }k\text{ and }j\in\hat{\mathcal{D}}\right\},$$

then we have

$$\begin{aligned}&\Pr\left(\hat{\delta}_h(\mathbf{X}_*)\neq\delta_{nsc}(\mathbf{X}_*)\middle|(Y_{tr},\mathbf{X}_{tr})\right)\\ =&\Pr\left(\{\hat{\delta}_h(\mathbf{X}_*)\neq\delta_{nsc}(\mathbf{X}_*)\}\cap\mathcal{B}\middle|(Y_{tr},\mathbf{X}_{tr})\right)\\ +&\Pr\left(\{\hat{\delta}_h(\mathbf{X}_*)\neq\delta_{nsc}(\mathbf{X}_*)\}\cap\mathcal{B}^c\middle|(Y_{tr},\mathbf{X}_{tr})\right)\\ \leq&\Pr\left(\{\hat{\delta}_h(\mathbf{X}_*)\neq\delta_{nsc}(\mathbf{X}_*)\}\cap\mathcal{B}\middle|(Y_{tr},\mathbf{X}_{tr})\right)\\ +&\Pr\left(\mathcal{B}^c\mid(Y_{tr},\mathbf{X}_{tr})\right).\end{aligned}$$

Denote the event

$$\hat{\mathcal{B}}' = \bigcap_{k=1}^{K} \left\{ \{|\tilde{X}_{\cdot jk} - \mu_{jk}| < 4\epsilon\} \cap \{|\hat{\pi}_k - \pi_k| < \epsilon\} \right\} \cap \{|\tilde{S}_j - \sigma_j| < C_1\epsilon\} \cap \{\hat{\mathcal{D}} = \mathcal{D}\}.$$

Since $X_{*,j} = \varepsilon_j^* + \mu_{jY_*}$ and $|\frac{X_{*,j} - \hat{X}'_{\cdot jk}}{\tilde{S}_j}| = |\frac{X_{*,j} - \mu_{jY_*} + \mu_{jY_*} - \mu_{jk} + \mu_{jk} - \hat{X}'_{\cdot jk}}{\tilde{S}_j}|$, it can be bounded by

$$\begin{aligned}
&\frac{\sigma_j}{\tilde{S}_j} \left(|\frac{\varepsilon_j^*}{\sigma_j}| - |\frac{\mu_{jY_*} - \mu_{jk}}{\sigma_j}| - |\frac{\mu_{jk} - \hat{X}'_{\cdot jk}}{\sigma_j}| \right) \\
\leq &|\frac{X_{*,j} - \mu_{jY_*} + \mu_{jY_*} - \mu_{jk} + \mu_{jk} - \hat{X}'_{\cdot jk}}{\tilde{S}_j}| \\
\leq &\frac{\sigma_j}{\tilde{S}_j} \left(|\frac{\varepsilon_j^*}{\sigma_j}| + |\frac{\mu_{jY_*} - \mu_{jk}}{\sigma_j}| + |\frac{\mu_{jk} - \hat{X}'_{\cdot jk}}{\sigma_j}| \right).
\end{aligned}$$

If the event $\hat{\mathcal{B}}'$ and Assumptions (A1) and (A2) hold, we have $|\mu_{jk}| \leq \zeta$ and $0 < u < \sigma_j$. Then when $\epsilon \leq \frac{1}{8}u$, we can show

$$\left\{ |\frac{\varepsilon_j^*}{\sigma_j}| \leq \frac{1}{2}h - \frac{2\zeta + C\epsilon}{u} \right\} \subset \left\{ |\frac{X_{*,j} - \hat{X}'_{\cdot jk}}{\tilde{S}_j}| \leq h \right\} \subset \left\{ |\frac{\varepsilon_j^*}{\sigma_j}| \leq 2h + \frac{2\zeta + C\epsilon}{u} \right\}$$

The conclusions above together show that when the event $\hat{\mathcal{B}}'$ and Assumptions (A1) and (A2) hold, we have

$$\mathcal{B}^c \subset \bigcup_{j \in \mathcal{D}} \left\{ |\frac{\varepsilon_j^*}{\sigma_j}| > \frac{1}{2}h - C \right\}$$

For some constants C. Since $\frac{\varepsilon_j^*}{\sigma_j}$ is a standard Gaussian random variable, by Proposition 6, we have

$$\Pr\left(|\frac{\varepsilon_j^*}{\sigma_j}| > \frac{1}{2}h - C \right) \leq \frac{\phi(Ch - C)}{Ch - C}$$

Hence, by Theorem 1 and Lemma 6, with probability greater than $1 - (q + p)C\exp\{-\frac{Cn\epsilon^2}{v^2}\}$, we have

$$\Pr\left(\mathcal{B}^c \mid (Y_{tr}, \mathbf{X}_{tr})\right) \leq q \left(\frac{\phi(Ch - C)}{Ch - C} \right).$$

We now consider the term $\Pr\left(\{\hat{\delta}_h(\mathbf{X}_*) \neq \delta_{nsc}(\mathbf{X}_*)\} \cap \mathcal{B} \middle| (Y_{tr}, \mathbf{X}_{tr}) \right)$. $\hat{\delta}_h$ and δ_{nsc} are the same as the classifiers in the Linear Discriminant Analysis (LDA) when the event $\mathcal{B}$ holds. We now can prove that this probability converges to 1.

For $\delta_{nsc}(\mathbf{X}_*)$, choose $k=1$ as a base line, then

$$\rho_k^{nsc}(\mathbf{X}_*)-\rho_1^{nsc}(\mathbf{X}_*)=2\sum_{j\in\mathcal{D}}\frac{(X_{*,j}-\frac{\mu_{jk}+\mu_{j1}}{2})(\mu_{j1}-\mu_{jk})}{\sigma_j^2}-2\log\frac{\pi_k}{\pi_1}=-2l_k,$$

and we have

$$\delta_{nsc}(\mathbf{X}_*)=\arg\max_k\{l_k\}.$$

Following the same procedure, let

$$\hat{l}_k=\sum_{j\in\hat{\mathcal{D}}}\frac{(X_{*,j}-\frac{\hat{X}'_{\cdot jk}+\hat{X}'_{\cdot j1}}{2})(\hat{X}'_{\cdot jk}-\hat{X}'_{\cdot j1})}{\tilde{S}_j^2}+\log\frac{\hat{\pi}_k}{\hat{\pi}_1},$$

then we have

$$\hat{\delta}_h(\mathbf{X}_*)=\arg\max_k\{\hat{l}_k\}$$

when event $\mathcal{B}$ holds.

Therefore, for any $\Gamma(\epsilon)>0$, we have

$$\begin{aligned}
&\Pr\left(\{\hat{\delta}_h(\mathbf{X}_*)\neq\delta_{nsc}(\mathbf{X}_*)\}\cap\mathcal{B}\Big|(Y_{tr},\mathbf{X}_{tr})\right)\\
\leq&1-\Pr\left(\{|\hat{l}_{k_1}-l_{k_1}|\leq\Gamma(\epsilon)/2,|l_{k_1}-l_{k_2}|\geq\Gamma(\epsilon),\text{for any } k_1,\,k_2\}\cap\mathcal{B}\Big|(Y_{tr},\mathbf{X}_{tr})\right)\\
\leq&\Pr\left(|\hat{l}_{k_1}-l_{k_1}|\geq\Gamma(\epsilon)/2,\text{for some } k_1\Big|(Y_{tr},\mathbf{X}_{tr})\right)\\
&+\Pr\left(|l_{k_1}-l_{k_2}|\leq\Gamma(\epsilon),\text{for some } k_1,\,k_2|(Y_{tr},\mathbf{X}_{tr})\right).
\end{aligned}$$

Denote $\theta_{jk}=\frac{\mu_{jk}-\mu_{j1}}{\sigma_j^2}$ and $\hat{\theta}_{jk}=\frac{(\hat{X}'_{\cdot jk}-\hat{X}'_{\cdot j1})}{\tilde{S}_j^2}$, then $l_{k_1}-l_{k_2}$ is normally distributed with variance $\sum_{j\in\mathcal{D}}\sigma_j^2(\theta_{jk_1}-\theta_{jk_2})^2$ given $Y_*=k$ for any k. Thus,

$$\begin{aligned}
&\Pr\left(|l_{k_1}-l_{k_2}|\leq\Gamma(\epsilon),\text{for some } k_1,\,k_2|(Y_{tr},\mathbf{X}_{tr})\right)\\
\leq&\sum_k\Pr\left(|l_{k_1}-l_{k_2}|\leq\Gamma(\epsilon)|Y_*=k,(Y_{tr},\mathbf{X}_{tr})\right)\pi_k\\
\leq&\sum_{k_1,k_2,k}\frac{C\Gamma(\epsilon)}{\sqrt{\sum_{j\in\mathcal{D}}\sigma_j^2(\theta_{jk_1}-\theta_{jk_2})^2}}\\
\leq&C\Gamma(\epsilon).
\end{aligned}$$

Similarly, for $\Pr\left(|\hat{l}_{k_1}-l_{k_1}|\geq\Gamma(\epsilon)/2,\text{for some } k_1\Big|(Y_{tr},\mathbf{X}_{tr})\right)$, under the condition $(Y_{tr},\mathbf{X}_{tr})$, $\hat{l}_{k_1}-l_{k_1}$ is normally distributed given $Y_*=k$ for any k. If the

event $\hat{\mathcal{B}}'$ holds, then $\hat{\mathcal{D}} = \mathcal{D}$ and its distribution has mean

$$\begin{aligned}\mu_{k_1,k} =& \sum_{j\in\mathcal{D}} \mu_{jk}(\hat{\theta}_{jk_1} - \theta_{jk_1}) \\ &+ \sum_{j\in\mathcal{D}} \frac{1}{2}\left[(\mu_{jk_1} + \mu_{j1})\theta_{jk_1} - (\hat{X}'_{\cdot jk_1}\hat{X}'_{\cdot j1})\hat{\theta}_{jk_1}\right] \\ &+ \log\frac{\pi_{k_1}}{\pi_1} - \log\frac{\hat{\pi}_{k_1}}{\hat{\pi}_1}\end{aligned}$$

and variance $\sum_{j\in\mathcal{D}} \sigma_j^2(\hat{\theta}_{jk_1} - \theta_{jk_1})^2$.

By Markov's inequality, we can obtain

$$\begin{aligned}&\Pr\left(|\hat{l}_{k_1} - l_{k_1}| \geq \Gamma(\epsilon)/2, \text{for some } k_1 \middle| (Y_{tr}, \mathbf{X}_{tr})\right) \\ \leq& \sum_k \Pr\left(|\hat{l}_{k_1} - l_{k_1}| \geq \Gamma(\epsilon)/2 \middle| Y_* = k, (Y_{tr}, \mathbf{X}_{tr})\right)\pi_k \\ \leq& C\mathrm{E}\left\{\frac{\max_{k_1}\{\sum_{j\in\mathcal{D}} \sigma_j^2(\hat{\theta}_{jk_1} - \theta_{jk_1})^2\}}{(\Gamma(\epsilon) - \mu_{k_1,k})^2}\right\}.\end{aligned}$$

We can show

$$|\hat{\theta}_{jk_1} - \theta_{jk_1}| \leq |\frac{\hat{X}'_{\cdot jk_1}}{\tilde{S}_j^2} - \frac{\mu_{jk_1}}{\sigma_j^2}| + |\frac{\hat{X}'_{\cdot j1}}{\tilde{S}_j^2} - \frac{\mu_{j1}}{\sigma_j^2}|$$

and

$$\begin{aligned}|\frac{\hat{X}'_{\cdot jk}}{\tilde{S}_j^2} - \frac{\mu_{jk}}{\sigma_j^2}| &\leq |\frac{\hat{X}'_{\cdot jk}}{\tilde{S}_j^2} - \frac{\mu_{jk}}{\tilde{S}_j^2}| + |\frac{\mu_{jk}}{\tilde{S}_j^2} - \frac{\mu_{jk}}{\sigma_j^2}| \\ &\leq \frac{1}{\tilde{S}_j^2}(4\epsilon + m_k\tilde{S}_j\Lambda) + \mu_{jk}\left(\frac{S_j \tilde{+} \sigma_j}{\tilde{S}_j^2\sigma_j^2}\right)C\epsilon\end{aligned}$$

for any k.

Follow the choice of Λ in Theorem 1 and let $\Lambda = \frac{8\epsilon}{m_0 w}$, then $m_k\tilde{S}_j\Lambda = \frac{m_k}{m_0}\frac{\tilde{S}_j}{w}8\epsilon$. If $|\tilde{S}_j - \sigma_j| < C_1\epsilon$ and $\epsilon \leq \frac{u}{3C_1} \leq \frac{\sigma_j}{3C_1}$, we have $\frac{\tilde{S}_j}{w} \leq 2$; If $|\hat{\pi}_k - \pi_k| < \pi_k\epsilon$ and $\epsilon \leq \frac{1}{2}$, we have $\frac{m_k}{m_0} \leq C_2$. Hence, $m_k\tilde{S}_j\Lambda < 16C_2\epsilon$ under these two conditions, which indicates $|\hat{\theta}_{jk_1} - \theta_{jk_1}| \leq C\epsilon$.

Now we can show that

$$\sum_{j\in\mathcal{D}} \sigma_j^2(\hat{\theta}_{jk_1} - \theta_{jk_1})^2 \leq qC\epsilon^2$$

and

$$|\mu_{k_1,k}| \leq C\epsilon$$

when the event $\hat{\mathcal{B}}'$ and Assumptions (A1) and (A2) hold.

Then for any $\epsilon < 1$, let $\Gamma(\epsilon) = (q\epsilon^2)^{\frac{1}{4}}$, we have $\Gamma(\epsilon) \geq C\epsilon$. Therefore,

$$\Pr\left(|\hat{l}_{k_1} - l_{k_1}| \geq \Gamma(\epsilon)/2, \text{for some } k_1 \middle| (Y_{tr}, \mathbf{X}_{tr})\right) \leq Cq^{\frac{1}{2}}\epsilon.$$

Combining the results above, by the result in Lemma 5 and Theorem 1, with probability greater than $1 - (q+p)C\exp\{-\frac{Cn\epsilon^2}{v^2}\}$, we have

$$\begin{aligned}\Pr\left(\hat{\delta}_h(\mathbf{X}_*) \neq \delta_{bayes}(\mathbf{X}_*) \middle| (Y_{tr}, \mathbf{X}_{tr})\right) \leq & \Gamma(\epsilon) + Cq^{\frac{1}{2}}\epsilon + q\left(\frac{\phi(Ch - C)}{Ch - C}\right) \\ \leq & C(q\epsilon^2)^{\frac{1}{4}} + q\left(\frac{\phi(Ch - C)}{Ch - C}\right).\end{aligned}$$

Furthermore, assume that $n \to \infty$ and $\frac{q\log p}{n} \to 0$. By letting $H \to \infty$ while $H \ll \sqrt{\frac{v^2 n}{\log p}}$, $C\frac{\Lambda}{\sqrt{n}} \to 0$ while $\Lambda \gg C\sqrt{\log p}$ and $h \to \infty$ while $h \gg \sqrt{\log q}$, we have $\epsilon \to 0$ and $\epsilon \gg \sqrt{\frac{\log p}{n}}$. Hence, we have $C(q\epsilon^2)^{\frac{1}{4}} \to 0$ and $q\left(\frac{\phi(Ch-C)}{Ch-C}\right) \to 0$, which indicates that our estimator for the classifier consistently gives the same result as the true one. □

Acknowledgments

The authors thank the editor, the associate editor, and the referee, whose comments led to significant improvements of this paper.

References

[1] Amin, K. M., Litzky, L. A., Smythe, W. R., Mooney, A. M., Morris, J. M., Mews, D. J., Pass, H. I., Kari, C., Rodeck, U., Rauscher III, F. J. et al. (1995). Wilms' tumor 1 susceptibility (WT1) gene products are selectively expressed in malignant mesothelioma. *The American Journal of Pathology* **146** 344–356.

[2] Avella-Medina, M., Battey, H. S., Fan, J. and Li, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika* **105** 271–284. MR3804402

[3] Belloni, A., Chernozhukov, V. et al. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics* **39** 82–130. MR2797841

[4] Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory. COLT '92* 144–152.

[5] Bradley, P. S. and Mangasarian, O. L. (1998). Feature selection via concave minimization and support vector machines. In *Machine Learning Proceedings of the Fifteenth International Conference* **98** 82–90.

[6] Branden, K. V. and Hubert, M. (2005). Robust classification in high dimensions based on the SIMCA method. *Chemometrics and Intelligent Laboratory Systems* **79** 10–21.
[7] Bubeck, S., Cesa-Bianchi, N. and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory* **59** 7711–7717. MR3124669
[8] Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association* **106** 1566–1577. MR2896857
[9] Cantoni, E. and Ronchetti, E. (2001). Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing* **11** 141–146. MR1837133
[10] Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **48** 1148–1185. MR3052407
[11] Cho, E., Cho, M. J. and Eltinge, J. (2005). The variance of sample variance from a finite population. *International Journal of Pure and Applied Mathematics* **21** 389–396. MR2153293
[12] Choi, B. Y., Bair, E. and Lee, J. W. (2017). Nearest shrunken centroids via alternative genewise shrinkages. *PLOS ONE* **12** 1–18.
[13] Clemmensen, L., Hastie, T., Witten, D. and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics* **53** 406–413. MR2850472
[14] Şimşek, G. Ö., Ağababaoğlu, İ., Dursun, D., Özekinci, S., Erçetin, P., Ellidokuz, H., Aktaş, S., Gürel, D., Öztop, İ. and Akkoçlu, A. (2020). Evaluation of gene expression levels in the diagnosis of lung adenocarcinoma and malignant pleural mesothelioma. *Turkish Journal of Thoracic and Cardiovascular Surgery* **28** 188–196.
[15] Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics* **36** 2605–2637. MR2485009
[16] Fan, J., Fan, Y. and Barut, E. (2014). Adaptive robust variable selection. *The Annals of Statistics* **42** 324–351. MR3189488
[17] Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 745–771. MR2965958
[18] Fan, J., Ke, Z. T., Liu, H. and Xia, L. (2015). QUADRO: A supervised dimension reduction method via Rayleigh quotient optimization. *The Annals of Statistics* **43** 1498–1534. MR3357869
[19] Fan, J., Li, Q. and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 247–265. MR3597972
[20] Feng, J., Xu, H., Mannor, S. and Yan, S. (2014). Robust logistic regression and classification. *Advances in Neural Information Processing Systems* **27** 253–261.

[21] Ghosh, A. K., Chaudhuri, P. et al. (2005). On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli* **11** 1–27. MR2121452

[22] Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J. and Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research* **62** 4963–4967.

[23] Hall, P., Titterington, D. and Xue, J.-H. (2009). Median-based classifiers for high-dimensional data. *Journal of the American Statistical Association* **104** 1597–1608. MR2597003

[24] Hastie, T., Tibshirani, R. and Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction. Springer Series in Statistics*. Springer, New York. MR2722294

[25] Huber, P. J. et al. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35** 73–101. MR0161415

[26] Jiang, B., Wang, X. and Leng, C. (2018). A direct approach for sparse quadratic discriminant analysis. *Journal of Machine Learning Research* **19** 1–37. MR3862438

[27] Joly, E. and Lugosi, G. (2016). Robust estimation of U-statistics. *Stochastic Processes and their Applications* **126** 3760–3773. MR3565476

[28] Jörnsten, R. (2004). Clustering and classification based on the L_1 data depth. *Journal of Multivariate Analysis* **90** 67–89. MR2064937

[29] Kettunen, E., Nicholson, A., Nagy, B., Wikman, H., Seppänen, J., Stjernvall, T., Ollikainen, T., Kinnula, V., Nordling, S., Hollmen, J. et al. (2005). L1CAM, INP10, P-cadherin, tPA and ITGB4 overexpression in malignant pleural mesotheliomas revealed by combined use of cDNA and tissue microarray. *Carcinogenesis* **26** 17–25.

[30] Kuramitsu, Y., Miyamoto, H., Tanaka, T., Zhang, X., Fujimoto, M., Ueda, K., Tanaka, T., Hamano, K. and Nakamura, K. (2009). Proteomic differential display analysis identified upregulated astrocytic phosphoprotein PEA-15 in human malignant pleural mesothelioma cell lines. *Proteomics* **9** 5078–5089.

[31] Lambert-Lacroix, S., Zwald, L. et al. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics* **5** 1015–1053. MR2836768

[32] Li, Q. and Shao, J. (2015). Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica* **25** 457–473. MR3379082

[33] Liu, H., Lafferty, J. and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research* **10** 2295–2328. MR2563983

[34] Loh, P.-L. et al. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics* **45** 866–896. MR3650403

[35] Mai, Q., He, D. and Zou, H. Coordinatewise Gaussianization: Theories and applications. *Journal of the American Statistical Association* **in press**.

[36] Mai, Q., Yang, Y. and Zou, H. (2019). Multiclass sparse discriminant analysis. *Statistica Sinica* **29** 97–111. MR3889359

[37] Mai, Q. and Zou, H. (2013). A note on the connection and equivalence of three sparse linear discriminant analysis methods. *Technometrics* **55** 243–246. MR3176524

[38] Mai, Q., Zou, H. and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99** 29–42. MR2899661

[39] Markovich, N. (2008). *Nonparametric analysis of univariate heavy-tailed data: Research and practice. Wiley Series in Probability and Statistics*. John Wiley & Sons, Ltd., Chichester. MR2364666

[40] Meier, L., Van De Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 53–71. MR2412631

[41] Melaiu, O., Melissari, E., Mutti, L., Bracci, E., De Santi, C., Iofrida, C., Di Russo, M., Cristaudo, A., Bonotti, A., Cipollini, M. et al. (2015). Expression status of candidate genes in mesothelioma tissues and cell lines. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **771** 6–12.

[42] Nemirovskij, A. S. and Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization. A Wiley-Interscience Publication*. John Wiley & Sons, Inc., New York. MR0702836

[43] Nolan, J. P. (2020). *Univariate stable distributions: Models for heavy tailed data. Springer Series in Operations Research and Financial Engineering*. Springer, Cham. MR4230105

[44] Pan, Y. and Mai, Q. (2020). Efficient computation for differential network analysis with applications to quadratic discriminant analysis. *Computational Statistics & Data Analysis* **144** 106884. MR4038668

[45] Pan, Y., Mai, Q. and Zhang, X. (2020). TULIP: A toolbox for linear discriminant analysis with penalties. *The R Journal* **12** 134–154.

[46] Sæbø, S., Almøy, T., Aarøe, J. and Aastveit, A. H. (2008). ST-PLS: A multi-directional nearest shrunken centroid type classifier via PLS. *Journal of Chemometrics* **22** 54–62.

[47] Sun, J., Frees, E. W. and Rosenberg, M. A. (2008). Heavy-tailed longitudinal data modeling using copulas. *Insurance: Mathematics and Economics* **42** 817–830. MR2712927

[48] Sun, Q., Zhou, W.-X. and Fan, J. (2020). Adaptive Huber regression. *Journal of the American Statistical Association* **115** 254–265. MR4078461

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

[50] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science* **18** 104 – 117. MR1997067

[51] Tong, T., Chen, L. and Zhao, H. (2012). Improved mean estimation and its application to diagonal discriminant analysis. *Bioinformatics* **28**

531–537.

[52] Veenman, C. J. and Reinders, M. J. (2005). The nearest subclass classifier: A compromise between the nearest mean and nearest neighbor classifier. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27** 1417–1429.

[53] Wang, D., Xiao, H., Devadas, S. and Xu, J. (2020). On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning* 10081–10091.

[54] Wang, L. (2013). The L_1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis* **120** 135-151. MR3072722

[55] Wang, L., Zheng, C., Zhou, W. and Zhou, W.-X. (2021). A new principle for tuning-free Huber regression. *Statistica Sinica* **31** 2153-2177. MR4328856

[56] Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 753–772. MR2867457

[57] Zhang, H. H., Ahn, J., Lin, X. and Park, C. (2006). Gene selection using support vector machines with non-convex penalty. *Bioinformatics* **22** 88–95. MR2408601

[58] Zhou, W., Bose, K., Fan, J. and Liu, H. (2018). A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *The Annals of Statistics* **46** 1904–1931. MR3845005