Biometrics, 2026, 82(1), ujag039
https://doi.org/10.1093/biomtc/ujag039
Published: 26 February 2026
Biometric Methodology

OXFORD

Sparse robust discriminant analysis for high-dimensional and heavy-tailed data

Weijian Huang[1], Qing Mai[2], Jing Zeng[1,*]

[1]Faculty of Business for Science & Technology, School of Management, University of Science and Technology of China, Hefei, Anhui 230026, China
[2]Department of Statistics, Florida State University, Tallahassee, FL 32312, United States

*Corresponding author: Jing Zeng, Faculty of Business for Science & Technology, School of Management, University of Science and Technology of China, Hefei, Anhui 230026, China (zengjxl@ustc.edu.cn).

Abstract

With advancements in data-collecting techniques, large-scale data have become increasingly prevalent in medical science. For instance, gene expression data provide information on tens of thousands of genes, while diagnostic imaging, such as the magnetic resonance imaging, generates a vast volume of pixels. While various sparse linear discriminant analysis methods have been developed to handle high-dimensional medical data, they often assume the light-tailed predictors, which is frequently violated in real applications. In this paper, we propose a robust classifier under an elliptically contoured discriminant analysis (EDA) model, which accommodates both light-tailed and heavy-tailed data. In addition, we assess the prediction accuracy using the balanced rate, a more appropriate metric when the data is imbalanced. Under the EDA model, we identify the intrinsic dimension-reduction subspace that captures all information from predictors for achieving the lowest balanced rate. By leveraging this dimension-reduction subspace, we propose a robust high-dimensional classifier, which reduces data dimensionality through subspace projection, followed by prediction on the reduced data. Theoretically, our proposal simultaneously enjoys the consistencies of subspace estimation, variable selection, and prediction accuracy under only finite fourth-moment condition of predictors. Numerically, we apply our method to synthetic data and three real datasets, including two lung cancer data and a leukemia data. The empirical findings support the superiority of our approach over other state-of-the-art methods.

Keywords: discriminant analysis, heavy-tailedness, high-dimensional classification, imbalanced data, variable selection

1 Introduction

Classification is one of the core statistical problems, finding important applications in medical science. With advancements of frontier data-collecting tools, it is increasingly common to encounter ultra-high dimensional data nowadays. For instance, in genomics, a critical task is diagnosing diseases using gene expression data, which often involves tens of thousands of genes (Min et al., 2018). In neuroimaging studies, medical images, such as magnetic resonance imaging, are used for disease diagnosis and prognosis, and vectorizing high-resolution medical images results in ultra-high dimensional data (Zhang and Li, 2017). In multi-omics data analysis, heterogeneous data is collected for the same subject from diverse resources, including genome, proteome, transcriptome, and metabolome, contributing to the high-dimensional nature of multi-omics data (Kaur et al., 2021). These overwhelming, high-dimensional medical data pose great challenges to classical classification tools.

Ever since the introduction of Fisher's linear discriminant analysis (LDA; Fisher, 1936), LDA has become one of the most prominent probabilistic model-based classifiers. In recent decades, a variety of sparse LDA methods have been developed to tackle high-dimensional data, see Clemmensen et al. (2011), Fan and Fan (2008), Witten and Tibshirani (2011), Mai et al. (2019), Cai and Zhang (2019), Ren and Mai (2022), and Zeng et al. (2023), for instance. However, LDA methods rely on the conditional normality for predictors, which is often violated in real world. For demonstration purposes, we take as an example a lung cancer data. The dataset contains expression levels of 12 533 gene transcripts from 181 surgical specimens in two classes. In Fig-

Received: 29 March 2025. **Revised:** 26 November 2025. **Accepted:** 28 January 2026
© The Author(s) 2026. Published by Oxford University Press on behalf of The International Biometric Society. All rights reserved. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site-for further information please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/biometrics/article/82/1/ujag039/8498964 by guest on 27 February 2026

ure 1, we present the histogram of excess kurtosis coefficients for 2000 selected variables. The histogram shows that approximately 72% of variables have excess kurtosis coefficients greater than 2, indicating the presence of heavy-tailedness.

Another critical challenge results from the fact that LDA minimizes the overall misclassification rate. On imbalanced data where some classes have much fewer observations, an overall misclassification rate based classifier tends to emphasize the predictions of the majority class. An extreme example would be a naive classifier that blindly classifies all observations to the majority class. If the majority class is sufficiently dominating, the naive classifier achieves a low overall misclassification rate but will be useless in detecting observations from the minority class. In many medical problems, accurately classifying samples from the minority class is far more important than classifying those from the majority class. Thus, it is essential to develop classifiers maintaining high accuracy on imbalanced data.

In this paper, we propose a high-dimensional classifier called SODA, short for **S**parse r**O**bust **D**iscriminant **A**nalysis. SODA is highly accurate even when data deviate from light-tailed distributions. Our proposal is developed under the elliptically contoured discriminant analysis (EDA) model, which assumes the predictors follow an elliptically contoured (EC) distribution within each class, widening the scope of discriminant analysis model. Moreover, we assess classification accuracy using the balanced rate, putting more emphasis on the classification of the minority class. Thus, our proposal achieves higher accuracy on minority classes. Under the EDA model, we identify the underlying dimension-reduction subspace containing all the information relevant to prediction accuracy measured by the balanced rate. This allows us to reduce the dimensionality of the predictors without sacrificing classification information, and facilitate the visualization of data distribution patterns. Then, we propose a robust estimation of the dimension-reduction subspace by truncating the predictors and reformulating the subspace estimation as a group-Lasso optimization problem. After projecting the data onto the dimension-reduction subspace, the subsequent inference is conducted on the reduced data. The robustness of our proposal is also supported from theoretical respective: it enjoys consistency in subspace estimation, variable selection, and prediction under only the finite fourth-order moment assumption of the predictors, which relaxes the light-tailed distribution assumption in sparse LDA methods.

Recently, a variety of works have also been devoted to robust classification. For example, Hall et al. (2009) proposed a component-wise median-based classifier, verified to behave well under high-dimensional heavy-tailed data. Hennig and Viroli (2016) developed a quantile-based classifier, which includes the median classifier as a special case. More recently, Xiong et al. (2025) introduce a component-wise mode-based classifier, leveraging structures revealed by mode that are missed by mean, median, and quantile. Ren and Mai (2022) also developed a robust distance-based method by replacing unstable sample estimates with more robust Huber-loss estimates. However, these methods lack a probabilistic interpretation and are unable to visualize data directly. In comparison, our discriminant analysis model offers a more refined characterization of data patterns within each class. Recovering the underlying low-dimensional representation, the high-dimensional data can be visualized directly in a low-dimensional space, providing deeper insights into data distribution.

2 Background

2.1 Notations

For a vector $\mathbf{v} = (v_i) \in \mathbb{R}^p$, the L_q-norm is defined as $\|\mathbf{v}\|_q = \left(\sum_{i=1}^p v_i^q\right)^{1/q}$, where $1 \le q < \infty$, the L_0-norm is $\|\mathbf{v}\|_0 = |\{i \mid v_i \neq 0\}|$, and the L_∞-norm is $\|\mathbf{v}\|_\infty = \max_i |v_i|$. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$, we use $\mathrm{span}(\mathbf{A})$ to denote the subspace spanned by the column vectors of $\mathbf{A}$, and define its $L_{2,1}$-norm, L_∞-norm, and Frobenius norm as $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^p (\sum_{j=1}^q a_{ij}^2)^{1/2}$, $\|\mathbf{A}\|_\infty = \max_i \sum_{j=1}^q |a_{ij}|$, and $\|\mathbf{A}\|_F = (\sum_{i=1}^p \sum_{j=1}^q a_{ij}^2)^{1/2}$, respectively. The nuclear norm and the spectral norm are defined as $\|\mathbf{A}\|_* = \sum_{i=1}^{\min\{p,q\}} \sigma_i(\mathbf{A})$ and $\|\mathbf{A}\| = \sigma_1(\mathbf{A})$, where $\sigma_1(\mathbf{A}) \ge \ldots \ge \sigma_{\min\{p,q\}}(\mathbf{A}) \ge 0$ are the ordered singular values of $\mathbf{A}$. For a positive semi-definite matrix $\mathbf{N} \in \mathbb{R}^{p\times p}$, let $\varphi_{\max}(\mathbf{N}) \equiv \varphi_1(\mathbf{N}) \ge \ldots \ge \varphi_p(\mathbf{N}) \equiv \varphi_{\min}(\mathbf{N})$ denote its eigenvalues. Suppose that $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ is a basis of a subspace $\mathcal{S} \subseteq \mathbb{R}^p$, then $\mathbf{P}_\mathcal{S} \equiv \mathbf{P}_\beta = \boldsymbol{\beta}\left(\boldsymbol{\beta}^\top\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}^\top$ denotes the projection matrix onto the subspace $\mathcal{S}$. For two scalar series $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we say $a_n \asymp b_n$ if there exist constants $0 < c_1 < c_2$ such that $c_1 a_n \le b_n \le c_2 a_n$ for all n, and $a_n \lesssim b_n$ if there exists some constant $M > 0$ such that $a_n \le M b_n$ for all n. For two values a and b, define $a \wedge b = \min\{a, b\}$.

2.2 EC distribution

Let $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ be deterministic, $\mathbf{U} \in \mathbb{R}^p$ a random vector uniformly distributed on the unit sphere, and $\xi > 0$ a scaler random variable independent of $\mathbf{U}$. Suppose that $\mathbf{X} = \boldsymbol{\mu} + \xi\boldsymbol{\Sigma}^{1/2}\mathbf{U}$, then $\mathbf{X} \in \mathbb{R}^p$ follows an EC distribution (Johnson, 1987). The density function of $\mathbf{X}$ is given by $f(\boldsymbol{x}) = k_p|\boldsymbol{\Sigma}|^{-1/2} g\{(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$, for $\boldsymbol{x} \in \mathbb{R}^p$, where $k_p > 0$ is a normalizing constant, and g is some positive link function. We denote $\mathbf{X} \sim \mathrm{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; g)$.

3 Probabilistic Model And Dimension-Reduction Subspace

For response $Y \in \{1, \ldots, K\}$ and predictors $\mathbf{X} \in \mathbb{R}^p$, we consider the *EDA analysis* model as follows:

$$\Pr(Y = k) = \pi_k, \quad \mathbf{X} \mid (Y = k) \sim \mathrm{EC}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}; g), \tag{1}$$

where we assume that $\mathrm{E}(\mathbf{X}) = \mathbf{0}$ and the link function $g(u)$ is monotonic and bounded for $u \ge 0$.

The EDA model (1) can be traced back to the work of Wakaki (1994), who investigated the properties of Fisher's LDA within this framework. Subsequent researches in discriminant analysis has utilized the EDA model or its special cases to study the classification problem involving heavy-tailed data, see, for example, Bose et al. (2015) and Ghosh

Downloaded from https://academic.oup.com/biometrics/article/82/1/ujag039/8498964 by guest on 27 February 2026

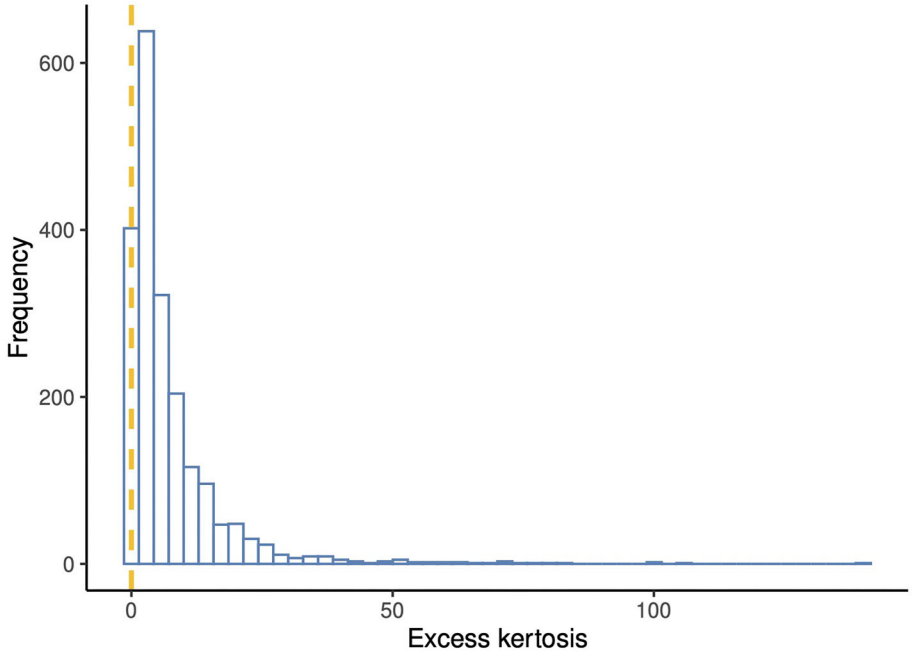

Figure 1 Histogram of the excess kurtosis coefficients of 2000 selected variables on the lung cancer data.

et al. (2021). The monotonicity assumption on g has been considered in literature, such as in works Wakaki (1994) and Shao et al. (2011). In fact, the EC family with monotonic and bounded $g(u)$ for $u \geq 0$ covers many common distributions, which we summarized in Table S.1 in the Supplementary Materials.

For a general classification rule $\delta : \mathbb{R}^p \to \{1, \ldots, K\}$, we propose to measure its prediction accuracy via the *balanced rate*, which is defined as

$$R(\delta) = \frac{1}{K} \sum_{k=1}^{K} \Pr\{\delta(\mathbf{X}) \neq k \mid Y = k\}. \tag{2}$$

The balanced rate has been widely adopted in literature, including in LDA methods (Cai and Liu, 2011; Shao et al., 2011) and quadratic discriminant analysis (QDA) methods (Ghosh et al., 2021; Li and Shao, 2015). They use the balanced rate mainly for simplifying theoretical analysis. Qiao and Liu (2009) used the balanced rate for addressing the issue of overall misclassification rate in the imbalanced problems.

Moreover, Tong et al. (2020) proposed a Neyman–Pearson (NP) classifier for binary classification, which was further generalized in Tian and Feng (2025) to multi-class model-free problems. The idea of the NP classifier is similar to the balanced rate, in that it protects the error rates of some specific classes.

Another commonly used criterion for quantifying the prediction accuracy is the *overall misclassification rate*, defined as $\widetilde{R}(\delta) = \sum_{k=1}^{K} \pi_k \Pr\{\delta(\mathbf{X}) \neq k \mid Y = k\}$, which is used by Bayes rule-based classifiers. However, when dealing with the imbalanced data, the overall misclassification rate is improper since the misclassification of a sample from a majority class will inflate it by the larger amount of prior π_k.

In contrast, the balanced rate takes into account a balance among the error rates of different classes, and is more favored when measuring the prediction accuracy on the imbalanced data.

We define the *optimal rule* as the rule achieving the lowest balanced rate, denoted by $\delta^*_{\mathbf{X}}$, where the subscript $\mathbf{X}$ implies the dependence of the rule on the conditional distribution of $\mathbf{X} \mid Y$. The optimal rule is used as the benchmark when evaluating the prediction accuracy of a rule. The following theorem gives the general form of the optimal rule and its specific form under the EDA model.

Theorem 1: The optimal rule $\delta^*_{\mathbf{X}}$ takes the form as $\delta^*_{\mathbf{X}}(\boldsymbol{x}) = \operatorname{argmax}_{k \in \{1,\ldots,K\}} f_k(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathbb{R}^p$ and f_k is the density of the conditional distribution $\mathbf{X} \mid (Y = k)$. Specifically, under the EDA model (1),

$$\delta^*_{\mathbf{X}}(\boldsymbol{x}) = \operatorname*{argmax}_{k \in \{1,\ldots,K\}} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} \{\boldsymbol{x} - (\boldsymbol{\mu}_k + \boldsymbol{\mu}_1)/2\}. \tag{3}$$

In (3), we choose to use the first class as the reference class for simplicity. However, any class can be used as the reference class to produce legitimate classifiers using our method.

As indicated by Theorem 1, under the EDA model, the optimal rule depends on $\boldsymbol{x}$ only through its $K-1$ linear combinations $(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}$, $k = 2, \ldots, K$.

Therefore, the $K-1$ directions $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$, $k = 2, \ldots, K$, contain all the information from $\mathbf{X}$ for prediction. However, this set of directions are not the only choice for dimension reduction. As demonstrated in the following lemma, any matrix spanning the same subspace as $\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1), k = 2, \ldots, K\}$ serves the same purpose in reducing the data. Let $\delta^*_{\boldsymbol{\beta}^\top \mathbf{X}}$ denote the optimal rule for $\boldsymbol{\beta}^\top \mathbf{X}$ such that $\delta^*_{\boldsymbol{\beta}^\top \mathbf{X}}(\boldsymbol{\beta}^\top \boldsymbol{x}) = \operatorname{argmax}_{k \in \{1,\ldots,K\}} h_k(\boldsymbol{\beta}^\top \boldsymbol{x})$, where h_k denotes the conditional distribution of $\boldsymbol{\beta}^\top \mathbf{X} \mid (Y = k)$.

Lemma 1: Under the EDA model (1), for any matrix $\boldsymbol{\beta} \in \mathbb{R}^{p \times (K-1)}$ such that

Downloaded from https://academic.oup.com/biometrics/article/82/1/ujag039/8498964 by guest on 27 February 2026

$\text{span}(\boldsymbol{\beta}) = \text{span}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1), \ldots, \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_K - \boldsymbol{\mu}_1)\}$, we have $\delta^*_{\mathbf{X}}(\boldsymbol{x}) = \delta^*_{\boldsymbol{\beta}^\top \mathbf{X}}(\boldsymbol{\beta}^\top \boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^p$.

We denote the dimension-reduction subspace in the EDA model by $\mathcal{S} = \text{span}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1), \ldots, \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_K - \boldsymbol{\mu}_1)\}$. According to Lemma 1, the subspace $\mathcal{S}$ fully characterizes the dimension-reduction structure inherent in the EDA model, containing all the information from $\mathbf{X}$ for achieving the optimal rate.

When p is less than n, the subspace $\mathcal{S}$ can be estimated by directly substituting the sample estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_k$, $k = 1, \ldots, K$.

However, in the ultra-high-dimensional setting, where p grows at an exponential rate with n, the sample estimate of $\boldsymbol{\Sigma}$ is singular, rendering the failure of the plug-in method. In the next section, we introduce the sparsity structure to $\mathcal{S}$, which enables its efficient estimation even under the ultra-high-dimensional setting.

4 Sparse Robust Discriminant Analysis

4.1 Main procedure

In high-dimensional statistics, it is a common strategy to assume the sparsity structure in modeling. In our setting, we assume that only a subset of predictors contributes to the classification. Let $\boldsymbol{\beta} = (\beta_{jk}) \in \mathbb{R}^{p\times(K-1)}$ denote a basis of $\mathcal{S}$. Intuitively, a covariate X_j is non-contributory in the optimal rule (3) if and only if $\beta_{jk} = 0$ for $k = 1, \ldots, K-1$. Accordingly, we define the active set $\mathcal{A} = \{j \mid \beta_{jk} \neq 0 \text{ for some } k\}$, denoting the index set of important variables. Let $s = |\mathcal{A}|$ denote the sparsity level. In fact, the definition of $\mathcal{A}$ is independent of how we choose the basis matrix of $\mathcal{S}$. Therefore, we have more options for parameterizing $\mathcal{S}$. One option is $\boldsymbol{\eta} = \boldsymbol{\Sigma}_x^{-1}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{K-1}) \in \mathbb{R}^{p\times(K-1)}$, where $\boldsymbol{\Sigma}_x = \text{Cov}(\mathbf{X}) \in \mathbb{R}^{p\times p}$. The following lemma validates that $\boldsymbol{\eta}$ equivalently spans $\mathcal{S}$:

Lemma 2: The subspace $\mathcal{S} = \text{span}(\boldsymbol{\eta})$ and $\delta^*_{\mathbf{X}}(\boldsymbol{x}) = \delta^*_{\boldsymbol{\eta}^\top \mathbf{X}}(\boldsymbol{\eta}^\top \boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^p$, where $\delta^*_{\boldsymbol{\eta}^\top \mathbf{X}}$ denotes the optimal rule for $\boldsymbol{\eta}^\top \mathbf{X}$ and is defined similar to $\delta^*_{\boldsymbol{\beta}^\top \mathbf{X}}$.

The new reparameterization of $\mathcal{S}$ in Lemma 2 lays the foundation for our estimation procedure. As we will see shortly, the sparse estimation of $\boldsymbol{\eta}$ can be cast as a group-Lasso problem, which allows us to leverage the computationally efficient group-Lasso algorithm. The active set $\mathcal{A}$ can be equivalently defined as $\mathcal{A} = \{j \mid \eta_{jk} \neq 0 \text{ for some } k\}$. Let $\mathbf{Y} = \{\pi_1^{-1} I(Y = 1), \ldots, \pi_{K-1}^{-1} I(Y = K-1)\}^\top \in \mathbb{R}^{K-1}$ denote the scaled indicator vector of response Y, where $I(\cdot)$ is the indicator function. The matrix $\boldsymbol{\eta}$ can be recovered from a least-square problem, stated in the following lemma:

Lemma 3: The matrix $\boldsymbol{\eta} = \text{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^{p\times(K-1)}} \text{E}\|\mathbf{Y} - \boldsymbol{\alpha}^\top \mathbf{X}\|_2^2$.

According to Lemma 3, when the sparsity structure for $\mathcal{S}$ is demanded, the least-square formulation in (3) motivates a sparse estimation of $\boldsymbol{\eta}$ through solving a group-Lasso penalized problem. Moreover, to alleviate the adverse impact of heavy-tailedness, we truncate the predictors before fitting the model with group-Lasso. For the predictor $\mathbf{X}$, we truncate it with a proper thresholding value $\tau > 0$, and obtain the truncated predictor $\widetilde{\mathbf{X}}$. Specifically, each element of the truncated predictor $\widetilde{\mathbf{X}}$ is defined as $\widetilde{X}_j = \text{sign}(X_j)(|X_j| \wedge \tau)$, for $j = 1, \ldots, p$. We choose the data truncation due to its simplicity and ease of implementation. We will show that by appropriately selecting the truncation value and capping the outliers, the group-Lasso estimator gains consistency in subspace estimation, variable selection, and prediction. In fact, the data truncation procedure has been widely studied in (generalized) linear models, see Fan et al. (2021) and Zhu and Zhou (2021), for example. However, these methods are specifically designed for regression and cannot be directly applied to our classification problem.

Let $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, denote n i.i.d. samples. For each response Y_i, we construct the scaled indicator vector $\widehat{\mathbf{Y}}_i = \{\widehat{\pi}_1^{-1} I(Y_i = 1), \ldots, \widehat{\pi}_{K-1}^{-1} I(Y_i = K-1)\}^\top$, where $\widehat{\pi}_k = n_k/n$ is the estimated prior for response and $n_k = \sum_{i=1}^n I(Y_i = k)$ denotes the sample size of the kth class. For each predictor $\mathbf{X}_i$, we truncate it with a proper thresholding value $\tau > 0$, and obtain the truncated predictor $\widetilde{\mathbf{X}}_i$. Then, let $\mathbb{Y}_n = (\widehat{\mathbf{Y}}_1, \ldots, \widehat{\mathbf{Y}}_n)^\top \in \mathbb{R}^{n\times(K-1)}$ and $\mathbb{X}_n = (\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_n)^\top \in \mathbb{R}^{n\times p}$ denote the data matrices, we propose the sparse estimation of $\boldsymbol{\eta}$ by solving the following group-Lasso penalized problem:

$$\widehat{\boldsymbol{\eta}} = \underset{\boldsymbol{\alpha}\in\mathbb{R}^{p\times(K-1)}}{\text{argmin}} \frac{1}{2n}\|\mathbb{Y}_n - \mathbb{X}_n\boldsymbol{\alpha}\|_F^2 + \lambda\|\boldsymbol{\alpha}\|_{2,1}, \tag{4}$$

where $\lambda > 0$ is the tuning parameter. Accordingly, the estimation of $\mathcal{S}$ is $\widehat{\mathcal{S}} = \text{span}(\widehat{\boldsymbol{\eta}})$. Based on the definition of $\mathcal{A}$, we estimate it by $\widehat{\mathcal{A}} = \{j \mid \widehat{\eta}_{jk} \neq 0 \text{ for some } k\}$. The group-Lasso penalized problem (4) can be readily solved by off-the-shelf group-Lasso algorithms. In our implementation, we use the R package **glmnet** (Friedman et al., 2010) for solving (4).

Once we obtain the sparse estimator $\widehat{\boldsymbol{\eta}}$ from the group-Lasso penalized optimization problem (4), we develop the classifier using the "projection-and-prediction" strategy. According to Lemma 2, given $\boldsymbol{\eta}$, the classification rule $\delta^*_{\boldsymbol{\eta}^\top \mathbf{X}}(\boldsymbol{\eta}^\top \boldsymbol{x})$ is equivalent to $\delta^*_{\mathbf{X}}(\boldsymbol{x})$, achieving the optimal rate. Let $\mathbf{Z} = \boldsymbol{\eta}^\top \mathbf{X} \in \mathbb{R}^{K-1}$ denote the reduced variates. The following results describe the distribution of $\mathbf{Z}$ and the corresponding optimal rule.

Lemma 4: Within each class $k \in \{1, \ldots, K\}$, the reduced variates $\mathbf{Z} = \boldsymbol{\eta}^\top \mathbf{X}$ follow the conditional EC distribution such that $\mathbf{Z} \mid (Y = k) \sim \text{EC}_{K-1}(\boldsymbol{\theta}_k, \boldsymbol{\Phi}; g)$, where $\boldsymbol{\theta}_k = \boldsymbol{\eta}^\top \boldsymbol{\mu}_k$ and $\boldsymbol{\Phi} = \boldsymbol{\eta}^\top \boldsymbol{\Sigma} \boldsymbol{\eta}$. For any observation $\boldsymbol{x} \in \mathbb{R}^p$, the optimal rule for $\boldsymbol{z} = \boldsymbol{\eta}^\top \boldsymbol{x}$ is given by

$$\delta^*_{\mathbf{Z}}(\boldsymbol{z}) \equiv \delta^*_{\mathbf{X}}(\boldsymbol{x}) = \underset{k\in\{1,\ldots,K\}}{\text{argmax}} (\boldsymbol{\theta}_k - \boldsymbol{\theta}_1)^\top \boldsymbol{\Phi}^{-1}\{\boldsymbol{z} - (\boldsymbol{\theta}_k + \boldsymbol{\theta}_1)/2\}. \tag{5}$$

With the estimator $\widehat{\boldsymbol{\eta}}$, the estimation of parameters $\boldsymbol{\Phi}$ and $\boldsymbol{\theta}_k$, for $k = 1, \ldots, K$, then follows as $\widehat{\boldsymbol{\Phi}} = \widehat{\boldsymbol{\eta}}^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\eta}}$ and $\widehat{\boldsymbol{\theta}}_k = \widehat{\boldsymbol{\eta}}^\top \widehat{\boldsymbol{\mu}}_k$, where $\widehat{\boldsymbol{\mu}}_k = n_k^{-1} \sum_{Y_i=k} \widetilde{\mathbf{X}}_i$ denotes the sample mean for class k and $\widehat{\boldsymbol{\Sigma}} = K^{-1} \sum_{k=1}^K \sum_{Y_i=k} (n_k - 1)^{-1} (\widetilde{\mathbf{X}}_i - \widehat{\boldsymbol{\mu}}_k)(\widetilde{\mathbf{X}}_i - \widehat{\boldsymbol{\mu}}_k)^\top$ is the pooled sample covariance matrix. By plugging the estimates $\widehat{\boldsymbol{\theta}}_k$ and $\widehat{\boldsymbol{\Phi}}$ into (5), we obtain the sample-version optimal rule, denoted by $\widehat{\delta}_n$. For a new observation $(Y^*, \mathbf{X}^*)$ independent of the training set, we reduce $\mathbf{X}^*$ to $\widehat{\mathbf{Z}}^* = \widehat{\boldsymbol{\eta}}^\top \mathbf{X}^*$ and make a

Downloaded from https://academic.oup.com/biometrics/article/82/1/ujag039/8498964 by guest on 27 February 2026

prediction as follows:

$$\widehat{\delta}_n(\mathbf{X}^*) = \underset{k\in\{1,\ldots,K\}}{\operatorname{argmax}} (\widehat{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_1)^\top \widehat{\boldsymbol{\Phi}}^{-1} \{\widehat{\mathbf{Z}}^* - (\widehat{\boldsymbol{\theta}}_k + \widehat{\boldsymbol{\theta}}_1)/2\}^\top. \tag{6}$$

Our proposed method benefits from the data projection both computationally and theoretically. Under the sparsity assumption of $\mathcal{S}$, or equivalently $\boldsymbol{\eta}$, only a proportion of elements in $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_k$ is involved in the optimal rule (5). Therefore, once the estimator $\widehat{\boldsymbol{\eta}}$ accurately recovers the active set $\mathcal{A}$, the computation in (6) becomes more efficient. The reduced number of parameters also facilitates the theoretical justification and simplifies our theoretical proof, so long as s grows with n at a slow rate.

Some remarks on formulation (4) are given as follows. Formulation (4) bears resemblance to MSDA method, introduced by Mai et al. (2019). However, there are several key distinctions between the two approaches. First, MSDA assumes the LDA model and relies on the normality of $\mathbf{X} \mid Y$. In contrast, our SODA method is designed for a more general EDA model, including the LDA model as a special case. Second, MSDA does not account for the heavy-tailedness of the data, which can lead to performance degradation when the LDA model assumption is violated. Third, MSDA implements a block-wise gradient descent algorithm while we solves a group-Lasso problem.

4.2 Tuning parameter selection

In SODA, we have two tuning parameters, λ and τ. While we could use cross-validation to tune these two parameters simultaneously, such an approach may be time-consuming. For efficient computation, we consider a sequential tuning procedure, where we first tune λ by fixing τ at a small value τ_0, and then fix λ at the chosen value to select τ. Our sequential tuning strategy is motivated by Fan et al. (2021), which tunes the sparsity-encouraging parameter and the truncation parameter sequentially.

In this sequential procedure, it is especially important to select a good λ in the first stage to ensure the success of the final prediction. To achieve this goal, we first use cross-validation to produce an appropriate range for candidate tuning parameters, and then use an information criterion to fine tune λ. In particular, we first select the baseline value λ_0 using the cross-validation function for group Lasso in the R package **glmnet**, then the candidate sequence for λ is constructed by multiplying the baseline value λ_0 by a sequence of factors. In the simulations, the factors consist of 15 equally spaced values from 1.5 to 0.01. For the real data analysis, we adopt finer tuning grids by setting the factors to 20 equally spaced values ranging from 1.5 to 0.001. Moreover, the value τ_0 is fixed at the 0.98 quantile of $\{|X_{i,j}| \mid i = 1, \ldots, n, j = 1, \ldots, p\}$. Note that the information criterion is used in fine tuning to speed up computation, where we only fit the solution path once. For the same reason, information criteria have been considered in high dimensions; see Fan and Tang (2013), Wang and Leng (2007), and Wang et al. (2009; 2007), for example. We consider the following information criterion:

$$\mathrm{IC}(\lambda) = n \cdot \log\left(\|\mathbb{Y}_n - \mathbb{X}_n \widehat{\boldsymbol{\eta}}(\tau_0, \lambda)\|_F^2/n\right) + \log p \cdot \log(\log n) \cdot k(K-1),$$

where $\widehat{\boldsymbol{\eta}}(\tau_0, \lambda)$ denotes the estimator of $\boldsymbol{\eta}$, $K-1$ and k denote the number of columns and non-zero rows of $\widehat{\boldsymbol{\eta}}(\tau_0, \lambda)$, respectively, such that $k(K-1)$ equals to the number of non-zero elements of $\widehat{\boldsymbol{\eta}}(\tau_0, \lambda)$. Our motivating references also account for sparsity in the information criteria only through the number of non-zero elements, too. Furthermore, the term $\log p \cdot \log(\log n)$ in $\mathrm{IC}(\lambda)$ is inspired by the generalized information criterion in Fan and Tang (2013), which adapts to the dimensionality p even when data are high-dimensional. The optimal λ is the one yielding the smallest $\mathrm{IC}(\lambda)$, denoted by λ_{opt}.

In the second step, with λ fixed at λ_{opt}, we tune τ via 5-fold cross-validation. In simulations, the candidate sequence for τ comprises of 10 quantiles of $\{|X_{i,j}| : i = 1, \ldots, n, j = 1, \ldots, p\}$, equally spaced from the 80th to the 99.9th percentiles, with infinity also included to represent no truncation. For the real applications, we consider more candidate values, including the infinity and 15 percentiles equally spaced between the 80th and 99.9th percentiles. For each τ from the candidate sequence, denote the SODA estimator on the mth fold by $\widehat{\boldsymbol{\eta}}^{(m)}(\tau, \lambda_{\mathrm{opt}})$, where $m = 1, \ldots, 5$. Then we record the averaged loss $(\sum_{m=1}^{5} \|\mathbb{Y}_n - \mathbb{X}_n \widehat{\boldsymbol{\eta}}^{(m)}(\tau, \lambda_{\mathrm{opt}})\|_F^2)/5$ and choose the optimal τ_{opt} as the τ yielding the smallest loss.

5 Theory

In this section, we provide theoretical justification for the consistency of the prediction accuracy of our proposal. Due to the space limit, the consistency results of subspace estimation and variable selection are provided in Section S.4 of the Supplementary Materials.

The following finite moment assumption is required for establishing the theoretical results of our proposal:

(1) For any $j, k = 1, \ldots, p$, there exists some constant $R > 0$ such that $\mathrm{E}(X_j^2 X_k^2) \le R$.

Assumption (1) requires the predictors to have a finite fourth-moment, which is less restrictive compared to the light-tailed distributional assumption in the classical LDA model. This finite moment assumption is also common in robust statistical inferences. For instance, Ren and Mai (2022) adopt the finite conditional fourth-moment assumption in high-dimensional classification, and Fan et al. (2021) impose the finite fourth-moment assumption in trace regression. To save space, other technical assumptions and their explanations are gathered in Section S.3 of the Supplementary Materials.

Consider a pair of new data $(Y^*, \mathbf{X}^*)$, independent of the training set that is used to obtain the estimated rule $\widehat{\delta}_n$ in (6). Let $R_n = (1/K)\sum_{k=1}^{K} \Pr\{\widehat{\delta}_n(\mathbf{X}^*) \neq k \mid Y^* = k\}$ denote the balanced rate of the estimated rule, and let $R^* = (1/K)\sum_{k=1}^{K} \Pr\{\delta^*_{\mathbf{X}}(\mathbf{X}^*) \neq k \mid Y^* = k\}$ denote the optimal rate. Note that R_n is a random variable, with its randomness arising from the stochastic $\widehat{\delta}_n$.

Downloaded from https://academic.oup.com/biometrics/article/82/1/ujag039/8498964 by guest on 27 February 2026

Theorem 2: Under Assumptions (A1)–(A9), assume that $C_1 s^2 \log p \le n$ for some sufficiently large constant $C_1 > 0$ and that $\log p \ge n^{1-2k}$ for some constant $0 < k < 1/2$. By taking $\lambda \asymp s(\log p/n)^{1/2}$ and $\tau \asymp (n/\log p)^{1/4}$, there exist some constants $C, C' > 0$ such that with probability at least $1 - C\exp(-C'n^{1-2k})$, we have $|R_n - R^*| \lesssim (s^4 \log p/n)^{1/6}$.

According to Theorem 2, compared with sparse LDA methods, which rely on the conditional normality assumption, our proposal attains consistency of the optimal rate under a much weaker moment Assumption (1). Therefore, our approach provides accurate classification performance even in the presence of heavy-tailed data, demonstrating its robustness.

6 Simulation

In this section, we compare our proposal with several competitors, including sparse logistic regression method (SLR; Friedman et al., 2010) and two off-the-shelf sparse LDA methods, namely, multiclass sparse discriminant analysis (MSDA; Mai et al., 2019) and adaptive linear discriminant analysis rule (AdaLDA; Cai and Zhang, 2019). Additionally, we also include into comparison two popular machine learning classifiers, including support vector machine (SVM; Cortes and Vapnik, 1995) and random forest (RF; Breiman, 2001). Moreover, as a benchmark for evaluating the prediction accuracy, the optimal rate is included. We implement SLR, MSDA, SVM, and RF using the R packages **glmnet**, **msda**, **kernlab**, and **randomForest**, respectively. The code for AdaLDA is publicly available from the authors' website. For SODA, tuning parameters are selected via the sequential procedure described in Section 4. For SLR, MSDA, and SVM, tuning parameters are chosen by 5-fold cross-validation using the default cross-validation functions in their respective R packages. For RF, the optimal parameters are determined by the built-in tuning function in the **randomForest** package, which by default searches for the parameters that minimize the out-of-bag error. As AdaLDA is tuning-free, no additional parameter selection is required.

6.1 Balanced data

We set the dimension $p = 1000$ and generate a training set consisting of $n = 100 \times K$ samples, with each class containing 100 samples. We also generate a separate testing set with a sample size of $250 \times K$. Two sparse precision matrices are considered, including the Erdős–Rényi random graph $\mathbf{\Omega}_1$ and the block sparse model $\mathbf{\Omega}_2$, which are also used in Cai and Zhang (2019) and Cai and Liu (2011). The definitions of $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ are given in Section S.1 of the Supplementary Materials. We take the covariance matrix $\mathbf{\Sigma}_i = \mathbf{\Omega}_i^{-1}$ for $i = 1, 2$.

We consider six EDA models where the conditional distribution of $\mathbf{X} \mid Y$ follows a multivariate t distribution. Let $\boldsymbol{\beta}_k = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_{k+1} - \boldsymbol{\mu}_1)$, for $k = 1, \ldots, K-1$. In each model, we first specify $\mathbf{\Sigma}$ and $\boldsymbol{\beta}_k$, then take $\boldsymbol{\mu}_1 = -\mathbf{\Sigma}(K^{-1}\sum_{k=1}^{K-1}\boldsymbol{\beta}_k)$ and $\boldsymbol{\mu}_{k+1} = \mathbf{\Sigma}\{\boldsymbol{\beta}_k - K^{-1}(\sum_{k=1}^{K-1}\boldsymbol{\beta}_k)\}$, for $k = 1, \ldots, K-1$, such that $\mathrm{E}(\mathbf{X}) = \mathbf{0}$. For $\boldsymbol{\beta}_k = (\beta_{k,i})$, we only describe the non-zero elements and the rest of the elements are set to zero. In addition, $\mathrm{Unif}(a, b)$ denotes the uniform distribution supported on the interval (a, b):

(1) $K = 2$, $\beta_{1,i}$ takes values from Unif(8,9) for $1 \le i \le 4$, $\mathbf{\Sigma} = \mathbf{\Sigma}_1$;
(2) $K = 2$, $\beta_{1,i}$ takes values from Unif(0.4,1.4) for $1 \le i \le 4$, $\mathbf{\Sigma} = \mathbf{\Sigma}_2$;
(3) $K = 3$, $\beta_{1,i}, \beta_{2,j}$ take values from Unif(5, 5.5), and $\beta_{1,j}, \beta_{2,i}$ take values from Unif(0.5, 1) for $1 \le i \le 4$, $5 \le j \le 8$, $\mathbf{\Sigma} = \mathbf{\Sigma}_1$;
(4) $K = 3$, $\beta_{1,i}, \beta_{2,j}$ take values from Unif(1, 1.5), and $\beta_{1,j}, \beta_{2,i}$ take values from Unif(0.5, 1) for $1 \le i \le 4$, $5 \le j \le 8$, $\mathbf{\Sigma} = \mathbf{\Sigma}_2$;
(5) $K = 4$, $\beta_{1,i}$ takes values from Unif(5, 5.5) for $1 \le i \le 4$, and Unif(0.5, 1) for $5 \le i \le 12$, $\beta_{2,j}$ takes values from Unif(0.5, 1) for $1 \le j \le 4$ and $9 \le j \le 12$, and Unif(5, 5.5) for $5 \le j \le 8$, and $\beta_{3,k}$ takes values from Unif(0.5, 1) for $1 \le k \le 8$, and Unif(5, 5.5) for $9 \le k \le 12$, $\mathbf{\Sigma} = \mathbf{\Sigma}_1$.
(6) $K = 4$, $\beta_{1,i}$ takes values from Unif(1, 1.5) for $1 \le i \le 4$, and Unif(0.5, 1) for $5 \le i \le 12$, $\beta_{2,j}$ takes values from Unif(0.5, 1) for $1 \le j \le 4$ and $9 \le j \le 12$, and Unif(1, 1.5) for $5 \le j \le 8$, and $\beta_{3,k}$ takes values from Unif(0.5, 1) for $1 \le k \le 8$, and Unif(1, 1.5) for $9 \le k \le 12$, $\mathbf{\Sigma} = \mathbf{\Sigma}_2$.

The values of the non-zero elements in $\boldsymbol{\beta}_k$ are chosen such that the optimal rate is controlled at an appropriate level.

We vary the degrees of freedom ν over the values $\{2.1, 4.1, 5, 7, \infty\}$ to investigate how the magnitude of heavy-tailedness affects the performance of each competitor. When $\nu = \infty$, the multivariate t distribution degenerates to the multivariate normal distribution.

The balanced rates are reported in Table 1. It can be seen that SODA has advantages over other competitors, and its balanced rate approaches the optimal rate across all settings. Furthermore, the smaller the value of ν, the greater the advantage of SODA, highlighting its robustness against heavy-tailedness. Specially, when $\nu = \infty$, the EDA model reduces to the LDA model, and SODA continuous to exhibit accurate prediction performance.

To illustrate these findings more clearly, we present the R versus ν plots for each competitor, which are relegated to Section S.2.3 of the Supplementary Materials.

Moreover, even when Assumption (A1) is violated when $\nu = 2.1$, SODA still maintains accurate predictive performance. The subspace estimation and variable selection results are presented in Sections S.2.1 and S.2.2 of the Supplementary Materials, respectively. The results show that SODA dominates sparse LDA methods in estimating the subspace, and such a discrepancy becomes more pronounced when ν decreases. Moreover, SODA achieves perfect variable selection even when the predictors have heavy tails.

We also explore EC distributions beyond the multivariate t distribution. In Section S.2.4 of the Supplementary Materials, we conduct simulations under the EDA model with $\mathbf{X} \mid Y$ following multivariate Laplace distribution. Although the Laplace distribution is not theoretically heavy-tailed, it exhibits heavier tails than the normal distribution. Once again,

Downloaded from https://academic.oup.com/biometrics/article/82/1/ujag039/8498964 by guest on 27 February 2026

Table 1 Means (and standard errors) of the balanced rate R(%) in Models (M1)–(M6) with the degree of freedom $\nu = 2.1, 4.1, 5, 7, \infty$.

Method	Balanced rate R (%)									
	$\nu = 2.1$	$\nu = 4.1$	$\nu = 5$	$\nu = 7$	$\nu = \infty$	$\nu = 2.1$	$\nu = 4.1$	$\nu = 5$	$\nu = 7$	$\nu = \infty$
	Model (M1)					Model (M2)				
Optimal	17.7 (0.2)	11.8 (0.1)	10.6 (0.1)	9.2 (0.1)	6.2 (0.1)	18.6 (0.2)	12.5 (0.1)	11.5 (0.1)	9.9 (0.1)	7.0 (0.1)
SODA	**18.9 (0.2)**	**12.5 (0.2)**	**11.2 (0.1)**	**9.7 (0.1)**	**6.9 (0.1)**	**20.1 (0.2)**	**13.4 (0.2)**	**12.4 (0.2)**	**11.0 (0.2)**	**8.0 (0.1)**
MSDA	37.6 (0.7)	14.5 (0.4)	13.2 (0.3)	11.5 (0.3)	8.6 (0.3)	30.1 (0.6)	15.0 (0.3)	14.0 (0.3)	12.6 (0.3)	8.9 (0.2)
AdaLDA	27.2 (0.2)	14.3 (0.2)	13.3 (0.2)	10.9 (0.2)	8.1 (0.1)	28.3 (0.2)	15.3 (0.2)	13.6 (0.2)	12.2 (0.2)	8.1 (0.1)
SLR	22.8 (0.3)	14.4 (0.2)	12.9 (0.2)	11.0 (0.2)	7.9 (0.2)	27.4 (0.6)	15.9 (0.2)	14.4 (0.2)	12.8 (0.2)	9.2 (0.2)
SVM	37.7 (0.2)	26.8 (0.2)	24.6 (0.2)	22.4 (0.2)	18.2 (0.2)	38.4 (0.3)	26.4 (0.2)	25.3 (0.3)	23.7 (0.3)	20.5 (0.2)
RF	24.9 (0.4)	15.8 (0.2)	15.1 (0.2)	13.6 (0.2)	11.0 (0.3)	28.5 (0.2)	22.6 (0.2)	21.0 (0.2)	19.4 (0.2)	16.7 (0.2)
	Model (M3)					Model (M4)				
Optimal	21.5 (0.1)	11.3 (0.1)	9.8 (0.1)	7.6 (0.1)	4.0 (0.1)	18.1 (0.2)	8.4 (0.1)	7.0 (0.1)	5.1 (0.1)	2.4 (0.1)
SODA	**23.8 (0.2)**	**12.5 (0.1)**	**10.7 (0.1)**	**8.9 (0.1)**	**4.6 (0.1)**	**19.5 (0.2)**	**9.4 (0.1)**	**8.0 (0.1)**	**6.2 (0.1)**	**3.0 (0.1)**
MSDA	42.3 (0.5)	19.8 (0.4)	17.9 (0.3)	15.9 (0.3)	10.5 (0.3)	24.8 (0.3)	12.4 (0.2)	11.0 (0.2)	9.4 (0.1)	5.8 (0.1)
AdaLDA	28.6 (0.2)	14.8 (0.1)	12.9 (0.1)	9.9 (0.1)	4.9 (0.1)	25.1 (0.2)	14.5 (0.1)	10.8 (0.1)	8.6 (0.1)	3.8 (0.1)
SLR	27.2 (0.2)	15.0 (0.2)	12.2 (0.2)	10.0 (0.2)	5.5 (0.1)	23.8 (0.3)	11.1 (0.1)	9.6 (0.1)	7.4 (0.1)	3.7 (0.1)
SVM	43.2 (0.6)	21.4 (0.2)	20.3 (0.2)	15.9 (0.3)	10.5 (0.3)	38.8 (0.4)	21.2 (0.2)	18.3 (0.2)	16.3 (0.2)	11.3 (0.1)
RF	26.1 (0.6)	15.4 (0.2)	14.3 (0.2)	14.0 (0.2)	9.1 (0.3)	27.3 (0.4)	14.6 (0.3)	13.1 (0.3)	11.9 (0.3)	8.1 (0.1)
	Model (M5)					Model (M6)				
Optimal	25.2 (0.1)	14.3 (0.1)	12.3 (0.1)	9.9 (0.1)	5.6 (0.1)	24.8 (0.1)	13.1 (0.1)	10.5 (0.1)	8.3 (0.1)	3.9 (0.1)
SODA	**28.9 (0.2)**	**15.6 (0.2)**	**14.3 (0.1)**	**11.8 (0.1)**	**6.4 (0.1)**	**27.8 (0.3)**	**14.9 (0.2)**	**11.9 (0.2)**	**9.9 (0.1)**	**4.9 (0.1)**
MSDA	49.2 (0.4)	27.5 (0.2)	25.6 (0.2)	23.4 (0.2)	18.6 (0.2)	34.4 (0.5)	17.0 (0.3)	13.9 (0.1)	12.0 (0.1)	6.2 (0.1)
AdaLDA	32.2 (0.3)	19.2 (0.2)	16.1 (0.1)	13.3 (0.1)	7.3 (0.1)	33.4 (0.2)	16.2 (0.2)	13.2 (0.1)	10.7 (0.1)	4.9 (0.1)
SLR	34.5 (0.3)	18.6 (0.1)	16.3 (0.1)	12.8 (0.1)	7.6 (0.1)	32.6 (0.3)	16.1 (0.2)	14.3 (0.1)	10.9 (0.1)	5.7 (0.1)
SVM	54.3 (0.2)	35.5 (0.2)	35.1 (0.4)	31.5 (0.3)	25.0 (0.2)	48.6 (0.2)	37.7 (0.3)	33.1 (0.2)	24.1 (0.2)	20.1 (0.1)
RF	36.2 (0.4)	28.1 (0.2)	27.0 (0.2)	25.0 (0.2)	18.1 (0.2)	37.6 (0.2)	24.7 (0.2)	21.4 (0.2)	20.0 (0.2)	16.4 (0.2)

The best result excluding optimal rate in each setting is highlighted.

our proposal demonstrates its robustness, achieving the best performance in all aspects.

6.2 Imbalanced data

In this section, we examine how the imbalance affects the performance of each competitor. We focus on Models (M1) and (M4), where the conditional distribution $\mathbf{X} \mid Y$ follows the multivariate t distribution, and Model (M8), where $\mathbf{X} \mid Y$ follows the multivariate Laplace distribution. The dimension p is fixed at 1000. For Models (M1) and (M8) with $K = 2$, we generate $n = 200$ samples in the training set. For Model (M4) with $K = 3$, the training set consists of $n = 300$ samples. To measure the magnitude of data imbalance, we introduce the imbalanced factor M. Specifically, for binary classification, M represents the ratio of sample sizes of the first and second classes. For three-class classification, we keep the sample size of the second class at 100 and that of the other two classes at 200, then M represents the ratio of sample sizes of the first and the third classes. A larger M indicates greater imbalance level. For example, when $M = 1$, the sample sizes are equal. When $M = 4$, the ratio of sample sizes is $4 : 1$ for $K = 2$ and $8 : 5 : 2$ for $K = 3$. We also generate a separate testing set of size $2.5n$, with the ratio of sample sizes following that in the training set.

In Table 2, we report the balanced rate R as the imbalanced factor M varies over 1, 2, 3, and 4. We also report the error rate in the minority class, denoted by R_m. Recall that the rate R_m is more crucial than R in certain real applications. It can be seen from Table 2 that SODA consistently outperforms all competitors across all settings, with its balanced rate R approaching the optimal rate. Additionally, as the imbalanced factor M increases, the balanced rate R of SODA remains robust. In comparison, the balanced rates of MSDA, SVM, and RF deteriorate significantly. Furthermore, SODA achieves the lowest R_m in all settings, indicating its superior prediction accuracy for the minority class on imbalanced data. As M increases, R_m of SODA remains stable while other competitors struggle with predicting the minority class. Particularly, when $M = 4$, where the minority class makes up 20% of the data in Models (M1) and (M8), and 13.3% in Model (M4), SODA maintains accurate prediction, especially for samples in the minority class.

The subspace estimation and variable selection results are detailed in Section S.2.5 of the Supplementary Materials. Our proposal outperforms other competitors by a significant margin in subspace estimation and achieves almost perfect variable selection.

7 Gene Expression Data Analysis

Distinguishing between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung is of great

Downloaded from https://academic.oup.com/biometrics/article/82/1/ujag039/8498964 by guest on 27 February 2026

Table 2 Means (and standard errors) of the balanced rate R(%) and the error rate in the minority class R_m(%) in Models (M1), (M4), and (M8).

Method	M=1		M=2		M=3		M=4	
	R (%)	R_m (%)	R (%)	R_m (%)	R (%)	R_m (%)	R (%)	R_m (%)
				Model (M1)				
Optimal	11.8 (0.1)	11.8 (0.2)	11.7 (0.1)	11.5 (0.2)	12.0 (0.2)	12.2 (0.3)	12.1 (0.2)	12.4 (0.1)
SODA	**12.5 (0.2)**	**12.4 (0.2)**	**12.5 (0.2)**	**11.1 (0.3)**	**13.1 (0.2)**	**11.2 (0.3)**	**13.1 (0.2)**	**11.1 (0.4)**
MSDA	14.5 (0.3)	14.4 (0.4)	15.0 (0.4)	16.1 (0.5)	16.5 (0.5)	18.1 (0.7)	16.6 (0.5)	18.7 (0.7)
AdaLDA	14.3 (0.2)	14.3 (0.3)	14.3 (0.2)	14.5 (0.2)	14.7 (0.2)	15.0 (0.2)	15.1 (0.2)	16.3 (0.2)
SLR	14.5 (0.2)	14.4 (0.3)	16.2 (0.2)	23.9 (0.3)	19.2 (0.3)	30.8 (0.3)	21.3 (0.3)	35.1 (0.4)
SVM	26.8 (0.2)	26.9 (0.3)	29.8 (0.2)	47.1 (0.3)	35.3 (0.3)	64.8 (0.4)	40.4 (0.3)	78.2 (0.5)
RF	15.8 (0.2)	15.1 (0.5)	21.6 (0.3)	29.2 (0.4)	23.4 (0.2)	35.0 (0.4)	25.5 (0.2)	39.2 (0.4)
				Model (M4)				
Optimal	8.4 (0.1)	7.0 (0.2)	8.5 (0.1)	7.2 (0.2)	8.6 (0.1)	7.1 (0.2)	8.4 (0.1)	6.8 (0.3)
SODA	**9.4 (0.1)**	**8.8 (0.2)**	**9.5 (0.1)**	**8.5 (0.2)**	**9.7 (0.1)**	**8.6 (0.3)**	**9.6 (0.1)**	**8.3 (0.4)**
MSDA	12.4 (0.2)	11.1 (0.3)	12.6 (0.2)	12.1 (0.3)	12.8 (0.2)	12.8 (0.2)	13.3 (0.2)	13.9 (0.5)
AdaLDA	14.5 (0.1)	13.6 (0.3)	14.8 (0.2)	13.9 (0.3)	14.9 (0.1)	14.1 (0.3)	15.0 (0.2)	15.9 (0.3)
SLR	11.1 (0.1)	9.1 (0.2)	11.9 (0.2)	9.3 (0.3)	12.6 (0.2)	11.1 (0.3)	13.7 (0.2)	13.6 (0.4)
SVM	21.2 (0.2)	20.5 (0.3)	23.3 (0.2)	35.9 (0.4)	26.8 (0.2)	48.6 (0.5)	29.9 (0.2)	59.4 (0.6)
RF	14.6 (0.3)	13.5 (0.4)	18.4 (0.4)	31.9 (1.0)	23.4 (0.6)	48.1 (1.6)	29.1 (0.2)	64.4 (0.7)
				Model (M8)				
Optimal	13.5 (0.2)	13.7 (0.2)	13.6 (0.2)	13.8 (0.3)	13.6 (0.2)	13.8 (0.3)	13.5 (0.2)	13.5 (0.3)
SODA	**14.2 (0.2)**	**14.3 (0.3)**	**14.5 (0.2)**	**14.5 (0.3)**	**15.0 (0.2)**	**12.1 (0.4)**	**16.0 (0.3)**	**12.6 (0.5)**
MSDA	15.3 (0.2)	15.2 (0.4)	16.1 (0.2)	17.6 (0.4)	16.9 (0.3)	19.5 (0.6)	18.0 (0.3)	22.7 (0.7)
AdaLDA	15.4 (0.2)	15.1 (0.3)	16.3 (0.3)	17.1 (0.4)	16.7 (0.3)	19.4 (0.5)	18.1 (0.3)	23.3 (0.8)
SLR	16.5 (0.2)	16.6 (0.3)	18.2 (0.2)	25.6 (0.3)	21.9 (0.3)	32.9 (0.3)	25.0 (0.4)	37.7 (0.3)
SVM	29.4 (0.2)	29.2 (0.4)	35.2 (0.3)	60.5 (0.6)	49.9 (0.0)	99.6 (0.1)	50.0 (0.0)	100.0 (0.0)
RF	25.3 (0.3)	25.4 (0.4)	26.5 (0.2)	20.7 (0.5)	28.1 (0.3)	17.7 (0.7)	29.7 (0.3)	16.8 (0.6)

The imbalanced factor M varies over 1, 2, 3, and 4. The best result excluding optimal rate in each setting is highlighted.

interest for medical diagnosis purpose. In this paper, we evaluate the proposed method by classifying between MPM and ADCA samples in the lung cancer data set studied by Gordon et al. (2002). The data set is composed of $n = 181$ samples, collected from patients who underwent surgery at Brigham and Women's Hospital between 1993 and 2001. Among these samples, $n_1 = 31$ are collected from MPM surgical specimens and $n_2 = 150$ are collected from ADCA surgical specimens, encoded by $Y = 1, 2$, respectively. For each sample, the expression levels of $p = 12533$ gene transcripts are recorded, denoted by $\mathbf{X} \in \mathbb{R}^p$. In Sections S.2.6 and S.2.7 of the Supplementary Materials, we provide additional analyses of a binary-class lung cancer data set and a three-class leukemia data set.

We randomly split the data into a training set and a testing set in an 85/15 ratio in a stratified manner for 100 times. In each data splitting, we first perform variable screening. On the training set, we compute the two-sample t-test statistic t_j for the jth gene X_j, $j = 1, \ldots, p$, defined as $t_j = (\hat{\mu}_{1j} - \hat{\mu}_{2j})/(s_j\sqrt{n_1^{-1} + n_2^{-1}})$, where $s_j = \sqrt{\{(n_1 - 1)s_{1j}^2 + (n_2 - 1)s_{2j}^2\}/(n_1 + n_2 - 2)}$, $\hat{\mu}_{kj}$ and s_{kj} are sample mean and standard deviation of X_j in class k, for $k = 1, 2$. Based on t_j, we conduct a t-test and keep top 2000 variables with the highest significance from the training set. The corresponding variables in the testing set are also kept. Then, we compare our proposal with MSDA, AdaLDA, SLR, SVM, and RF. The balanced rate, error rate in the minority class, and estimated sparsity level are shown in Table 3. Since the variable selection is not directly implemented by SVM and RF, we omit the estimated sparsity level of SVM and RF from the table.

Table 3 shows that SODA achieves the highest prediction accuracy with fewer variables than competing methods. Notably, SODA also attains the most accurate prediction on the minority class, whereas other methods perform poorly, leading to their high balanced rates. To further examine the tail behavior of the gene expression data, we select the 2000 genes on the full data set following the same variable screening procedure, and make the histogram of excess kurtosis coefficients of these genes in Figure 1. Approximately 72% exhibit excess kurtosis greater than 2, indicating heavy-tailed distributions. These results demonstrate that SODA is not only robust to heavy-tailed lung cancer data but also yields a more interpretable model through parsimonious variable selection.

Furthermore, we provide biological explanations for variables most frequently selected by SODA. In Figure 2, the histogram displays the selection frequency of all variables across 100 replicates, with four genes standing out: *S100P, CD36, ETV1*, and *IGHA1*. Shu et al. (2023) reported that high ex-

Downloaded from https://academic.oup.com/biometrics/article/82/1/ujag039/8498964 by guest on 27 February 2026

Table 3 Means (and standard errors) of the balanced rate $R(\%)$, the error rate in the minority class $R_m(\%)$, and the estimated sparsity level on the lung cancer data.

	SODA	MSDA	AdaLDA	SLR	SVM	RF
$R(\%)$	**0.8 (0.2)**	4.7 (1.2)	7.9 (1.2)	5.3 (0.5)	4.6 (0.6)	1.4 (0.3)
$R_m(\%)$	**1.1 (0.4)**	8.5 (1.2)	7.5 (0.9)	10.6 (1.0)	9.1 (1.2)	2.9 (0.6)
Sparsity	20.4 (0.4)	43.6 (2.7)	57.1 (3.3)	22.1 (0.3)	–	–

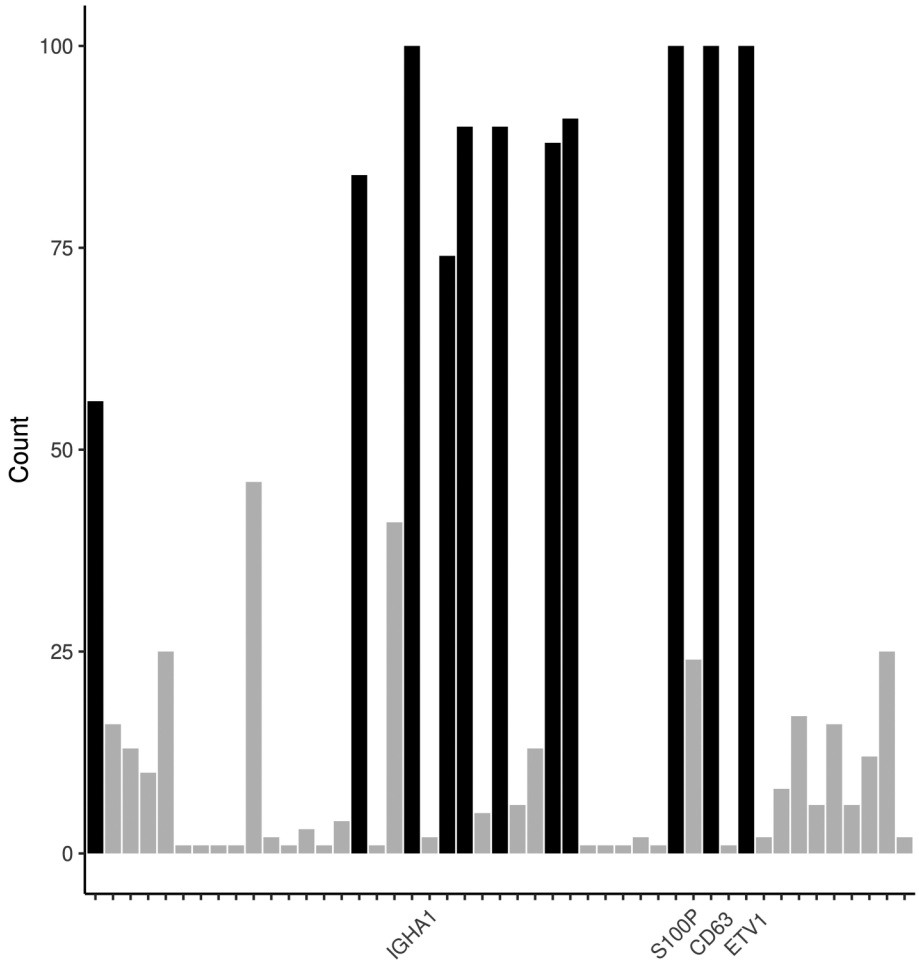

Figure 2 Appearance frequency of variables from the lung cancer data selected by SODA in 100 replicates. Bars with frequencies greater than 50 are highlighted.

pression of *S100P* is associated with tumor infiltration and poor prognosis, suggesting its role in promoting lung cancer progression through modulation of the tumor microenvironment. Liu et al. (2024) experimentally confirmed that inhibiting *CD36* reduces tumor growth and metastasis, highlighting its key role in lung cancer development. Similarly, *ETV1* has been shown to be upregulated in lung ADCA, where it promotes tumor proliferation and metastasis, indicating its potential as a therapeutic target (Zuo et al., 2020). Moreover, Lu et al. (2023) found that *IGHA1* expression is significantly elevated in tumor-associated immune cell populations, suggesting its involvement in immune regulation and its impact on patient prognosis.

We also provide the visualization of the dimension reduction effect of SODA. On the 2000 genes selected from the full data, we implement SODA and obtain the reduced data $\widehat{\boldsymbol{\eta}}^{\top}\mathbf{X}$ in the two classes. The estimated density curves are presented in Figure 3, in which the estimated direction separates the two classes well, facilitating accurate follow-up classification.

8 Discussion

In this paper, we propose a robust high-dimensional classifier that maintains efficiency when predictors have heavy tails. One constraint in the EDA model, also presents in the LDA

Downloaded from https://academic.oup.com/biometrics/article/82/1/ujag039/8498964 by guest on 27 February 2026

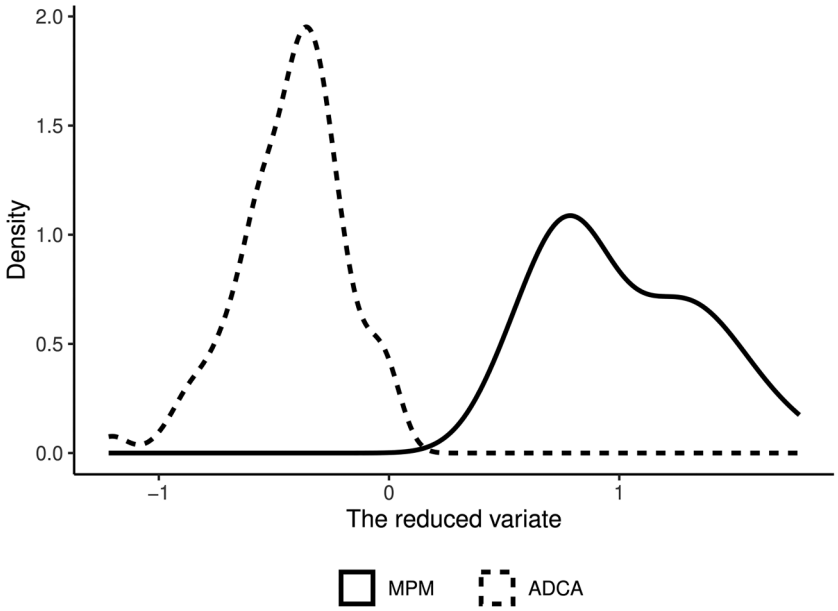

Figure 3 Density curves of the reduced variates $\widehat{\boldsymbol{\eta}}^{\top}\mathbf{X}$ in the lung cancer data.

model, is the homoscedasticity assumption of the conditional distribution $\mathbf{X} \mid Y$, which is violated in certain scenarios. It is known that the QDA model addresses this by assuming heteroscedastic variance. Therefore, a promising avenue for future research is to extend our proposal to a heteroscedastic model.

Acknowledgments

The authors thank the Associate Editor and an anonymous referee for very insightful comments that improved the overall quality of the paper.

Supplementary materials

Supplementary material is available at *Biometrics* online.

Web Appendices, Tables, and Figures referenced in Sections 3, 5, 6, and 7, and code are available with this paper at the Biometrics website on Oxford Academic.

Funding

J.Z. was partially supported by Grant 2024YFA1012200 from National Key R&D Program of China, Grant 12301365 from National Natural Science Foundation of China, and Grants WK2040000075 and WK2040250139 from Fundamental Research Funds for the Central Universities.

Conflicts of Interest

None declared.

Data Availability

The data that support the findings in this paper are available in **R** package **propOverlap**, and can be loaded via the **R** command **data(lung)**. The original data can also be downloaded from https://data.mendeley.com/datasets/ynp2tst2hh/4.

References

Bose, S., Pal, A., SahaRay, R. and Nayak, J. (2015). Generalized quadratic discriminant analysis. *Pattern Recognition*, 48, 2676–2684.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

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

Cai, T. and Zhang, L. (2019). High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81, 675–705.

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

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.

Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36, 2605–2637.

Fan, J., Wang, W. and Zhu, Z. (2021). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *The Annals of Statistics*, 49, 1239–1266.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal*

Downloaded from https://academic.oup.com/biometrics/article/82/1/ujag039/8498964 by guest on 27 February 2026

Statistical Society Series B: Statistical Methodology, 75, 531–552.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179–188.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.

Ghosh, A., SahaRay, R., Chakrabarty, S. and Bhadra, S. (2021). Robust generalised quadratic discriminant analysis. *Pattern Recognition*, 117, 107981.

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S. and et al. (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.

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.

Hennig, C. and Viroli, C. (2016). Quantile-based classifiers. *Biometrika*, 103, 435–446.

Johnson, M. (1987). *Multivariate Statistical Simulation: A Guide to Selecting and Generating Continuous Multivariate Distributions. Wiley Series in Probability and Statistics*. New York: Wiley.

Kaur, P., Singh, A. and Chana, I. (2021). Computational techniques and tools for omics data analysis: state-of-the-art, challenges, and future directions. *Archives of Computational Methods in Engineering*, 28, 4595–4631.

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

Liu, H., Guo, W., Wang, T., Cao, P., Zou, T., Peng, Y. and et al. (2024). CD36 inhibition reduces non-small-cell lung cancer development through AKT-mTOR pathway. *Cell Biology and Toxicology*, 40, 10.

Lu, H., Qian, J., Cheng, L., Shen, Y., Chu, T. and Zhao, C. (2023). Single-cell RNA-sequencing uncovers the dynamic changes of tumour immune microenvironment in advanced lung adenocarcinoma. *BMJ Open Respiratory Research*, 10, e001878.

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

Min, W., Liu, J. and Zhang, S. (2018). Edge-group sparse PCA for network-guided high dimensional data analysis. *Bioinformatics (Oxford, England)*, 34, 3479–3487.

Qiao, X. and Liu, Y. (2009). Adaptive weighted learning for unbalanced multicategory classification. *Biometrics*, 65, 159–168.

Ren, S. and Mai, Q. (2022). The robust nearest shrunken centroids classifier for high-dimensional heavy-tailed data. *Electronic Journal of Statistics*, 16, 3343–3384.

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

Shu, J., Jiang, J. and Zhao, G. (2023). Identification of novel gene signature for lung adenocarcinoma by machine learning to predict immunotherapy and prognosis. *Frontiers in Immunology*, 14, 1177847.

Tian, Y. and Feng, Y. (2025). Neyman-Pearson multi-class classification via cost-sensitive learning. *Journal of the American Statistical Association*, 120, 1164–1177.

Tong, X., Xia, L., Wang, J. and Feng, Y. (2020). Neyman-Pearson classification: parametrics and sample size requirement. *Journal of Machine Learning Research*, 21, 1–48.

Wakaki, H. (1994). Discriminant analysis under elliptical populations. *Hiroshima Mathematical Journal*, 24, 257–298.

Wang, H. and Leng, C. (2007). Unified LASSO estimation by least squares approximation. *Journal of the American Statistical Association*, 102, 1039–1048.

Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71, 671–683.

Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94, 553–568.

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.

Xiong, W., Härdle, W., Wang, J., Yu, K. and Tian, M. (2025). Mode-based classifier: A robust and flexible discriminant analysis for high-dimensional data. *Statistica Sinica*, 35, 1391–1422.

Zeng, J., Zhang, X. and Mai, Q. (2023). An efficient convex formulation for reduced-rank linear discriminant analysis in high dimensions. *Statistica Sinica*, 33, 1249–1270.

Zhang, X. and Li, L. (2017). Tensor envelope partial least-squares regression. *Technometrics*, 59, 426–436.

Zhu, Z. and Zhou, W. (2021). Taming heavy-tailed features by shrinkage. *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 130, 3268–3276.

Zuo, Y., Shen, W., Wang, C., Niu, N. and Pu, J. (2020). Circular RNA Circ-ZNF609 promotes lung adenocarcinoma proliferation by modulating miR-1224-3p/ETV1 signaling. *Cancer Management and Research*, 12, 2471–2479.

Downloaded from https://academic.oup.com/biometrics/article/82/1/ujag039/8498964 by guest on 27 February 2026

Received: 29 March 2025. **Revised:** 26 November 2025. **Accepted:** 28 January 2026

© The Author(s) 2026. Published by Oxford University Press on behalf of The International Biometric Society. All rights reserved. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site-for further information please contact journals.permissions@oup.com