Received: Added at production | Revised: Added at production | Accepted: Added at production
DOI: xxx/xxxx

ORIGINAL ARTICLE

A High-Dimensional Classification Rule Using Sample Covariance Matrix Equipped With Adjusted Estimated Eigenvalues

Seungchul Baek[1] | Hoyoung Park[2] | Junyong Park*[2]

[1]Department of Mathematics and Statistics, University of Maryland Baltimore County, Maryland, U.S.A.
[2]Department of Statistics, Seoul National University, Seoul, Korea

Correspondence
*Corresponding author Junyong Park, Department of Statistics, Seoul National University, Gwanak-ro 1, Gwanak-gu, Seoul, 08826 Korea. Email: junyongpark@snu.ac.kr

Present Address
Department of Statistics
Seoul National University
Gwanak-ro 1, Gwanak-gu
Seoul, 08826 Korea

Abstract

High-dimensional classification have had challenges mainly due to the singularity issue of the sample covariance matrix. In this work, we propose a different approach to get a more reliable sample covariance matrix by adjusting the estimated eigenvalues. This procedure also brings us a non-singular matrix as a by-product. We improve the optimization procedure to obtain a linear classifier by incorporating the adjusted sample covariance matrix and a shrinkage mean vector into the original optimization problem. We have showed that our proposed binary classification rule is better than some other rules in terms of misclassification rule through most of various synthetic data and real data sets.

KEYWORDS:
Adjusted estimated eigenvalue; High dimensional classification; Linear discriminant analysis

1 | INTRODUCTION

Classification problems have had a broad range of applications in many fields such as biomedical studies, pattern recognition, bioinformatics, as well as in practical subjects in statistics. With technology innovation, high throughput data are common in a lot of areas. However, in high-dimensional problems which can be characterized by $p > n$ where p is the number of covariates and n is the sample size, the classical statistical methods may be not applicable or they don't show a good performance. For an overview of challenges in high-dimensional setting, see Fan and Li (2006) and Aoshima et al. (2018).

In high-dimensional classification problems, a seminal paper, Bickel and Levina (2004), showed that the performance of Fisher discriminant analysis is not better than random guessing under $p/n \to \infty$. As indicated by Bickel and Levina (2004) and Fan and Fan (2008), this phenomenon is from the diverging spectra, which generally arises due to the singularity of the sample covariance matrix in high-dimensional setting. As a remedy, independence rule (IR) or naive Bayes (NB) rule based on ignoring correlations among all covariates have been used in high dimension and its performance has been successful in practical use. Bickel and Levina (2004) explained the superiority of IR over Fisher's discriminant rule in high dimension in theoretical point of view. As variations of IR, there have been many different types of linear classifier in high dimension by modification of the coefficients in IR. In lieu of a subset of important covariates, Fan and Fan (2008) developed a classification method equipped with a variable selection, named as FAIR. Greenshtein, Park, and Lebanon (2009) proposed a method using a variable selection incorporated with the conditional maximum likelihood. Greenshtein and Park (2009) proposed a new naive Bayes type classifier which has a different flavor by using nonparametric empirical Bayes (NPEB) procedure.

In many cases, the independence assumption is too stringent to investigate the true structure of dependency among covariates. For example, in the Gene Set Enrichment Analysis (GSEA) suggested in Mootha et al. (2003), the consideration of the correlations among covariates is essential. More detailed information can be found in Ackermann and Strimmer (2009). There are a lot of literature dealing with high-dimensional classification, which are not based on independence assumption. The shrunken centroids regularized discriminant analysis (SCRDA) proposed in Guo, Hastie, and Tibshirani (2007) used a regularized sample covariance matrix. Shao, Wang, Deng, Wang, et al. (2011) proposed a linear classifier by using regularized sample covariance matrix and mean difference vector. Bair, Hastie, Paul, and Tibshirani (2006) handled a general regression problems in a high-dimensional setting via principal component analysis.

Ahn and Marron (2010) proposed the maximal data piling (MDP) direction vector which makes the distance between the group means largest. Since the MDP forces the data to be completely piled, it is not anticipated to be good, and in this sense we may view it as an extreme version of regularized rule. As a way to adjust the extent of regularization of the data piling, Lee, Ahn, and Jeon (2013) presented a new regularization approach, the regularized data piling (RDP) method, which does an L_2-norm regularization on within-class projections. In addition, they showed that the optimization problem to find a RDP vector led to the solution of a similar type of ridge linear discriminant analysis (rLDA). The ridge linear discriminant analysis is not to represent a specific classifier, but to stand for a class of classifiers whose sample covariance matrix is slightly biased while being made non-singular and stabilzingly estimated. The idea dates back to Hoerl and Kennard (1970) and Di Pillo (1976) in the ridge regression and the fixation of ill-conditioned inverse problem, respectively. Studies for the LDA problems with the ridge correction have been broadly conducted, e.g., Friedman (1989); Guo et al. (2007); Hastie and Tibshirani (2004). In numerical studies, Sections 3 and 4, we compare the performances of our proposed classification rule directly with Naïve Bayes, RDP, and rLDA.

Linear classifiers in high dimension are obtained in two ways: one direction is the modifications of IR by considering different types of estimators of coefficients in IR, for example Fan and Fan (2008) and Greenshtein and Park (2009). These studies focus on the estimation of mean vectors. The other one is based on regularization of sample covariance matrix such as MDP, RDP, rLDA, and related studies, e.g., Ahn and Jeon (2015); Ahn and Marron (2010); Lee et al. (2013). These approaches take into account the regularization of covariance matrix, however the mean vectors are simply estimated using sample mean vectors which may not perform well in high dimension. It is natural to consider a combination of two approaches–the regularization on both of mean vectors and covariance matrix, e.g., Guo et al. (2007). As a further step to this integrated approach, in this paper we propose a way to restore eigenvalues of the covariance matrix more accurately. In other words, we adjust the eigenvalues of the sample covariance matrix so that they are expected to be closer to true ones.

The rest of this article is organized as follows. In Section 2, we describe our proposed binary classification rule with rationales. In order to show a performance of our proposed method with some other existing methods, we conduct simulations under a broad range of cases in Section 3, and do analyses for real data sets in Section 4. We conclude this work with a brief discussion in Section 5.

2 | METHODOLOGY

We consider a binary classification problem. Suppose $\mathbf{X}|\mathsf{Y}=1 \sim \mathsf{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathbf{X}|\mathsf{Y}=2 \sim \mathsf{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ and $\mathbf{x} \in \mathbb{R}^{\mathsf{p}}$. Assuming $\mathsf{P}(\mathsf{Y}=1) = \mathsf{P}(\mathsf{Y}=2) = 1/2$, any linear classification rule δ is

$$\delta(\mathbf{X}) = \mathbb{I}\{\mathbf{w}^{\mathrm{T}}(\mathbf{X} - \boldsymbol{\mu}_{\mathsf{a}}) > 0\}, \tag{1}$$

where $\boldsymbol{\mu}_{\mathsf{a}} = (\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)/2$. A well known criterion, Rayleigh Quotient, to determine a $\mathbf{w}$ is

$$\begin{aligned} \widehat{\mathbf{w}} &= \underset{\mathbf{w}\neq 0}{\mathsf{argmax}} \frac{\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}_{\mathsf{d}}}{\sqrt{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}}} \\ &= \underset{\mathbf{w}\neq 0}{\mathsf{argmax}} \frac{\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}_{\mathsf{d}}\boldsymbol{\mu}_{\mathsf{d}}^{\mathrm{T}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}} \\ &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathsf{d}}, \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}_{\mathsf{d}} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ mean difference vector. This optimization problem is justified in sense that the misclassification rate of δ, which is $1 - \Phi\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}_{\mathsf{d}}/(\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w})^{1/2}\}$, will be minimized.

In practice, $\boldsymbol{\mu}_{\mathsf{d}}$ and $\boldsymbol{\Sigma}$ are unknown and they should be estimated using the data. The IR is obtained by using $\mathbf{D} = \mathsf{diag}(\widehat{\boldsymbol{\Sigma}})$ in (2). The IR is obtained through

$$\widehat{\mathbf{w}}_{\mathsf{IR}} = \underset{\mathbf{w}^*\neq 0}{\mathsf{argmax}} \frac{(\mathbf{w}^*)^{\mathrm{T}}\mathbf{D}^{-1/2}\widehat{\boldsymbol{\mu}}_{\mathsf{d}}\widehat{\boldsymbol{\mu}}_{\mathsf{d}}^{\mathrm{T}}\mathbf{D}^{-1/2}\mathbf{w}^*}{(\mathbf{w}^*)^{\mathrm{T}}\mathbf{w}^*} \tag{3}$$

where $\mathbf{w}^* = \mathbf{D}^{-1/2}\hat{\boldsymbol{\mu}}_{\mathsf{d}}$ and $\widehat{\boldsymbol{\mu}}_{\mathsf{d}} = \overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2$. In addition, the FAIR considers the truncation of $\widehat{\mathbf{w}}_{\mathsf{IR}} = (\hat{\mathsf{w}}_{1,\mathsf{IR}}, \ldots, \hat{\mathsf{w}}_{\mathsf{p},\mathsf{IR}})^{\mathrm{T}}$ which has the form of $\widehat{\mathbf{w}}_{\mathsf{FAIR}} = (\hat{\mathsf{w}}_{1,\mathsf{IR}}\mathsf{I}(|\hat{\mathsf{w}}_{1,\mathsf{IR}}| \geq \mathsf{c}), \ldots, \hat{\mathsf{w}}_{\mathsf{p},\mathsf{IR}}\mathsf{I}(|\mathsf{w}_{\mathsf{p},\mathsf{IR}}| \geq \mathsf{c}))^{\mathrm{T}}$, where c is determined based on some criterion in Fan and Fan (2008). Similarly, Greenshtein and Park (2009) estimated $\mathsf{diag}(\boldsymbol{\Sigma})^{-1/2}\boldsymbol{\mu}_{\mathsf{d}} = \boldsymbol{\xi}$ based on $\mathbf{D}^{-1/2}\widehat{\boldsymbol{\mu}}_{\mathsf{d}}$ by nonparametric empirical Bayes method minimizing square loss and then take $\widehat{\mathbf{w}}_{\mathsf{NPEB}} = \widehat{\boldsymbol{\xi}}$. All these methods such as IR, FAIR and NPEB are assuming the independence of covariates leading to removing the off-diagonal terms in $\boldsymbol{\Sigma}$. Such a simplification of covariance matrix may lead to the poor performance of classification in the presence of correlations among variables. On the other hand, Ahn and Marron (2010), Lee et al. (2013) and Friedman (1989) paid attention to the regularization of estimation of covariance matrix, but they simply used $\widehat{\boldsymbol{\mu}}_{\mathsf{d}} = \overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2$ in (2) which is a bad estimator for the high-dimensional mean vector.

In a high-dimensional setting $p > n$, the optimization problem (2) gets largely involved in the quality of plug-in estimates for $\boldsymbol{\mu}_d$ and $\boldsymbol{\Sigma}$. To obtain better estimates, we employ the regularized mean vector in Section 2.1, and propose a novel way to obtain an adjusted version of the sample covariance matrix in Sections 2.2 and 2.3. Those two estimates lead to coming up with a new classification rule in Section 2.4.

Before we propose our regularization of mean vector and covariance matrix, we introduce some additional notations. We denote $\widehat{\boldsymbol{\Sigma}}$ to be the pooled sample covarince matrix, and let the eigen-decomposition of it as $\widehat{\boldsymbol{\Sigma}} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$, where $\mathbf{D}$ is a $p \times p$ diagonal matrix with elements of ordered eigenvaules d_i, i.e., $d_1 \geq d_2 \geq \cdots \geq d_p$, and $\mathbf{U}$ is a $p \times p$ matrix whose columns are p eigenvectors corresponding to ordered eigenvalues. Similarly, we have the eigen-decomposition of the true covariance as $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix whose elements are ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, and the columns of $\mathbf{V}$ consist of corresponding eigenvectors.

2.1 | Regularizing Mean Vectors

In the optimization problem (2), existing approaches use the estimated direction of $\boldsymbol{\mu}_d$ as $\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2$, however, this direction is not a good direction in general when p is large, which is pointed out by Fan and Fan (2008) and Greenshtein et al. (2009). There are many studies on estimators of the mean vector $\boldsymbol{\mu}$ which are better than the maximum likelihood estimator. Although Shao et al. (2011) proposed a consistent estimator of $\boldsymbol{\mu}$ under sparse structure and some regularity conditions, in practice, there is no guarantee to have a sparse structure. Since a seminal paper of James and Stein (1961), many studies on James-Stein type shrinkage estimator have flourished. See Efron and Morris (1973), Gleser (1986), Fourdrinier, Strawderman, and Wells (2003), and Chételat, Wells, et al. (2012) for more details. Moreover in existing studies, they generally assumed that $\boldsymbol{\Sigma}$ is known or an invertible estimator of $\boldsymbol{\Sigma}$ does exist. It is against common situations in a high-dimensional setting.

Following Tong, Chen, and Zhao (2012), in this work we use a shrinkage mean vector estimator, which is

$$\widehat{\boldsymbol{\mu}}_g = \left\{1 - \frac{(p-2)(n_g-1)}{n_g(n_g-3)\overline{\mathbf{X}}_g^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_d^{-1}\overline{\mathbf{X}}_g}\right\}\overline{\mathbf{X}}_g, \tag{4}$$

where $g = 1, 2$ is a group index and $\widehat{\boldsymbol{\Sigma}}_d$ is a diagonal matrix with elements that are diagonal terms of $\widehat{\boldsymbol{\Sigma}}$. In fact, (4) is a James-Stein type estimator assuming that $\boldsymbol{\Sigma}$ is a diagonal matrix. Tong et al. (2012) showed that even though the independence assumption is too strong to be applied in practice, the usage of the shrinkage mean estimator improved the classification prediction. In addition we reflect a dependent structure of the covariance matrix in a different way, which is described through the following sections, Sections 2.2 and 2.3.

2.2 | Adjusting Estimated Eigenvalues

In a high-dimensional setting, $p > n$, $\widehat{\boldsymbol{\Sigma}}$ is singular, which implies that it is not a good approximation for $\boldsymbol{\Sigma}$. In terms of eigendecomposition, two types of uncertainty are involved in estimating $\boldsymbol{\Sigma}$, which turns out to be errors of eigenvalues and errors of eigenvectors. With a limited data resource, $p > n$, it is impossible to exactly recover all structure of the true covariance matrix. As a remedy that is not complete but meaningful, we may adjust the estimated eigenvalues so that the adjusted eigenvalues are closer to the true eigenvalues component-wise.

Ledoit and Wolf (2004) showed in Lemma 2.1 eigenvalues from the sample covariance are more spread around the mean of true eigenvalues $\bar{\lambda}$ compared to those from the true covariance matrix with the extent of $E\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F^2$, i.e., the following relationship holds

$$E\left\{\sum_{i=1}^{p}(d_i - \bar{\lambda})^2\right\} = \sum_{i=1}^{p}(\lambda_i - \bar{\lambda})^2 + E\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F^2, \tag{5}$$

where $\bar{\lambda} = (1/p)\sum_{i=1}^{p}\lambda_i$, and $\|\cdot\|_F$ is the Frobenius norm. They proposed to regularize the sample covariance by tapering its eigenvalues to its mean, see Ledoit and Wolf (2004) for more details. This rationale seems appealing, but it is limited at the same time. Since (5) is true only for the grand mean $\bar{\lambda}$, it is impossible to assess the differences between individual d_i's and λ_i's. In the case of an identity covariance matrix, a regularization toward the grand mean would be a good way to refine the sample covariance, however, there is an open space to improve if with non-identity covariance matrix. In particular, Ledoit and Wolf (2004) claimed that a tendency of the sample eigenvalues, i.e., the largest sample eigenvalues are positively biased and the smallest ones have an opposite direction, however, it is not always a case. For example, with $p = 500$ and $n_1 = n_2 = 100$, we generate a covariance matrix $\boldsymbol{\Sigma}$ whose diagonal elements is one and other elements are 0.5, and compute the sample covariance matrix. As seen in Figure 1, the largest sample eigenvalue is greater than the largest true one, i.e., $\max_i \lambda_i = 250.5$ and $\max_i d_i = 240.5$. In other words, the largest eigenvalue is negatively biased in this example. Therefore, a shrinkage of eigenvalues to the grand mean proposed in Ledoit and Wolf (2004) is not always justified.

Our regularization for estimation of covariance matrix is done through imposing a condition on the sum of sample eigenvalues from sample covariance matrix. The motivation of this regularization comes from well known results of sample covariance matrix in high dimension. $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}})$ is an unbiased estimator of $\mathrm{tr}(\boldsymbol{\Sigma})$, i.e., $E\{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}})\} = \mathrm{tr}(\boldsymbol{\Sigma})$. Beyond this unbiasedness, $\mathrm{tr}(\boldsymbol{\Sigma})$ has some desirable properties under some conditions such as the ratio consistency, $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}})/\mathrm{tr}(\boldsymbol{\Sigma}) \to 1$ in probability. See Jonsson (1982) for more detail.

With these motivations, we propose the following algorithm to obtain more closer to true eigenvalues, λ_i for $1 \le i \le p$ component-wise. We provided some theoretical justification for this algorithm in Remark 2 in what follows.

Algorithm for Adjustment to the Estimated Eigenvalues

1. $\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} d_i = \text{tr}(\widehat{\boldsymbol{\Sigma}})$.

2. Even though the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ is singular when $p > n$, its structure of eigenvalues $d_1 \ge d_2 \ge \cdots \ge d_{n^*}$ is close to the one of eigenvalues of the true covariance matrix over $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p$, where $p > n \ge n^*$. In general, one is able to determine an appropriate n^* by taking $\text{rank}(\widehat{\boldsymbol{\Sigma}})$. In fact, since $p - \text{rank}(\widehat{\boldsymbol{\Sigma}})$ eigenvalues are exactly zero, we don't need to consider some cases of $n^* > \text{rank}(\widehat{\boldsymbol{\Sigma}})$.

3. In order to identify more accurate structure of eigenvalues, we use a Kernel regression–Nadaraya-Watson by assuming $E(V|Z = z) = m(z)$, where $z \in [1, n^*]$. Now we have

$$\hat{m}(z) = \frac{\sum_{i=1}^{n^*} K_h(z - z_i)v_i}{\sum_{j=1}^{n^*} K_h(z - z_j)},$$

where K_h is a kernel with a bandwidth h. Here v_i is no more than $d_i, i = 1, \ldots, n^*$. Then we extend this structure to p, while holding $\text{tr}(\widehat{\boldsymbol{\Sigma}}) = \sum_{i=1}^{p} d_i$ unchanged. For $z \in [1, p]$, we propose a new estimator for λ's,

$$\hat{\lambda}(z) = \frac{\sum_{i=1}^{n^*} K_h(z - z_i)\tilde{m}(z_i)}{\sum_{j=1}^{n^*} K_h(z - z_j)},$$

for $z = \{1, 2, \ldots, p\}$ and where

$$\tilde{m}(z_i) = \frac{\sum_{k=1}^{p} d_k}{\sum_{j=1}^{n^*} (p/n^*)\hat{m}(z_j)} \hat{m}(z_i),$$

$z_i = \{1, 2, \ldots, n^*\}$. The effect of this process is distinct in sense that it resembles a trend of $d_1, \ldots, d_{n^*}$ which is close to the one of true eigenvalues, instead of simply shrinking the eigenvalues toward the grand mean.

4. With adjusted eigenvalues $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge \hat{\lambda}_p$, we obtain an adjusted covariance matrix $\widehat{\boldsymbol{\Sigma}}_A = \mathbf{U}\mathbf{D}_A\mathbf{U}^T$, where $\mathbf{D}_A$ is a diagonal matrix with elements of $\hat{\lambda}_i$'s.

In summary, we adjust the estimated eigenvalues from the sample covariance matrix so that the structure of the adjusted eigenvalues is similar to the one of the sample covariance matrix for $n \le n^*$, while preserving $\text{tr}(\widehat{\boldsymbol{\Sigma}})$ to be unchanged.

Remark 1. The proposed algorithm is based on kernel smoothing. In some cases, for example one with a dominated eigenvalue as Figure 1, it may be unlikely to identify the spiky structure because kernel smoothing makes a seemingly smooth line connecting the dominated eigenvalue with following eigenvalues.

Remark 2. The above algorithm is based on some covariance estimation methods that hold sample eigenvectors and replace the sample eigenvalues, e.g., Lin and Perlman (1985); Stein (1975 1986). Compared to true eigenvalues, the sample eigenvalues may have a tendency of being more spread out (Dey & Srinivasan 1985). This fact provides a rationale that one should adjust the sample eigenvalues. Using this fact, we come up with a simple and convincing idea: the sample eigenvalues may be spread out further in smaller sample cases, but its decreasing pattern may be maintained, i.e., lack of sample could affect the amount of being spread out in the sample eigenvalues. Therefore, based on the consistency of $\text{tr}(\widehat{\boldsymbol{\Sigma}})$, to restore the true eigenvalue structure is no more than to adjust the sample eigenvalues while holding $\text{tr}(\widehat{\boldsymbol{\Sigma}})$ unchanged. In other words, maintaining $\text{tr}(\widehat{\boldsymbol{\Sigma}})$ plays a role in holding patterns with true eigenvalue structure, and kernel smoothing adjusts eigenvalues within the constant $\text{tr}(\widehat{\boldsymbol{\Sigma}})$. This idea is backed up by numerical studies in Sections 3 and 4.

2.3 | Regularizing $\widehat{\boldsymbol{\Sigma}}_A$

As a by-product of adjusting the estimated eigenvalues in the preceding subsection, the singularity issue has been resolved as we construct a newly estimated eigenvalues which are strictly greater than zero. However, the estimation of eigenvectors might still remain unsatisfactory in high-dimensional settings. With $p > n$ setting, it is challenging to improve the estimation of eigenvectos without any prior knowledge about the true covariance structure. Hence instead of sticking to improvement of eigenvectors, we proceed with a regularization so that an empirically optimal regularization parameter is selected in terms of minimizing misclassification rate. A common approach to regularization for covariance matrix is that

$$\widehat{\boldsymbol{\Sigma}}^* = a\widehat{\boldsymbol{\Sigma}} + (1 - a)\mathbf{T}, \tag{6}$$

where $\mathsf{a} \in [0, 1]$ and $\mathbf{T}$ is predetermined well-conditioned target matrix, see Friedman (1989), Ledoit and Wolf (2004), Fisher and Sun (2011), etc. For example, we set $\mathbf{T}$ as a diagonal matrix with diagonal terms of $\widehat{\boldsymbol{\Sigma}}$ or an identity matrix $\mathbf{I}_\mathsf{p}$.

In place of (6), we consider a different type of regularization technique in the sense that (i) we do regularize not $\widehat{\boldsymbol{\Sigma}}$ but $\widehat{\boldsymbol{\Sigma}}_\mathsf{A}$ whose eigenvalues are adjusted, $\hat{\lambda}_\mathsf{i}$, $\mathsf{i} = 1, \ldots, \mathsf{p}$ and (ii) we will not do any harm on the trace of $\widehat{\boldsymbol{\Sigma}}_\mathsf{A}$, i.e., preserving the summation of estimated eigenvalues, which leads to regularizing off-diagonal elements of $\widehat{\boldsymbol{\Sigma}}_\mathsf{A}$. Define a hollow matrix, $\widehat{\boldsymbol{\Sigma}}^\mathsf{H}_\mathsf{A}$, which has zero diagonal elements with same off-diagonal elements as $\widehat{\boldsymbol{\Sigma}}_\mathsf{A}$, and $\widehat{\boldsymbol{\Sigma}}^\mathsf{D}_\mathsf{A}$ to be a diagonal matrix of $\widehat{\boldsymbol{\Sigma}}_\mathsf{A}$. For $\mathsf{a} \in [0, 1]$, we come up with a regularized covariance matrix $\widetilde{\boldsymbol{\Sigma}}_\mathsf{A}$

$$\widetilde{\boldsymbol{\Sigma}}_\mathsf{A} = \mathsf{a}\widehat{\boldsymbol{\Sigma}}^\mathsf{H}_\mathsf{A} + \widehat{\boldsymbol{\Sigma}}^\mathsf{D}_\mathsf{A}, \tag{7}$$

in which $\widetilde{\boldsymbol{\Sigma}}_\mathsf{A} = \widehat{\boldsymbol{\Sigma}}_\mathsf{A}$ when $\mathsf{a} = 1$. We are able to choose a tuning parameter a empirically by minimizing cross-validation error rate.

2.4 | A New Classification Rule

Combining all of components we described in the aforementioned subsections, we propose a new classification rule as follows:

$$\hat{\delta}_\mathsf{new}(\mathbf{X}) = \mathbb{I}\{\widetilde{\mathbf{w}}^\mathrm{T}(\mathbf{X} - \widetilde{\boldsymbol{\mu}}_\mathsf{a}) > 0\}, \tag{8}$$

where $\widetilde{\boldsymbol{\mu}}_\mathsf{a} = (\widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_2)/2$. We obtain $\widetilde{\mathbf{w}}$ by solving the optimization problem,

$$\widetilde{\mathbf{w}} \ = \ \underset{\mathbf{w} \neq 0}{\mathsf{argmax}} \frac{\mathbf{w}^\mathrm{T}\widehat{\boldsymbol{\mu}}_\mathsf{d}}{(\mathbf{w}^\mathrm{T}\widetilde{\boldsymbol{\Sigma}}_\mathsf{A}\mathbf{w})^{1/2}}, \tag{9}$$

where $\widehat{\boldsymbol{\mu}}_\mathsf{d} = \widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2$. In terms of implementation of (9), which is a standard optimization problem, one may employ one of computer packages to solve the problem. For example, one may use an R package, **CVXR** (Fu, Narasimhan, & Boyd 2017) or its Matlab version, **CVX** (Grant & Boyd 2014), etc. In fact, since singularity issue is resolved, i.e., $\widetilde{\boldsymbol{\Sigma}}_\mathsf{A}$ is invertible, we have a closed form of solution to the optimization problem of (9):

$$\widetilde{\mathbf{w}} \ = \ \widetilde{\boldsymbol{\Sigma}}_\mathsf{A}^{-1}\widehat{\boldsymbol{\mu}}_\mathsf{d}.$$

In view of the original optimization problem in (2), we maintain the rationale behind it, i.e., (9) is to find out a direction vector while minimizing the misclassification error with replaced mean difference vector and sample covariance matrix, $\widehat{\boldsymbol{\mu}}_\mathsf{d}$ and $\widetilde{\boldsymbol{\Sigma}}_\mathsf{A}^{-1}$. Only difference is that plug-in estimates are replaced with a shrinkage mean vector estimate $\widehat{\boldsymbol{\mu}}_\mathsf{d}$ and an adjusted and regularized sample covariance matrix $\widetilde{\boldsymbol{\Sigma}}_\mathsf{A}$. In general situation where one may use the sample mean difference vector and the sample covariance matrix in high dimensions, one should encounter two essential obstacles: the one is is the noise accumulation from estimating mean vector and the other is singularity of the sample covariance matrix (Bickel & Levina 2004; Fan & Fan 2008; Fan, Feng, & Tong 2012). Since these two issues are resolved, we expect that in general situation the direction vector selected from (9) should be better than others in existing methods, and this is supported from our numerical studies in Sections 3 and 4.

3 | SIMULATIONS

We carry out a broad range of simulations to see the performances of our proposed classification rule based on finite samples. Throughout this section, we assume that $\mathsf{p} = 500, 1000$, $\mathsf{n}_1 = \mathsf{n}_2 = 50, 100, 200$, and $\boldsymbol{\mu}_1 = \mathbf{0}$. We do each simulation 1000 times, and evaluate performances by comparing mis-classification rates which are computed based on $\mathsf{n}_1 + \mathsf{n}_2$ testing samples for each group.

- Case 1: The covariance matrix $\boldsymbol{\Sigma}$ is an identity matrix $\mathbf{I}_\mathsf{p}$. We set $\boldsymbol{\mu}_2 = (-\mathbf{1}^\mathrm{T}_{10}, \mathbf{1}^\mathrm{T}_{10}, \mathbf{0}^\mathrm{T}_{\mathsf{p}-20})^\mathrm{T}$.

- Case 2: We set the (i, j)-th element of $\mathbf{V}$ to be $\sqrt{2/(\mathsf{p}+1)}\sin\{\mathsf{ij}\pi/(\mathsf{p}+1)\}$, the elements of $\boldsymbol{\Lambda}$ to be $\lambda_\mathsf{j} = 2^{-3\mathsf{j}/(\mathsf{p}-1)+3/(\mathsf{p}-1)+3}$, where $\mathsf{j} = 1, \ldots, \mathsf{p}$, and $(\boldsymbol{\mu}_2)_\mathsf{j} = 5(-1)^\mathsf{j}(\mathbf{v}_2)_\mathsf{j}$, where $\mathsf{j} = 1, \ldots, \mathsf{p}$ and $\mathbf{v}_2$ is the eigenvector of $\boldsymbol{\Sigma}$ corresponding to the second largest eigenvalue of it.

- Case 3: We denote the mean vector to be $\boldsymbol{\mu}_2 = (-\mathbf{1}^\mathrm{T}_{10}, \mathbf{1}^\mathrm{T}_{10}, \mathbf{0}^\mathrm{T}_{\mathsf{p}-20})^\mathrm{T}$. We randomly generate every component of a $\mathsf{p} \times \mathsf{p}$ matrix $\mathbf{A}$ from $\mathsf{U}(-0.12, 0.12)$, and take $\boldsymbol{\Sigma} = \mathbf{A}^\mathrm{T}\mathbf{A}$.

- Case 4: The covariance matrix is a Toeplitz matrix with the (i, j)-th element of $1/(|\mathsf{i}-\mathsf{j}|+1)$, and $(\boldsymbol{\mu}_2)_\mathsf{j} = 3(-1)^\mathsf{j}(\mathbf{v}_1)_\mathsf{j}$, where $\mathsf{j} = 1, \ldots, \mathsf{p}$ and $\mathbf{v}_1$ is the eigenvector of $\boldsymbol{\Sigma}$ corresponding to the largest eigenvalue of it.

- Case 5: The diagonal terms of the covariance matrix $\boldsymbol{\Sigma}$ is 1 and its off-diagonal elements are $\rho = 0.25, 0.5, 0.75$. We set $\boldsymbol{\mu}_2 = 0.5\sqrt{\mathsf{p}}\mathbf{v}_1$, where $\mathbf{v}_1$ is the eigenvector of $\boldsymbol{\Sigma}$ corresponding the largest eigenvalue from the eigendecomposition of the covariance matrix $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\mathrm{T}$.

Case 1 shows an independent structure among covariates. Cases 2-4 include a various type of covariance structures: Cases 2 and 3 consider gradually decaying eigenvalues in which our adjusted eigenvalues are much more similar to true values, and Case 4 has exponentially decayed eigenvalues in which the adjusted eigenvalues also bring a better approximation to true eigenvalues. In Figure 2, we can verify that the structures of eigenvalues are well adjusted via our proposed method, in particular, in Cases 2 and 3. The covariance matrix in Case 5 represents a structure where a single eigenvalue dominates others. As seen in Figure 3, we generate plots for each $\rho = 0.25, 0.5, 0.75$, where the empty red triangles are true eigenvalues, the empty black circles are estimated eigenvalues from the sample covariance matrix, and the blue diamonds are from $\widetilde{\boldsymbol{\Sigma}}_{\mathsf{A}}$. Overall the pattern of adjusted eigenvalues are closer to the true one, however, in Case 5 where there is a single dominated eigenvalue, as seen Figure 3, there exist a noticeable difference for some small eigenvalues from the sample covariance matrix. This is due to intrinsic property of kernel smoothing, which is described in Remark 1 in Section 2.2.

We make comparisons between our proposed method and other procedures such as a simple classification rule, Naïve Bayes, and two data piling regularization methods proposed in Lee et al. (2013), which are a ridge linear discriminant analysis (rLDA) and the regularized data piling (RDP). The settings for Case 2 and Case 5 come from Lee et al. (2013), but they are slightly modified.

As pointed out by a referee, it would be helpful to assess the effects of the regularized mean difference vector and the sample covariance matrix with adjusted eigenvalues separately. In doing so, we conducted additional simulations in which we considered two separate cases: the one is to replace $\widetilde{\boldsymbol{\Sigma}}_{\mathsf{A}}$ with a ridge type covariance matrix and the other is to use the sample mean difference vector instead of the regularized one $\widehat{\boldsymbol{\mu}}_{\mathsf{d}}$. Details are relegated to the Supporting Information. As noted in Tables S.1 and S.2 in the Supporting Information, in almost all settings, the performances based on our proposed method are better than those two classifiers.

The results are in Table 1 for $\mathsf{p} = 500$ and Table 2 for $\mathsf{p} = 1000$, with varying sample sizes. Throughout Cases 1-4, our proposed method outperforms all of other rules. In Case 5, our proposed classification rule's performance is almost good as rLDA and RDP. As a matter of fact, compared to Case 5, Cases 1-4 would be more feasible settings in practice. In this reason, simulation results support a superiority of our proposed method in terms of misclassification rate.

4 | APPLICATIONS

In order to show the performance of our proposed classification rule in practice, we use three data sets which are in Dettling (2004), and they were re-analyzed in Lee et al. (2013). "Colon" data set has $\mathsf{n}_0 = 22$ and $\mathsf{n}_1 = 40$ observations with $\mathsf{p} = 2000$, "Leukemia" $\mathsf{n}_0 = 47$ and $\mathsf{n}_1 = 25$ observations with $\mathsf{p} = 3571$, and "Prostate" $\mathsf{n}_0 = 50$ and $\mathsf{n}_1 = 52$ observations with $\mathsf{p} = 6033$. Following Dettling (2004), we analyzed these data sets with 3-fold cross-validation, i.e., we randomly selected 2/3 observations from each data set as training sample, and then 1/3 were used for testing sample. In Table 3, the misclassification rates and its standard errors are provided based on 50 iterations. In Colon and Leukemia data sets, our proposed classification rule is slightly better than all other discriminant rules in terms of misclassificaiton rate. However, for Prostate data set, our method yields a little worse result than other classification rules. It may be because very few of eigenvalues of the true covariance matrix dominates others.

5 | DISCUSSION

We have proposed a new binary classification rule in high-dimensional data by replacing components which are used in a conventional optimization problem for classification with more reliable plug-in estimates. Above all, it has advantages over some existing high-dimensional classification rules in sense that (i) the sample covariance matrix is more reliable as we adjusts the estimated eigenvalues so that the trajectory of sample covariance eigenvalues is close to one of the true covariance matrix, (ii) this adjustment of estimated eigenvalues enables us to avoid the singularity issue easily, and (iii) the mean vectors are also regularized based on the shrinkage estimator instead of sample mean vector. In a problem where one and/or very few eigenvalues dominates others, the simulation studies show that our proposed method shows a similar or little worse than other methods. One main reason is that the proposed method of regularizing sample eigenvalues actually tends to smooth them, so our classification rule seems to fail in recognizing the spiky structure of eigenvalues of covariance matrix. Since the structure of covariance matrix such as spiky or not is unknown in practice, we may need to develop more adaptive procedures which can take into account those unknown structures of covariance matrix. We believe that the development of a better estimator of covariance matrix will improve our proposed classification rule for more various situations. In addition, for more rigorous treatment for the algorithm proposed in Section 2.2, our paramount goal to achieve is to show that $\hat{\lambda}_{\mathsf{i}} \to \lambda_{\mathsf{i}}$ in probability, but it would be very challenging and at the same time it deserve careful investigation in future research. We leave those issues as future work.

ACKNOWLEDGEMENT

We thank the Associate Editor and two anonymous reviewers for helpful comments and suggestions. This work was supported by the New Faculty Startup Fund from Seoul National University and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.2020R1A2C1A01100526).

References

Ackermann, M., & Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC bioinformatics*, *10*(1), 47.

Ahn, J., & Jeon, Y. (2015). Sparse hdlss discrimination with constrained data piling. *Computational Statistics & Data Analysis*, *90*, 74–83.

Ahn, J., & Marron, J. (2010). The maximal data piling direction for discrimination. *Biometrika*, *97*(1), 254–259.

Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y.-H., & Marron, J. (2018). A survey of high dimension low sample size asymptotics. *Australian & New Zealand journal of statistics*, *60*(1), 4–19.

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, *101*(473), 119–137.

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

Chételat, D., Wells, M. T., et al. (2012). Improved multivariate normal mean estimation with unknown covariance when p is greater than n. *The Annals of Statistics*, *40*(6), 3137–3160.

Dettling, M. (2004). Bagboosting for tumor classification with gene expression data. *Bioinformatics*, *20*(18), 3583–3593.

Dey, D. K., & Srinivasan, C. (1985). Estimation of a covariance matrix under stein's loss. *The Annals of Statistics*, *13*(4), 1581–1591.

Di Pillo, P. J. (1976). The application of bias to discriminant analysis. *Communications in Statistics-Theory and Methods*, *5*(9), 843–854.

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors—an empirical bayes approach. *Journal of the American Statistical Association*, *68*(341), 117–130.

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

Fan, J., Feng, Y., & 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*(4), 745–771.

Fan, J., & Li, R. (2006). Statistical challenges with high dimensionality. In *Proceedings of the international congress of mathematicians. madrid, spain.*

Fisher, T. J., & Sun, X. (2011). Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Computational Statistics & Data Analysis*, *55*(5), 1909–1918.

Fourdrinier, D., Strawderman, W. E., & Wells, M. T. (2003). Robust shrinkage estimation for elliptically symmetric distributions with unknown covariance matrix. *Journal of multivariate analysis*, *85*(1), 24–39.

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

Fu, A., Narasimhan, B., & Boyd, S. (2017). CVXR: An R package for disciplined convex optimization. *arXiv preprint arXiv:1711.07582*.

Gleser, L. J. (1986). Minimax estimators of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *The Annals of Statistics*, 1625–1633.

Grant, M. C., & Boyd, S. P. (2014). *CVX: MATLAB Software for Disciplined Convex Programming, version 2.1.* `http://cvxr.com/cvx`.

Greenshtein, E., & Park, J. (2009). Application of non parametric empirical Bayes estimation to high dimensional classification. *Journal of Machine Learning Research*, *10*(Jul), 1687–1704.

Greenshtein, E., Park, J., & Lebanon, G. (2009). Regularization through variable selection and conditional mle with application to classification in high dimensions. *Journal of statistical planning and inference*, *139*(2), 385–395.

Guo, Y., Hastie, T., & Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, *8*(1), 86–100.

Hastie, T., & Tibshirani, R. (2004). Efficient quadratic regularization for expression arrays. *Biostatistics*, *5*(3), 329–340.

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

James, W., & Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 361–379.

Jonsson, D. (1982). Some limit theorems for the eigenvalues of a sample covariance matrix. *J. Multivariate Anal.*, *12*, 1–38.

Ledoit, O., & Wolf, M. (2004). A well conditioned estimator for large dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411.

Lee, M. H., Ahn, J., & Jeon, Y. (2013). HDLSS discrimination with adaptive data piling. *Journal of Computational and Graphical Statistics*, *22*(2), 433–451.

Lin, S., & Perlman, M. (1985). *A monte carlo comparison of four estimators for a covariance matrix, multivariate analysis vi (ed. pr krishnaiah), 411-429.* North Holland, Amsterdam.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., … others (2003). Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, *34*(3), 267–273.

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

Stein, C. (1975). Estimation of a covariance matrix, rietz lecture. In *39th annual meeting ims, atlanta, ga, 1975.*

Stein, C. (1986). Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics*, *34*(1), 1373–1403.

Tong, T., Chen, L., & Zhao, H. (2012). Improved mean estimation and its application to diagonal discriminant analysis. *Bioinformatics*, *28*(4), 531–537.

TABLE 1 Simulation results with $p = 500$ and varying sample sizes: misclassification rates and their sample standard errors within braces based on 1000 iterations. “New” stands for our proposed classification rule, “NB” for Naïve Bayes, “rLDA” for the ridge linear discriminant analysis (rLDA), and “RDP” for the regularized data piling (RDP).

		New	NB	rLDA	RDP
$p = 500, n_1 = n_2 = 50$					
Case 1	-	0.0384(0.0204)	0.0590(0.0246)	0.0618(0.0266)	0.0615(0.0254)
Case 2	-	0.1750(0.0405)	0.2461(0.0446)	0.2265(0.0452)	0.2256(0.0447)
Case 3	-	0.2123(0.0448)	0.2797(0.0491)	0.2643(0.0469)	0.2624(0.0465)
Case 4	-	0.1580(0.0410)	0.2587(0.0496)	0.1952(0.0475)	0.1991(0.0479)
Case 5	$\rho = 0.25$	0.3198(0.0458)	0.3107(0.0462)	0.3171(0.0482)	0.3153(0.0470)
	$\rho = 0.50$	0.3716(0.0485)	0.3628(0.0485)	0.3720(0.0516)	0.3689(0.0516)
	$\rho = 0.75$	0.3967(0.0522)	0.3882(0.0493)	0.3985(0.0531)	0.3944(0.0523)
$p = 500, n_1 = n_2 = 100$					
Case 1	-	0.0252(0.0116)	0.0348(0.0134)	0.0354(0.0136)	0.0354(0.0136)
Case 2	-	0.0949(0.0235)	0.1751(0.0306)	0.1377(0.0289)	0.1406(0.0289)
Case 3	-	0.1531(0.0275)	0.2258(0.0309)	0.1885(0.0309)	0.1883(0.0300)
Case 4	-	0.0759(0.0210)	0.1889(0.0355)	0.1063(0.0259)	0.1164(0.0264)
Case 5	$\rho = 0.25$	0.3390(0.0336)	0.3095(0.0318)	0.3125(0.0326)	0.3118(0.0323)
	$\rho = 0.50$	0.3850(0.0376)	0.3624(0.0345)	0.3681(0.0363)	0.3642(0.0361)
	$\rho = 0.75$	0.4054(0.0376)	0.3871(0.0344)	0.3913(0.0355)	0.3883(0.0354)
$p = 500, n_1 = n_2 = 200$					
Case 1	-	0.0216(0.0081)	0.0231(0.0075)	0.0235(0.0078)	0.0235(0.0077)
Case 2	-	0.0403(0.0102)	0.1074(0.0181)	0.0638(0.0140)	0.0905(0.0171)
Case 3	-	0.0708(0.0147)	0.1745(0.0202)	0.0877(0.0172)	0.0959(0.0185)
Case 4	-	0.0337(0.0097)	0.1203(0.0115)	0.0488(0.0115)	0.0960(0.0200)
Case 5	$\rho = 0.25$	0.3546(0.0257)	0.3090(0.0231)	0.3108(0.0238)	0.3100(0.0230)
	$\rho = 0.50$	0.3914(0.0319)	0.3621(0.0235)	0.3649(0.0242)	0.3623(0.0234)
	$\rho = 0.75$	0.4109(0.0341)	0.3866(0.0246)	0.3898(0.0260)	0.3868(0.0252)

TABLE 2 Simulation results with $p = 1000$ and varying sample sizes: misclassification rates and their sample standard errors within braces based on 1000 iterations.

		New	NB	rLDA	RDP
$p = 1000, n_1 = n_2 = 50$					
Case 1	-	0.0638(0.0243)	0.1012(0.0329)	0.1014(0.0327)	0.1011(0.0324)
Case 2	-	0.2529(0.0452)	0.3141(0.0462)	0.3062(0.0457)	0.3059(0.0454)
Case 3	-	0.3818(0.0531)	0.4138(0.0479)	0.4080(0.0485)	0.4074(0.0484)
Case 4	-	0.2486(0.0461)	0.3241(0.0489)	0.2935(0.0481)	0.2942(0.0477)
Case 5	$\rho = 0.25$	0.3132(0.0477)	0.3086(0.0461)	0.3153(0.0467)	0.3128(0.0464)
	$\rho = 0.50$	0.3681(0.0482)	0.3642(0.0490)	0.3709(0.0505)	0.3694(0.0506)
	$\rho = 0.75$	0.3931(0.0480)	0.3895(0.0486)	0.3973(0.0509)	0.3948(0.0512)
$p = 1000, n_1 = n_2 = 100$					
Case 1	-	0.0379(0.0144)	0.0584(0.0167)	0.0592(0.0173)	0.0593(0.0174)
Case 2	-	0.1755(0.0277)	0.2455(0.0308)	0.2247(0.0313)	0.2246(0.0315)
Case 3	-	0.3270(0.0342)	0.3795(0.0361)	0.3661(0.0350)	0.3657(0.0345)
Case 4	-	0.1534(0.0286)	0.2615(0.0358)	0.1935(0.0313)	0.1954(0.0316)
Case 5	$\rho = 0.25$	0.3129(0.0326)	0.3081(0.0327)	0.3125(0.0340)	0.3106(0.0339)
	$\rho = 0.50$	0.3683(0.0337)	0.3636(0.0338)	0.3676(0.0355)	0.3654(0.0350)
	$\rho = 0.75$	0.3906(0.0349)	0.3864(0.0353)	0.3928(0.0384)	0.3894(0.0383)
$p = 1000, n_1 = n_2 = 200$					
Case 1	-	0.0264(0.0083)	0.0345(0.0090)	0.0346(0.0090)	0.0349(0.0093)
Case 2	-	0.0957(0.0186)	0.1736(0.0207)	0.1354(0.0200)	0.1396(0.0205)
Case 3	-	0.2720(0.0262)	0.3389(0.0250)	0.2973(0.0260)	0.2982(0.0262)
Case 4	-	0.0734(0.0147)	0.1900(0.0271)	0.1040(0.0177)	0.1148(0.0180)
Case 5	$\rho = 0.25$	0.3218(0.0228)	0.3089(0.0230)	0.3110(0.0232)	0.3098(0.0232)
	$\rho = 0.50$	0.3734(0.0251)	0.3606(0.0236)	0.3635(0.0245)	0.3609(0.0237)
	$\rho = 0.75$	0.3944(0.0259)	0.3850(0.0249)	0.3877(0.0258)	0.3855(0.0255)

TABLE 3 Misclassification rates and its sample standard errors within braces for three gene expression data sets.

	New	NB	rLDA	RDP
Colon	0.1248(0.0600)	0.1760(0.0679)	0.1360(0.0572)	0.1440(0.0611)
Leukemia	0.0192(0.0210)	0.0208(0.0242)	0.0300(0.0280)	0.0225(0.0255)
Prostate	0.1029(0.0532)	0.3612(0.1046)	0.0859(0.0398)	0.0906(0.0424)

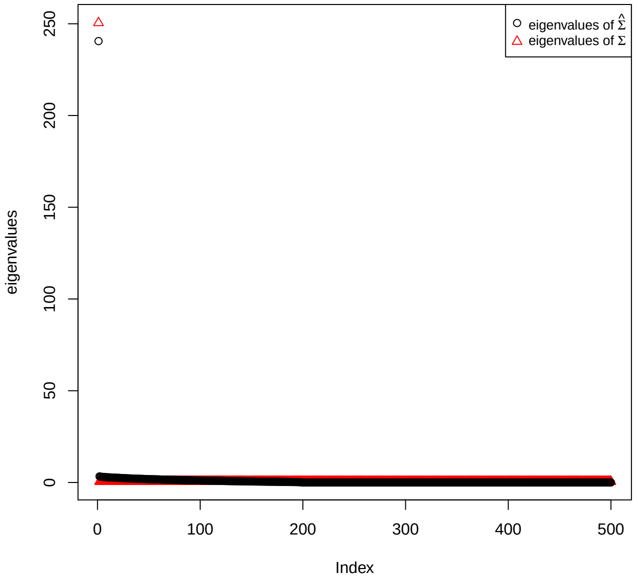

FIGURE 1 A example that the largest sample eigenvalue is greater than the true one. The x-axis represents indices for ordered eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$, i.e., index=i for $i = 1, \ldots, 500$ where $\max_i d_i = d_1 \geq \cdots \geq d_{500} = \min_i d_i$ and $\max_i \lambda_i = \lambda_1 \geq \cdots \geq \lambda_{500} = \min_i \lambda_i$.

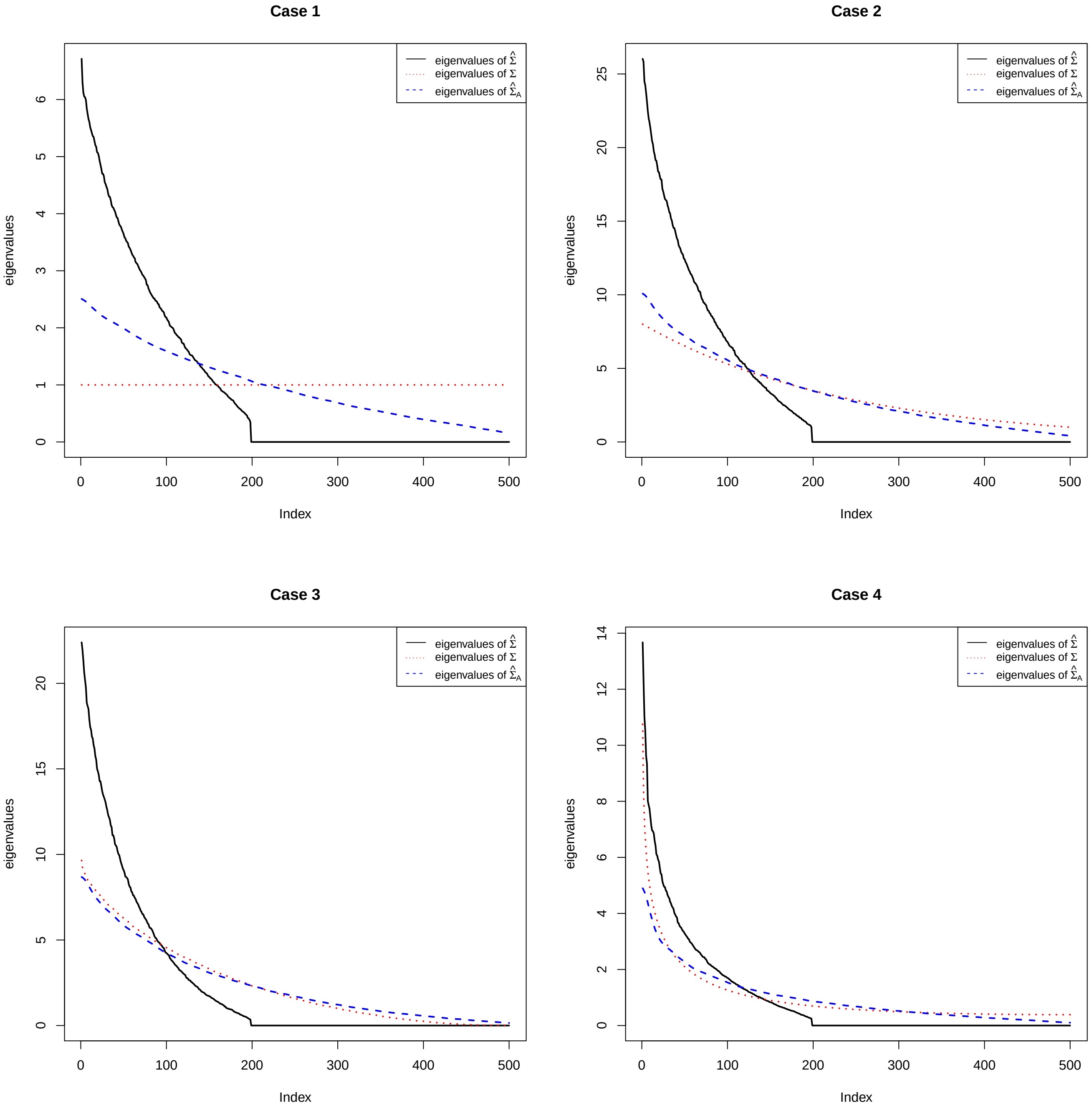

FIGURE 2 Plots of eigenvalues of the true covariance matrix, the sample covariance matrix, and the adjusted convariance matrix for each of Cases 1, 2, 3, and 4 when $p = 500$ and $n_1 = n_2 = 100$.

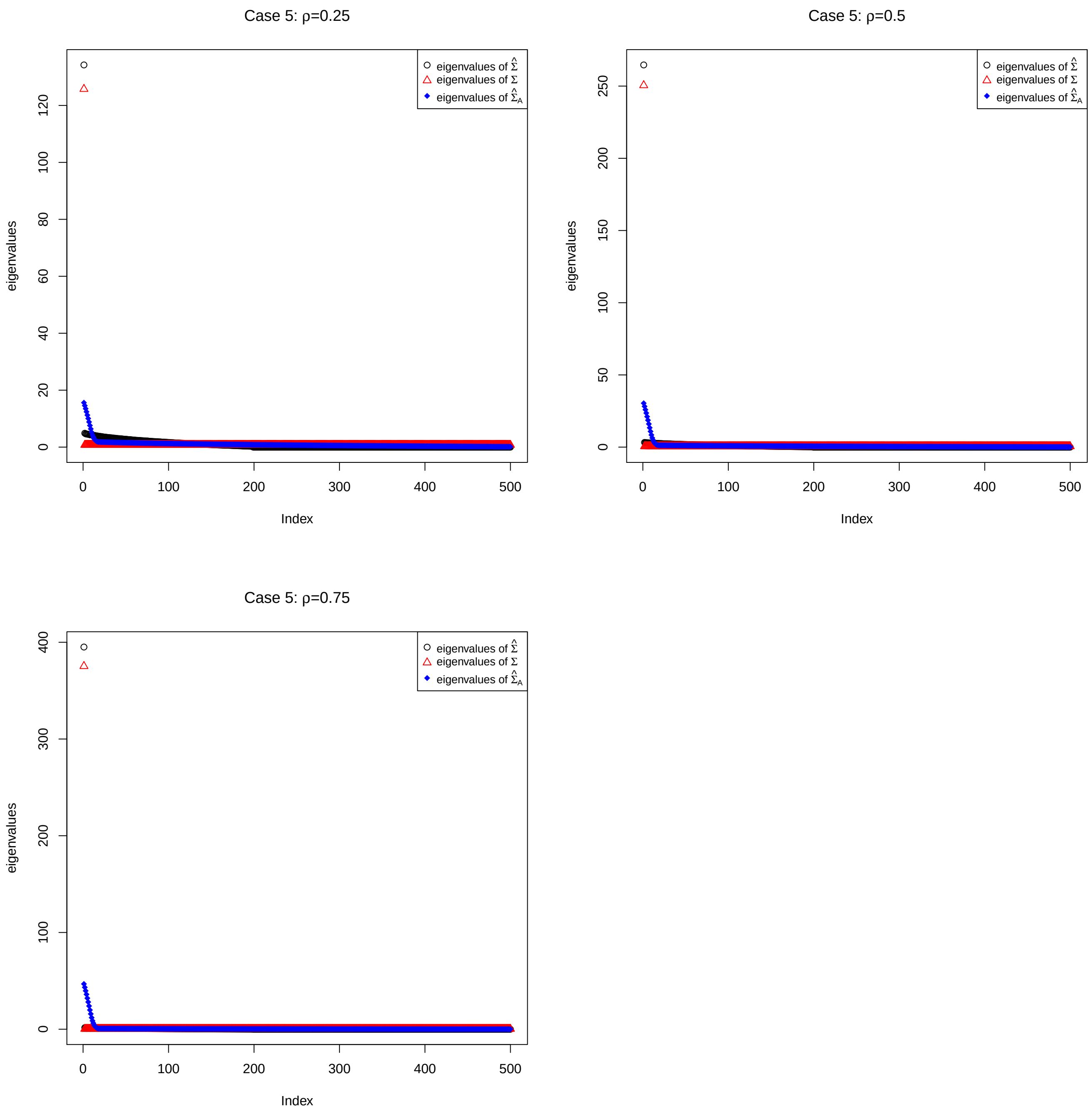

FIGURE 3 Plots of eigenvalues of the true covariance matrix, the sample covariance matrix, and the adjusted convariance matrix for Case 5 when $p = 500$ and $n_1 = n_2 = 100$.

How to cite this article: S. Baek, H. Park, J. Park (2020+), A High-Dimensional Classification Rule Using Sample Covariance Matrix Equipped With Adjusted Estimated Eigenvalues, *STAT*, *XXXX*.