Communications for Statistical Applications and Methods
2024, Vol. 31, No. 4, 393–408

https://doi.org/10.29220/CSAM.2024.31.4.393
Print ISSN 2287-7843 / Online ISSN 2383-4757

Concave penalized linear discriminant analysis on high dimensions

Sunghoon Kwon[a], Hyebin Kim[a], Dongha Kim[b], Sangin Lee[1,c]

[a]Department of Applied Statistics, Konkuk University, Korea;
[b]Department of Statistics, Sungshin Women's University, Korea;
[c]Department of Information and Statistics, Chungnam National University, Korea

Abstract

The sparse linear discriminant analysis can be incorporated into the penalized linear regression framework, but most studies have been limited to specific convex penalties, including the least absolute selection and shrinkage operator and its variants. Within this framework, concave penalties can serve as natural counterparts of the convex penalties. Implementing the concave penalized direction vector of discrimination appears to be straightforward, but developing its theoretical properties remains challenging. In this paper, we explore a class of concave penalties that covers the smoothly clipped absolute deviation and minimax concave penalties as examples. We prove that employing concave penalties guarantees an oracle property uniformly within this penalty class, even for high-dimensional samples. Here, the oracle property implies that an ideal direction vector of discrimination can be exactly recovered through concave penalized least squares estimation. Numerical studies confirm that the theoretical results hold with finite samples.

Keywords: concave penalties, linear discriminant analysis, direction vector, oracle property, high dimension

1. Introduction

The Linear Discriminant Analysis (LDA) stands out as a popular technique for classification tasks. However, its applicability falters when the sample is high-dimensional; that is, the number of input variables p is larger than the sample size n, which is a situation that has become increasingly prevalent in practical scenarios. In these cases, the classical LDA faces a critical limitation in estimating the population covariance matrix due to the singularity of the pooled sample covariance matrix. For example, Bickel and Levina (2004) pointed out that the performance of the classical LDA can be as poor as a random guess for a high-dimensional sample.

Many authors have studied modifications of the classical LDA concerning high-dimensional sample, with early proposals based on the independence rule and variable selection. Bickel and Levina (2004) proposed to use the independence rule or the diagonal LDA paradigm for estimating the population covariance matrix by ignoring the correlation structure of the input variables. Tibshirani *et al.* (2002) and Fan and Fan (2008) introduced the nearest shrunken centroid estimation and the features annealed independent rule, in which the variable selection is done by soft and hard thresholding rules,

This paper was supported by Konkuk University in 2021.

[1] Corresponding author: Department of Information and Statistics, Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon 34134, Korea. E-mail: sanginlee44@gmail.com

Published 31 July 2024 / journal homepage: http://csam.or.kr

respectively. However, these methods do not achieve Bayes classification error for the classification task unless the true common covariance matrix is diagonal.

Recently, the sparse LDA approaches have been proposed as an alternative to the independence rule. Cai and Liu (2011) proposed the linear programming discriminant rule that finds a sparse estimate of the population covariance matrix by using the Dantzig selector (James *et al.*, 2009), and Fan *et al.* (2012) proposed the regularized optimal affine discriminant method, which is based on the ℓ_1-Fisher's discriminant analysis. Similar ℓ_1-penalized linear discriminant analysis was studied in Trendafilov and Jolliffe (2007) and Witten and Tibshirani (2011). Clemmensen *et al.* (2011) proposed the sparse optimal scoring method which solves the optimal scoring formulation (Hastie *et al.*, 1994) with the ℓ_1-penalty. The Direct Sparse Discriminant Analysis (DSDA) studied in Mai *et al.* (2012) is another popular sparse LDA, which is computationally efficient and easier to understand, because it reformulates the high-dimensional LDA into a penalized linear regression framework. In general, the sparse LDA approaches have theoretical advantages over other methods of independence rules since they allow non-diagonal population covariance matrices, enabling the resulting classifier to achieve Bayes classification error. However, most previous works have been limited to specific convex penalties such as the Least Absolute Selection and Shrinkage Operator (LASSO) and its variants.

In this study, we focus on the DSDA applied with a class of concave penalties, including the smoothly clipped absolute deviation (Fan and Li, 2001), minimax concave (Zhang, 2010), and truncate d-ℓ_1 (Shen *et al.*, 2013) penalties as examples. Within this framework, implementing the concave penalized direction vector of discrimination appears to be straightforward, but developing the theoretical properties of the estimated direction vector still remains challenging. For this purpose, we prove that the concave penalized direction vector satisfies an oracle property uniformly in the class, which is the main contribution of the paper. Here, the oracle property implies that an ideal LDA direction vector of discrimination can be exactly recovered through concave penalized least square estimation. In addition, we provide various numerical studies to confirm whether the theoretical results hold with finite samples.

The rest of the paper is organized as follows. Section 2 introduces the sparse LDA. Section 3 introduces the concave penalized LDA and presents the related theoretical results. Section 4 provides numerical studies to confirm the theoretical results, and concluding remarks are given in Section 5. Technical details and proofs are provided in Appendix.

2. Sparse linear discriminant analysis

2.1. Fisher's linear discriminant analysis

Fisher's Linear Discriminant Analysis (LDA) (Fisher, 1936) is an efficient technique for discriminating a binary class label $C \in \{0, 1\}$ given an input vector $X \in \mathbb{R}^p$. The LDA assumes that X is a random vector with multivariate normal distribution given a binary random variable C such that

$$X \mid C = c \sim N(\mu_c, \Sigma_c), \quad c \in \{0, 1\}, \tag{2.1}$$

independently, where μ_c and Σ_c are the mean vector and covariance matrix of the normal distribution, respectively. Let $\phi(\cdot; \mu_c, \Sigma_c), c \in \{0, 1\}$ be the density function of $N(\mu_c, \Sigma_c)$ and $\pi_c = \mathbf{P}(C = c)$. The ratio between two conditional density values of $C|X = x$ provides a natural rule for discriminating the class label as $C = 1$ given $X = x$ as follows.

$$\frac{\pi_1 \phi(x; \mu_1, \Sigma_1)}{\pi_0 \phi(x; \mu_0, \Sigma_0)} \geq 1, \quad x \in \mathbb{R}^p. \tag{2.2}$$

In addition, when the covariance matrices in (2.1) are assumed to be common, $\Sigma_0 = \Sigma_1 = \Sigma$, the discrimination rule in (2.2) can be simplified into a linear form,

$$\psi(x) = \{x - (\mu_0 + \mu_1)/2\}^T \beta^{\mathrm{BS}} + \log(\pi_1/\pi_0) \geq 0, \quad x \in \mathbb{R}^p, \tag{2.3}$$

where $\beta^{\mathrm{BS}} = \Sigma^{-1}(\mu_1 - \mu_0)$ is Bayes direction vector.

Let $(x_i, c_i), i \leq n$ be n random samples of (X, C) then the LDA estimate $\hat{\psi}$ of ψ in (2.3) can be obtained by using simple moment type estimates:

$$\hat{\psi}^{\mathrm{LDA}}(x) = \{x - (\hat{\mu}_0 + \hat{\mu}_1)/2\}^T \hat{\beta}^{\mathrm{LDA}} + \log(\hat{\pi}_1/\hat{\pi}_0) \geq 0, \quad x \in \mathbb{R}^p, \tag{2.4}$$

where $\hat{\beta}^{\mathrm{LDA}} = \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)$ is the LDA direction vector,

$$\begin{aligned} \hat{\mu}_c &= \sum_{c_i = c} x_i/n_c, \\ \hat{\Sigma}_c &= \sum_{c_i = c} (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^T/(n_c - 1), \\ \hat{\Sigma} &= \sum_{c \in \{0,1\}} (n_c - 1)\hat{\Sigma}_c/(n - 2), \\ \hat{\pi}_c &= n_c/n, \end{aligned} \tag{2.5}$$

and n_c is the number of samples with $c_i = c$.

2.2. Sparse linear discriminant analysis

In many real fields, application of the LDA raises two challenging problems because of sparsity of the model and high-dimensionality of the samples. The sparsity of the model implies that Bayes direction vector satisfies $\beta_j^{\mathrm{BS}} = 0$ for some $j \leq p$ but, in general, the LDA estimate in (2.4) produces $\hat{\beta}_j^{\mathrm{LDA}} \neq 0$ for all $j \leq p$. The high-dimensionality of the samples implies $p > n$, where the pooled sample covariance matrix $\hat{\Sigma}$ in (2.5) becomes singular. In both cases, the discriminant analysis may produce unexpected bad results in prediction as well as interpretation. There are many literatures that figure out these problems and the penalized approach via the least squares estimation can be a nice solution (Mai *et al.*, 2012).

Let $y_i = (-1)^{1+c_i} n/n_{c_i}, i \leq n, c_i \in \{0, 1\}$ then the LDA direction vector in (2.4) can be interpreted as a Least Squares Estimator (LSE) as follows:

$$\hat{\beta}^{\mathrm{LSE}} = k\hat{\beta}^{\mathrm{LDA}} \tag{2.6}$$

for some positive constant k unless $\hat{\mu}_0 \neq \hat{\mu}_1$ (Hastie *et al.*, 2009), where

$$\left(\hat{\alpha}^{\mathrm{LSE}}, \hat{\beta}^{\mathrm{LSE}}\right) = \arg\min_{\alpha, \beta} \sum_{i=1}^{n} \left(y_i - \alpha - x_i^T \beta\right)^2 /2n. \tag{2.7}$$

Hence the LDA rule in (2.4) can be cast into the LSE rule:

$$\hat{\psi}^{\mathrm{LSE}}(x) = \{x - (\hat{\mu}_0 + \hat{\mu}_1)/2\}^T \hat{\beta}^{\mathrm{LSE}} + k\log(n_1/n_0) \geq 0, \quad x \in \mathbb{R}^p. \tag{2.8}$$

Given the sparsity and high-dimensionality, it is natural to employ the penalized LSE:

$$\left(\hat{\alpha}^{\lambda},\hat{\beta}^{\lambda}\right) = \arg\min_{\alpha,\beta}\left\{\sum_{i=1}^{n}\left(y_i - \alpha - x_i^T\beta\right)^2/2n + \sum_{j=1}^{p} J_{\lambda}\left(|\beta_j|\right)\right\},$$

where J_λ is a penalty equipped with a tuning parameter λ. In this case, the LSE rule in (2.8) can be replaced with the penalized LSE rule again:

$$\hat{\psi}^{\lambda}(x) = \{x - (\hat{\mu}_1 + \hat{\mu}_0)/2\}^T\hat{\beta}^{\lambda} + k_\lambda \log(n_1/n_0) \geq 0, \tag{2.9}$$

where $k_\lambda = \hat{\beta}^{\lambda^T}\hat{\Sigma}\hat{\beta}^{\lambda}/(\hat{\mu}_1 - \hat{\mu}_0)^T\hat{\beta}^{\lambda}$ is given by Mai *et al.* (2012) that is optimal to the discrimination whenever $(\hat{\mu}_1 - \hat{\mu}_0)^T\hat{\beta}^{\lambda} > 0$.

3. Concave penalized linear discriminant analysis

3.1. Concave penalized estimation

Let J_λ be a penalty function with tuning parameter $\lambda > 0$ and J'_λ be the first derivative of J_λ. We assume that following conditions hold.

(J1) $J_\lambda(t)$ is concave and non-decreasing over $t \in [0,\infty)$ and $J_\lambda(0) = 0$.

(J2) $J'_\lambda(t)$ is non-increasing and continuous over $t \in (0,\infty)$ and $\lim_{t\to 0+} J'_\lambda(t) = \lambda$.

(J3) $J'_\lambda(t) \geq \lambda - t/a$ over $t \in (0, a\lambda)$ and $J'_\lambda(t) = 0$ over $t \in (a\lambda, \infty)$ for some $a > 0$.

The class of penalties that satisfy (J1), (J2), and (J3) has been studied as a representative concave penalty class (Kim and Kwon, 2012; Zhang and Zhang, 2012), including the Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li, 2001),

$$J'_\lambda(t) = \min\{\lambda, (a\lambda - t)_+/(a-1)\}, \quad a > 2,$$

minimax concave (MCP) (Zhang, 2010),

$$J'_\lambda(t) = (\lambda - t/a)_+, \quad a > 1,$$

and truncated-ℓ_1 penalties (TLP) (Shen *et al.*, 2013)

$$J_\lambda(t) = \lambda\min\{t, a\}, \quad a > 1,$$

as examples, where $x_+ = \max\{0, x\}$.

Let $Z = (Z_1, \ldots, Z_p)$ be the centered design matrix in (2.7) then the LSE direction vector in (2.7) can be simplified into

$$\hat{\beta}^{\text{LSE}} = \arg\min_\beta\left\{\|Z\beta\|_2^2/2n - (\hat{\mu}_1 - \hat{\mu}_0)^T\beta\right\}.$$

Now, considering the true signal set $\mathcal{S} = \{j|\beta_j^{\text{BS}} \neq 0\}$, one of the best estimators for the sparse Bayes direction vector is the oracle LSE,

$$\hat{\beta}^{\text{OR}} = \arg\min_{\beta_j=0, j\in\mathcal{S}^c}\left\{\|Z\beta\|_2^2/2n - (\hat{\mu}_1 - \hat{\mu}_0)^T\beta\right\}, \tag{3.1}$$

which is unavailable in practice without the knowledge of $\mathcal{S}$. Hence, the main objective of the concave penalized LDA is to recover $\hat{\beta}^{\rm OR}$ through the penalized LSE:

$$\hat{\beta}^{\lambda} = \arg\min_{\beta} L_{\lambda}(\beta), \tag{3.2}$$

where

$$L_{\lambda}(\beta) = \|Z\beta\|_2^2/2n - (\hat{\mu}_1 - \hat{\mu}_0)^T\beta + \sum_{j=1}^{p} J_{\lambda}\left(|\beta_j|\right).$$

3.2. Optimality conditions

In this subsection, we introduce three lemmas that provide non-asymptotic sufficient conditions for a given estimator to be a local or unique local minimizer of L_λ. Let Ξ_λ be the set of all local minimizers of L_λ and $\hat{\theta} = \hat{\mu}_1 - \hat{\mu}_0$.

Lemma 1. *If $\hat{\beta} \in \mathbb{R}^p$ satisfies $\min_{\hat{\beta}_j \neq 0} |\hat{\beta}_j| > a\lambda$ and $\max_{\hat{\beta}_j = 0} |Z_j^T Z\hat{\beta}/n - \hat{\theta}_j| \leq \lambda$ then $\hat{\beta} \in \Xi_\lambda$.*

The conditions in Lemma 1 are simply the sub-gradient optimality conditions for the penalty class that satisfy penalty conditions (J1), (J2), and (J3), under which a given estimator becomes a local minimizer (Kim *et al.*, 2008; Kwon *et al.*, 2021) of L_λ. We say that $\hat{\beta} \in \Xi_\lambda$ satisfies the uniqueness condition (Kim and Kwon, 2012) with $\rho > 0$ if

$$\max_{\hat{\beta}_j=0} \left|Z_j^T Z\hat{\beta}/n - \hat{\theta}_j\right| < \lambda \min\{1, a\rho\} \leq \lambda(1 + a\rho) < \rho \min_{\hat{\beta}_j \neq 0} \left|\hat{\beta}_j\right|. \tag{3.3}$$

The following lemma states that the uniqueness condition forms a sufficient condition for a local minimizer to be unique. Let $\rho_{\min}(A)$ be the minimum eigenvalue of a square matrix A.

Lemma 2. *If $\hat{\beta} \in \Xi_\lambda$ satisfies the uniqueness condition with $\rho = \rho_{\min}(Z^TZ/n) > 0$ then $\hat{\beta}$ is a unique local minimizer of L_λ, that is, $\Xi_\lambda = \{\hat{\beta}\}$.*

Lemma 2 requires Z^TZ/n is non-singular which is impossible to hold for high-dimensional samples with $p > n$. However, we can extend the result to the cases by assuming the true model is sparse enough to estimate. For any subset $\mathcal{D} \subset \{1, \ldots, p\}$, let $Z_{\mathcal{D}}$ be a sub-design matrix of Z constructed by the columns $Z_j, j \in \mathcal{D}$. We say that Z satisfies the Sparse Riesz condition (Zhang, 2010) with $\rho > 0$ and rank $r > 0$ if

$$\min_{|\mathcal{D}| \leq r} \rho_{\min}\left(Z_{\mathcal{D}}^T Z_{\mathcal{D}}/n\right) > \rho. \tag{3.4}$$

Let $\Xi_\lambda^\kappa = \{\hat{\beta} \in \Xi_\lambda : \|\hat{\beta}\|_0 \leq \kappa\} \subseteq \Xi_\lambda$ be a restricted set of local minimizers of L_λ for some $\kappa \leq p$, where $\|a\|_0$ denotes the number of nonzero elements of a vector a. The following lemma extends the result of Lemma 2 to high-dimensional cases.

Lemma 3. *If Z satisfies the Sparse Riesz condition with $\rho > 0$ and rank $r \geq 2\kappa$ and $\hat{\beta} \in \Xi_\lambda$ satisfies the uniqueness condition with ρ then $\Xi_\lambda^\kappa = \{\hat{\beta}\}$.*

Remark 1. (a) The uniqueness and Sparse Riesz conditions in (3.3) and (3.4) are modified versions for the penalized LDA from those for the penalized linear regression (Kim and Kwon, 2012; Zhang, 2010), respectively. Lemma 1,2, and 3 are direct applications of Theorem 2 and 3 in Kim and Kwon (2012). (b) In Lemma 3, $\kappa \leq p$ represents the maximum number of candidate input variables that can be included in the final model, expecting that the corresponding local minimizer is unique in Ξ_λ^κ, which is impossible for Ξ_λ in general. Hence, it is reasonable to assume that $2\kappa \leq r \leq n$ since κ should be assumed or expected to be significantly less than the sample size.

3.3. Oracle property

In this subsection, we present the main results of the paper: The concave penalized LSE in (3.2) is unique, and hence, the global minimizer of L_λ, and the same as the oracle LSE in (3.1) asymptotically, in the sense that $\Xi_\lambda^\kappa = \{\hat{\beta}^\lambda\}$ and $\hat{\beta}^\lambda = \hat{\beta}^{\mathrm{OR}}$ with probability tending to 1.

3.3.1. Notations

We introduce some notations. For any matrix A, let $\|A\|_\infty = \max_i \sum_j |A_{ij}|$, where A_{ij} is the (i, j) entry of A, and let $A_{\mathcal{D}\mathcal{D}'}$ be a sub-matrix of A constructed by the entry $A_{ij}, i \in \mathcal{D}, j \in \mathcal{D}'$. For any vector a, let $\|a\|_\infty = \max_i |a_i|$ and $\|a\|_1 = \sum_i |a_i|$, where a_i is the ith element of a, and let $a_{\mathcal{D}}$ be a sub-vector of a constructed by the elements $a_i, i \in \mathcal{D}$. Let $|\mathcal{A}|$ be the cardinality of a set $\mathcal{A}$. For any positive sequences x_n and y_n, we write $x_n \gg y_n$ if $x_n/y_n \to \infty$ and $x_n \asymp y_n$ if $x_n/y_n \to a$ for some constant a as $n \to \infty$.

Recall the definition of the oracle LSE $\hat{\beta}^{\mathrm{OR}}$ obtained under the knowledge of the true signal set $\mathcal{S} = \{j|\beta_j^{\mathrm{BS}} \neq 0\}$. It is easy to see that

$$\hat{\beta}_{\mathcal{S}}^{\mathrm{OR}} = \hat{\Omega}_{\mathcal{S}\mathcal{S}}^{-1}\hat{\theta}_{\mathcal{S}} \quad \text{and} \quad \hat{\beta}_{\mathcal{N}}^{\mathrm{OR}} = 0,$$

where $\hat{\Omega} = Z^T Z/n$, $\hat{\theta} = \hat{\mu}_1 - \hat{\mu}_0$, and $\mathcal{N} = \{j|\beta_j^{\mathrm{BS}} = 0\}$ is the true noisy set. Hence, we can construct a population counterpart β^{OR} of $\hat{\beta}^{\mathrm{OR}}$ by defining

$$\beta_{\mathcal{S}}^{\mathrm{OR}} = \Omega_{\mathcal{S}\mathcal{S}}^{-1}\theta_{\mathcal{S}} \quad \text{and} \quad \beta_{\mathcal{N}}^{\mathrm{OR}} = 0,$$

where $\Omega = \mathrm{Cov}(X)$ and $\theta = \mu_1 - \mu_0$. Note that β^{OR} is exclusively considered for the development of theoretical properties since $\beta^{\mathrm{OR}} = k\beta^{\mathrm{BS}}$ for some constant $k \neq 0$. For further details, refer to Proposition 3 in Mai *et al.* (2012).

3.3.2. Oracle property

The main objective of the concave penalized LSE is to recover $\hat{\beta}^{\mathrm{OR}}$ by using $\hat{\beta}^\lambda$ in (3.2), which is often referred to the strong oracle property (Fan and Li, 2001; Kim *et al.*, 2016). From the lemmas in previous subsection, it is sufficient to prove that $\hat{\beta}^{\mathrm{OR}}$ satisfies the outlined uniqueness and Sparse Riesz conditions with probability tending to 1 of which the proofs are provided in Appendix. We assume the following regularity conditions:

(C1) There exist positive constants, $b_i, i \leq 4$, such that

$$\rho_{\min}(\Omega) > b_1, \quad \left\|\Omega_{\mathcal{N}\mathcal{S}}\Omega_{\mathcal{S}\mathcal{S}}^{-1}\right\|_\infty < b_2, \quad \left\|\Omega_{\mathcal{S}\mathcal{S}}^{-1}\right\|_\infty < b_3, \quad \text{and} \quad \|\theta_{\mathcal{S}}\|_\infty < b_4,$$

for any n.

Remark 2. The conditions in (C1) specify technical requirements for the oracle property that are slightly weaker than those applied with the Least Absolute Selection and Shrinkage Operator (LASSO) in Mai *et al.* (2012). For the LASSO, we need $b_1 = 1$ that plays as the strong irrepresentable condition in Zhao and Yu (2006) for linear regression.

Theorem 1. *Assume that (C1) holds. The oracle LSE is unique local, and hence, the global minimizer of L_λ with probability tending to one, in the sense that*

$$\lim_{n\to\infty} \mathbf{P}\left(\Xi_\lambda^\kappa = \left\{\hat{\beta}^{\mathrm{OR}}\right\}\right) = 1,$$

Table 1: Estimated probabilities of including correct model

r	p	q	n	Oracle property			Sign consistency			
				TLP	MCP	SCAD	LASSO	TLP	MCP	SCAD
0.25	300	10	500	0.38	0.37	0	0.53	0.38	0.38	0.14
			1000	0.95	0.97	0.13	0.99	0.96	0.97	0.78
			2000	1	0.98	0.79	1	1	1	0.99
		20	500	0	0	0	0	0	0	0
			1000	0.11	0.11	0	0.13	0.11	0.11	0.03
			2000	0.87	0.87	0.01	0.83	0.87	0.87	0.43
	3000	10	500	0.05	0.07	0	0.10	0.05	0.08	0.02
			1000	0.89	0.89	0	0.90	0.89	0.90	0.48
			2000	1	0.98	0.33	1	1	1	0.99
		20	500	0	0	0	0	0	0	0
			1000	0	0	0	0.01	0	0	0
			2000	0.50	0.50	0	0.50	0.50	0.51	0.09
0.5	300	10	500	0.05	0.11	0	0	0.06	0.11	0
			1000	0.72	0.76	0.01	0.30	0.72	0.76	0.18
			2000	1	1	0.53	0.88	1	1	0.86
		20	500	0	0	0	0	0	0	0
			1000	0	0.01	0	0	0	0.01	0
			2000	0.38	0.38	0	0.04	0.38	0.38	0.01
	3000	10	500	0	0	0	0	0	0	0
			1000	0.32	0.37	0	0.03	0.32	0.37	0
			2000	0.96	0.97	0.14	0.71	0.96	0.97	0.36
		20	500	0	0	0	0	0	0	0
			1000	0	0	0	0	0	0	0
			2000	0.08	0.08	0	0	0.08	0.08	0

for any sequence $\kappa = \kappa(n)$ that satisfies $2q \le \kappa \le n$, provided that

$$\lambda = o(m_{\mathcal{S}}), \quad \log p = o\left(\min\left\{n/\kappa^3, n\lambda^2/q^2\right\}\right) \quad \textit{and} \quad \min\left\{n/\kappa^2, n\lambda^2/q^2\right\} \to \infty,$$

as $n \to \infty$, where $q = |\mathcal{S}|$ and $m_{\mathcal{S}} = \min_{j\in\mathcal{S}} |\beta_j^{\mathrm{OR}}|$.

Remark 3. If $n/\kappa^3 \asymp n\lambda^2/q^2 \to \infty$ then conditions in Theorem 1 can be simplified as

$$m_{\mathcal{S}} \gg \lambda \gg q\sqrt{\log p/n} \quad \text{and} \quad \log p = o\left(n\lambda^2/q^2\right)$$

as $n \to \infty$. Therefore, we can observe the following theoretical properties:

(a) The penalized LDA allows for exponentially many input variables, polynomially many signal variables, and diminishing regression coefficients, satisfying that $p = o(\exp(n\lambda^2/q^2))$, $q = o(n^{1/3})$, and $m_{\mathcal{S}} \to 0$ at a slower rate than $q\sqrt{\log p/n}$.

(b) Compared with the ordinary penalized linear regression (Zhang, 2010), the requirements are stronger since $p = o(\exp(n\lambda^2))$, $q = o(n)$, and $m_{\mathcal{S}} \to 0$ at a slower rate than $\sqrt{\log p/n}$.

(c) The stronger requirements mainly comes from the random design matrix Z in (2.7) and the random Sparse Riesz condition in (3.4). We can check similar results in many literatures, for example, $p = o(n^{1/2})$ for the autoregressive model (Na and Kwon, 2018) and $q = o(n^{1/5})$ for general maximum likelihood estimation (Fan and Peng, 2004; Kwon and Kim, 2012).

4. Numerical studies

In this section, we report some results of simulation studies and real data analysis.

4.1. Simulation studies

We generated n random samples of C from $\mathrm{Ber}(\pi)$ with $\pi = 1/2$, and then generated n random samples of $X|C = c, c \in \{0, 1\}$ from $N(\mu_c, \Sigma)$. We set $\mu_0 = 0$ and $\mu_1 = \Sigma\beta^{\mathrm{BS}}$, where $\Sigma_{jk} = r^{|j-k|}$ and $\beta_j^{\mathrm{BS}} = (2/\sqrt{j})(-1)^j I(j \leq q), j, k \leq p$. We consider three concave penalties, TLP with $a = 0.001$, MCP with $a = 1.001$, and SCAD with $a = 2.001$, as examples in the penalty class, and set $n \in \{500, 1000, 2000\}$, $p \in \{300, 3000\}$, and $q \in \{10, 20\}$ with $r \in \{0.25, 0.5\}$ for each simulation. We repeated the simulation 200 times by using R package ncpen (Kim *et al.*, 2020).

We first investigated whether the oracle property can hold with finite samples and Table 1 shows the estimated probabilities of achieving the oracle property. For each simulation, we first found the interval $[\lambda_{\max}, \lambda_{\min}]$ that satisfies $\|\hat{\beta}^{\lambda_{\max}}\|_0 = 0$ and $\|\hat{\beta}^{\lambda_{\min}}\|_0 = 50$. Then we checked whether there exists a λ that satisfies $\hat{\beta}^{\lambda} = \hat{\beta}^{\mathrm{OR}}$ by investigating 200 values of λ decreasing with log-scale in the interval. From the table, we observed the followings:

- The ratios of the TLP and MCP approaches nearly 1 for some cases, while the SCAD has lower ratios, with the largest ratio being 0.79. In cases where both p and q are large, it is rare for the oracle property to hold, but we believe that this occurs due to limitations in the simulation settings. In general, we can conclude that the ratio increases as n increases and p, q, and r decreases, regardless of the simulation settings and penalties, which confirms the result of Theorem 1.

- In addition, we checked the sign consistency, $\mathrm{sign}(\hat{\beta}^{\lambda}) = \mathrm{sign}(\hat{\beta}^{\mathrm{OR}})$ for some λ, which is a slightly less stringent condition than the oracle property. The results exhibit similar pattern to that of the oracle property but the ratios are slightly larger. We found that the LASSO also achieves the sign consistency as Mai *et al.* (2012) proved. In general, the sign consistency of the LASSO does not hold even for the ordinary linear regression since it requires the strong irrepresentale condition on the design matrix. See Section 3 and example 3 in Zhao and Yu (2006) for some details.

Second, we compared the finite sample performance of the concave penalized estimators, using the LASSO as a benchmark method. The primary objective of the simulation is to check whether the oracle property can be realized through typical tuning parameter selection criteria. Note that there are two natural tuning parameter selection criteria from the characteristic of the framework: Minimizing the regression error in (2.7) and the classification error in (2.9). We used n training samples for estimating the penalized direction vector and $n/2$ independent validation samples for selecting an optimal tuning parameter λ_{opt} by minimizing the two criteria. We calculated three measures of selection accuracy: The numbers of true and false positive selection (TPS and FPS) and the ratio of the correct model selection (CMS): $\mathrm{TPS} = \sum_j I(\hat{\beta}_j^{\lambda_{\mathrm{opt}}} \neq 0, \beta_j^{\mathrm{BS}} \neq 0)$, $\mathrm{FPS} = \sum_j I(\hat{\beta}_j^{\lambda_{\mathrm{opt}}} \neq 0, \beta_j^{\mathrm{BS}} = 0)$, and CMS is the ratio of the cases when $\mathrm{TPS} = q$ and $\mathrm{FPS} = p-q$. In addition, we calculated the miss-classification error rate (ERR) by using $2n$ independent test samples.

Tables 2 and 3 summarizes the results. We first discuss the cases when λ_{opt} was chosen based on the regression errors of the validation samples:

- TPS increases to q as n increases, regardless of the penalties and simulation settings. This suggests that the signal variables tend to be selected as the sample size increases. As p or r increases while keeping n fixed, TPS decreases. This implies that higher model complexity and stronger correlation

Table 2: Averages of the four measures: Validation by regression errors

r	p	q	n	TPS				FPS			
				LASSO	TLP	MCP	SCAD	LASSO	TLP	MCP	SCAD
0.25	300	10	500	9.97	9.31	9.33	9.65	29.29	1.15	1.04	14.01
			1000	10.00	9.99	9.99	10.00	30.48	0.66	0.51	8.08
			2000	10.00	10.00	10.00	10.00	29.40	0.48	0.37	4.92
		20	500	18.66	13.44	13.42	15.62	29.74	2.36	2.18	12.64
			1000	19.94	18.94	18.98	19.05	30.02	2.50	2.34	10.43
			2000	19.99	19.94	19.94	19.98	30.26	0.72	0.51	8.46
	3000	10	500	9.74	7.69	7.80	8.81	36.88	1.12	0.96	18.79
			1000	10.00	9.91	9.91	9.96	38.56	0.57	0.44	17.28
			2000	10.00	10.00	10.00	10.00	37.88	0.47	0.30	6.08
		20	500	15.64	8.55	8.63	11.67	34.84	1.31	1.15	17.91
			1000	19.02	15.37	15.32	16.55	32.12	2.21	1.75	13.99
			2000	19.97	19.64	19.64	19.52	31.38	1.21	0.98	10.87
0.5	300	10	500	9.74	8.20	8.31	8.86	37.05	2.06	1.80	13.90
			1000	9.99	9.99	9.97	10.00	38.49	1.02	0.90	10.03
			2000	10.00	10.00	10.00	10.00	37.44	0.42	0.35	4.27
		20	500	16.12	10.21	10.30	11.83	33.76	2.62	2.62	13.12
			1000	19.42	17.31	17.25	16.71	31.20	4.72	4.05	12.86
			2000	19.97	19.71	19.74	19.51	30.68	1.74	1.60	9.65
	3000	10	500	7.94	5.87	5.81	6.47	41.96	1.15	0.94	19.51
			1000	9.78	9.17	9.16	9.11	41.51	1.01	0.88	20.05
			2000	10.00	9.99	9.99	10.00	41.88	0.39	0.31	8.20
		20	500	9.88	5.93	5.77	7.19	39.89	1.32	1.10	18.78
			1000	14.72	10.76	10.78	11.15	36.70	2.00	1.94	18.46
			2000	18.98	17.82	17.82	15.88	32.21	2.56	2.60	14.61
r	p	q	n	CMS				ERR			
				LASSO	TLP	MCP	SCAD	LASSO	TLP	MCP	SCAD
0.25	300	10	500	0	0.28	0.29	0	0.093	0.092	0.092	0.093
			1000	0	0.57	0.66	0.04	0.088	0.085	0.085	0.085
			2000	0	0.71	0.79	0.34	0.086	0.084	0.084	0.084
		20	500	0	0	0	0	0.082	0.088	0.088	0.089
			1000	0	0.06	0.07	0	0.072	0.071	0.071	0.075
			2000	0	0.53	0.67	0	0.068	0.066	0.066	0.067
	3000	10	500	0	0.03	0.05	0	0.103	0.105	0.103	0.109
			1000	0	0.65	0.69	0	0.091	0.086	0.086	0.090
			2000	0	0.66	0.80	0.14	0.087	0.084	0.084	0.084
		20	500	0	0	0	0	0.097	0.104	0.103	0.107
			1000	0	0	0	0	0.080	0.079	0.078	0.087
			2000	0	0.35	0.42	0	0.072	0.067	0.067	0.073
0.5	300	10	500	0	0.04	0.08	0	0.173	0.168	0.167	0.166
			1000	0	0.49	0.53	0	0.159	0.150	0.150	0.149
			2000	0	0.72	0.77	0.28	0.152	0.148	0.148	0.148
		20	500	0	0	0	0	0.167	0.168	0.167	0.165
			1000	0	0	0.01	0	0.147	0.141	0.141	0.148
			2000	0	0.28	0.32	0	0.135	0.130	0.130	0.132
	3000	10	500	0	0	0	0	0.202	0.187	0.185	0.196
			1000	0	0.25	0.32	0	0.171	0.156	0.155	0.164
			2000	0	0.67	0.76	0.05	0.157	0.148	0.148	0.148
		20	500	0	0	0	0	0.203	0.188	0.186	0.196
			1000	0	0	0	0	0.171	0.159	0.158	0.166
			2000	0	0.07	0.07	0	0.149	0.136	0.136	0.147

among the input variables have a negative effect on selection of signal variables. Among the concave penalties, the SCAD produces the largest TPS but the difference does not seem significantly large.

- For the concave penalties, FPS decreases as n increases for almost all cases. This implies that the noisy variables tend to be excluded as the sample size increases, which is not true for the LASSO. FPS of the SCAD is smaller than that of the LASSO but it is not at a negligible level even when $n = 2000$. However, FPS for the TLP and MCP are near 0 when $n = 2000$, implying that they excluded nearly all noisy variables.

- For the concave penalties, CMS increases as n increases regardless of the penalties and simulation settings, achieving the largest score 0.80, 0.77, and 0.34 for the TLP, MCP, and SCAD, respectively. For the LASSO, CMS is consistently 0, indicating that the regression errors does not work for the LASSO to select correct models.

- One of the reasons for the poor FPS and CMS of the SCAD seems to be because its shape or concavity on the interval $[0, a\lambda]$. Since the SCAD penalty is the same as the LASSO on $[0, \lambda]$ and $a > 2$ is limited below, the maximum concavity on $[0, a\lambda]$ is smaller than that of the MCP with $a > 1$, which results in selecting more variables with higher FPS and smaller CMS than those of the MCP. We refer to Zhang (2010) for some detailed discussion on the maximum concavity of the concave penalties.

- ERR decreases to that of the Bayes, the best performance, as n increases but increases as p or r increases with fixed n, regardless of the penalties and simulation settings.

For the cases of classification errors, we observed the followings:

- TPS shows similar patterns as n increases, resembling the case of regression errors. This occurs regardless of the penalties and simulation settings.

- The LASSO and SCAD select significantly less noisy variables, producing lower FPS and slightly higher CMS. However, the TLP and MCP show opposite results.

To sum up, (a) we can conclude that tuning parameter selection based on the regression errors exhibits better selection performance for the TLP and MCP. However, for the LASSO and SCAD, the classification errors seems to be more informative, as indicated by smaller FPS and larger CMS. (b) The two tuning parameter selection criteria can function to some extent in the simulation settings designed in this paper. (c) However, the probabilities of correctly identifying the true model in Table 1 consistently reach nearly 1, regardless of the penalties and simulation settings. This suggests that we need to develop other alternatives such as the information criterion based on the underling probability structure.

4.2. Real data analysis

We apply the penalized LDA methods to two benchmark datasets: The prostate and lung cancer datasets, which is available from the R packages `datamicroarray`. The prostate cancer dataset consists of the expression levels of approximately 12,600 genes obtained from 52 prostate tumour samples and 50 non-tumour prostate samples (Singh *et al.*, 2002). Likewise, the lung cancer dataset comprises 12,533 genes from 150 patients with adenocarcinoma and 31 patients with malignant pleural mesothelioma (Gordon *et al.*, 2002). The main task is to predict whether an observation is the specific tumor tissue or not, and to identify the genes associated with each type of cancer. We first

Table 3: Averages of the four measures: Validation by classification errors

r	p	q	n	TPS				FPS			
				LASSO	TLP	MCP	SCAD	LASSO	TLP	MCP	SCAD
0.25	300	10	500	9.80	8.79	8.79	9.30	14.67	1.87	1.84	8.96
			1000	10.00	9.82	9.79	9.95	14.08	2.44	2.39	6.28
			2000	10.00	9.95	9.93	9.99	14.54	3.31	2.91	5.10
		20	500	17.41	12.57	12.62	13.72	14.54	2.65	2.86	6.95
			1000	19.54	17.37	17.43	17.90	15.52	2.31	2.28	5.80
			2000	19.98	19.50	19.53	19.86	14.33	1.98	2.01	5.63
	3000	10	500	9.36	7.36	7.51	8.08	15.34	1.37	1.23	10.29
			1000	10.00	9.74	9.72	9.80	18.61	1.13	1.31	10.81
			2000	10.00	9.98	9.96	9.99	15.93	1.94	1.96	5.40
		20	500	13.98	8.29	8.44	10.23	16.91	1.50	1.80	9.29
			1000	18.47	14.84	14.93	15.36	17.62	2.47	2.54	8.24
			2000	19.93	18.95	19.08	19.14	18.73	1.50	1.65	6.87
0.5	300	10	500	9.20	7.98	8.06	8.45	21.94	2.74	3.13	10.80
			1000	9.95	9.77	9.71	9.84	23.10	2.82	2.34	8.84
			2000	10.00	9.94	9.96	9.97	21.13	2.83	2.76	6.34
		20	500	14.69	10.14	10.63	11.00	22.63	3.32	4.23	10.12
			1000	18.99	15.96	15.98	15.19	23.22	4.24	4.15	8.63
			2000	19.87	19.27	19.19	19.05	21.89	2.97	2.93	7.27
	3000	10	500	7.12	5.60	5.66	5.91	23.50	1.62	1.54	12.70
			1000	9.51	8.97	8.97	8.43	24.92	1.97	2.12	12.08
			2000	9.99	9.86	9.85	9.93	25.88	1.20	1.75	8.47
		20	500	8.29	5.77	5.69	6.10	21.49	1.69	1.56	11.51
			1000	13.72	10.48	10.32	9.89	24.71	3.29	2.59	10.57
			2000	18.76	17.51	17.41	14.88	26.59	3.52	3.42	9.78
r	p	q	n	CMS				ERR			
				LASSO	TLP	MCP	SCAD	LASSO	TLP	MCP	SCAD
0.25	300	10	500	0.03	0.15	0.14	0.02	0.097	0.097	0.097	0.098
			1000	0.09	0.33	0.29	0.15	0.089	0.088	0.088	0.088
			2000	0.20	0.43	0.44	0.42	0.086	0.086	0.086	0.085
		20	500	0	0	0	0	0.086	0.092	0.091	0.095
			1000	0	0.04	0.04	0	0.075	0.074	0.074	0.078
			2000	0.03	0.29	0.23	0.07	0.069	0.067	0.067	0.068
	3000	10	500	0	0.05	0.05	0	0.107	0.108	0.106	0.113
			1000	0.11	0.52	0.41	0.02	0.092	0.088	0.088	0.092
			2000	0.07	0.43	0.43	0.24	0.087	0.086	0.086	0.085
		20	500	0	0	0	0	0.102	0.106	0.105	0.112
			1000	0	0	0	0	0.082	0.081	0.081	0.090
			2000	0.04	0.26	0.25	0	0.073	0.069	0.069	0.075
0.5	300	10	500	0	0.04	0.04	0	0.178	0.171	0.171	0.170
			1000	0	0.25	0.23	0.03	0.160	0.154	0.155	0.153
			2000	0	0.50	0.41	0.17	0.153	0.150	0.150	0.150
		20	500	0	0	0	0	0.171	0.170	0.170	0.169
			1000	0	0	0	0	0.149	0.145	0.145	0.150
			2000	0	0.14	0.11	0	0.137	0.132	0.132	0.134
	3000	10	500	0	0	0	0	0.207	0.190	0.189	0.200
			1000	0	0.19	0.21	0	0.174	0.160	0.160	0.168
			2000	0	0.54	0.47	0.08	0.158	0.150	0.150	0.150
		20	500	0	0	0	0	0.209	0.191	0.189	0.201
			1000	0	0	0	0	0.174	0.162	0.162	0.171
			2000	0	0.05	0.04	0	0.150	0.138	0.138	0.149

choose top $p \in \{500, 1000, 2000\}$ genes with the largest t-statistics across two classes. We then applied the penalized LDA methods with two classes as a response variable and the chosen top p genes

Table 4: Averages of the measures in prostate and lung cancer data analysis

Data	Measure	p	Cross-validation by regression error				Cross-validation by classification error			
			LASSO	TLP	MCP	SCAD	LASSO	TLP	MCP	SCAD
Prostate	ERR(%)	500	6.34	9.32	9.44	9.04	7.28	9.64	9.48	9.06
		1000	7.06	9.40	9.24	9.12	7.84	10.28	9.80	9.16
		2000	7.76	9.40	9.32	9.30	8.00	10.38	10.56	9.14
	#MIS	500	1.58	2.33	2.36	2.26	1.82	2.41	2.37	2.27
		1000	1.76	2.35	2.31	2.28	1.96	2.57	2.45	2.29
		2000	1.94	2.35	2.33	2.33	2.00	2.60	2.64	2.28
	SIZE	500	25.31	4.60	4.00	11.24	12.25	5.98	5.17	8.45
		1000	25.65	3.08	3.10	13.40	9.12	6.43	6.68	9.03
		2000	25.50	3.01	2.21	12.94	6.28	4.46	5.21	6.40
Lung	ERR(%)	500	0.94	1.69	1.66	1.64	1.57	2.03	2.09	1.84
		1000	0.93	1.84	1.86	1.59	1.56	2.02	2.23	1.79
		2000	0.94	1.80	2.00	1.50	1.54	2.20	2.26	1.91
	#MIS	500	0.43	0.76	0.75	0.74	0.71	0.92	0.94	0.83
		1000	0.42	0.83	0.84	0.72	0.70	0.91	1.01	0.81
		2000	0.43	0.81	0.90	0.68	0.70	0.99	1.02	0.86
	SIZE	500	58.57	21.36	18.03	21.60	12.54	11.31	9.76	9.39
		1000	60.05	17.73	17.57	25.31	12.95	11.23	10.26	9.85
		2000	60.81	17.45	18.09	27.88	12.51	11.95	11.80	10.10

as predictive variables. For comparison, we split the data into the training and test sets by randomly choosing 3/4 samples and 1/4 samples, respectively. For each training set, the tuning parameter λ is selected by the 10-fold cross-validation methods based on regression error and classification error as in the simulations. We compute the mis-classification error rate (ERR), the number of mis-classified samples (#MIS) on the test set, and the model size (SIZE) representing the number of selected variables on the training set.

Table 4 presents the average values of the three measures obtained from 200 random partitions of data. In most cases, the LASSO shows the best prediction performance but selects the most variables. The TLP, MCP, and SCAD show similar prediction performances and they substantially selects fewer variables than the LASSO. Among the concave penalized methods, the SCAD selects more variables than other methods. Similar to the simulation results, the methods based on regression error exhibit better prediction performances. For model size, the LASSO and SCAD based on the classification error produce more sparse models while the methods based on regression error produce more sparse model for TLP and MCP. These results suggest that the concave penalized LDA method could be a good alternative when we wish to construct the sparse model without losing the prediction accuracy much.

5. Concluding remarks

In this paper, we studied the high-dimensional LDA based on the concave penalized linear regression. We proved that an oracle property holds uniformly on a class of concave penalties, including the TLP, SCAD, and MCP as examples. The primary advantage of the concave penalized approach lies in its superior selection performance compared to the convex penalized approach, as supported by the simulation studies. In addition, we found that the tuning parameter selection criteria based on the prediction errors work to some extent, and hence, we may use the criteria in practice. However, we believe that there are better alternatives, such as the Bayesian information criterion, which has been proven to be useful for other penalized approaches, including the penalized linear regression (Fan and Tang, 2012).

Appendix

Let $\|A\|_2 = \sup_{\|u\|_2=1} \|Au\|_2$ denote the spectral norm of a matrix A. Let $\hat{\delta} = \hat{\theta} - \theta$, $\hat{\Delta} = \hat{\Omega} - \Omega$, and $\hat{\Gamma} = \hat{\Omega}^{-1} - \Omega^{-1}$.

Lemma 4. *(Mai et al., 2012) There exist positive constants ϵ_0 and $c_i, i \leq 4$ such that*

$$\begin{aligned} \mathbf{P}\left(\left|\hat{\delta}_{\mathcal{D}}\right|_\infty \geq \epsilon\right) &\leq 2r \exp\left(-c_1 n\epsilon^2\right), \\ \mathbf{P}\left(\|\hat{\Delta}_{\mathcal{D}\mathcal{D}}\|_\infty \geq \epsilon\right) &\leq 2r^2 \exp\left(-c_2 nr^{-2}\epsilon^2\right), \\ \mathbf{P}\left(\|\hat{\Gamma}_{\mathcal{D}\mathcal{D}}\|_\infty \geq \epsilon\right) &\leq 2r^2 \exp\left(-c_4 nr^{-2}\epsilon^2\right), \\ \mathbf{P}\left(\|\hat{\Lambda}_{\mathcal{D}^c\mathcal{D}}\|_\infty \geq \epsilon\right) &\leq 2pr \exp\left(-c_3 nr^{-2}\epsilon^2\right), \end{aligned}$$

for any $\epsilon \leq \epsilon_0$ and subset $\mathcal{D} \subseteq \{1, \ldots, p\}$, where $r = |\mathcal{D}|$ and $\hat{\Lambda}_{\mathcal{D}^c\mathcal{D}} = \hat{\Omega}_{\mathcal{D}^c\mathcal{D}}\hat{\Omega}^{-1}_{\mathcal{D}\mathcal{D}} - \Omega_{\mathcal{D}^c\mathcal{D}}\Omega^{-1}_{\mathcal{D}\mathcal{D}}$.

Proof of Theorem 1 Let $\rho = \rho_{\min}(\Omega)$ and κ be a sequence with $2q \leq \kappa \leq n$. First, we will show that Z satisfies the Sparse Riesz condition with $\rho/2$ and rank κ with probability tending to 1. For any subset $\mathcal{D} \subset \{1, \ldots, p\}$,

$$\rho_{\min}\left(\hat{\Omega}_{\mathcal{D}\mathcal{D}}\right) = \inf_{\|\mathbf{u}\|_2=1} \left\{u^T\Omega_{\mathcal{D}\mathcal{D}}u - u^T\left(\Omega_{\mathcal{D}\mathcal{D}} - \hat{\Omega}_{\mathcal{D}\mathcal{D}}\right)u\right\} \geq \rho_{\min}(\Omega_{\mathcal{D}\mathcal{D}}) - \left\|\Omega_{\mathcal{D}\mathcal{D}} - \hat{\Omega}_{\mathcal{D}\mathcal{D}}\right\|_2.$$

From Cauchy's interlacing theorem, $\min_{|\mathcal{D}|\leq\kappa} \rho_{\min}(\Omega_{\mathcal{D}\mathcal{D}}) \geq \rho$. By Lemma 4, it follows that

$$\begin{aligned} \mathbf{P}\left(\min_{|\mathcal{D}|\leq\kappa} \rho_{\min}\left(\hat{\Omega}_{\mathcal{D}\mathcal{D}}\right) \leq \rho/2\right) &\leq \mathbf{P}\left(\max_{|\mathcal{D}|\leq\kappa} \left\|\Omega_{\mathcal{D}\mathcal{D}} - \hat{\Omega}_{\mathcal{D}\mathcal{D}}\right\|_2 \geq \min_{|\mathcal{D}|\leq\kappa} \rho_{\min}\left(\Omega_{\mathcal{D}\mathcal{D}}\right) - \rho/2\right) \\ &\leq \sum_{|\mathcal{D}|\leq\kappa} \mathbf{P}\left(\left\|\hat{\Delta}_{\mathcal{D}\mathcal{D}}\right\|_\infty \geq \rho/2\right) \\ &\leq p^\kappa 2\kappa^2 \exp\left(-c_2 n\kappa^{-2}(\rho/2)^2\right) \to 0, \end{aligned} \tag{5.1}$$

provided that $n/\kappa^2 \to \infty$ and $n/\kappa^2 \gg \kappa \log p$ as $n \to \infty$.

Second, we will prove that $\hat{\beta}^{\text{OR}}$ satisfies the first inequality in the uniqueness condition with $\rho/2$. On the event $\min_{|\mathcal{D}|\leq\kappa} \rho_{\min}(\hat{\Omega}_{\mathcal{D}\mathcal{D}}) > \rho/2$, $\hat{\Omega}_{\mathcal{S}\mathcal{S}}$ is invertible since $|\mathcal{S}| = q \leq \kappa/2$. Hence, (5.1) implies that $\hat{\beta}^{\text{OR}}_{\mathcal{S}}$ satisfies

$$\hat{\beta}^{\text{OR}} = \hat{\Omega}^{-1}_{\mathcal{S}\mathcal{S}}\hat{\theta}_{\mathcal{S}}$$

with probability tending to 1. By the triangular inequality,

$$\min_{j\in\mathcal{S}} \left|\hat{\beta}^{\text{OR}}_j\right| \geq \min_{j\in\mathcal{S}} \left|\beta^{\text{OR}}_j\right| - \max_{j\in\mathcal{S}} \left|\hat{\beta}^{\text{OR}}_j - \beta^{\text{OR}}_j\right| \geq m_{\mathcal{S}} - \left\|\hat{\beta}^{\text{OR}}_{\mathcal{S}} - \beta^{\text{OR}}_{\mathcal{S}}\right\|_\infty.$$

Since $\beta^{\text{OR}}_{\mathcal{S}} = \Omega^{-1}_{\mathcal{S}\mathcal{S}}\theta_{\mathcal{S}}$,

$$\hat{\beta}^{\text{OR}}_{\mathcal{S}} - \beta^{\text{OR}}_{\mathcal{S}} = \hat{\Omega}^{-1}_{\mathcal{S}\mathcal{S}}\hat{\theta}_{\mathcal{S}} - \Omega^{-1}_{\mathcal{S}\mathcal{S}}\theta_{\mathcal{S}} = \hat{\Gamma}_{\mathcal{S}\mathcal{S}}\hat{\delta}_{\mathcal{S}} + \Omega^{-1}_{\mathcal{S}\mathcal{S}}\hat{\delta}_{\mathcal{S}} + \hat{\Gamma}_{\mathcal{S}\mathcal{S}}\theta_{\mathcal{S}}.$$

From Lemma 4, there exist positive constants d_1 and d_2 such that

$$\begin{aligned} \|\hat{\beta}^{\text{OR}}_{\mathcal{S}} - \beta^{\text{OR}}_{\mathcal{S}}\|_\infty &\leq \|\hat{\Gamma}_{\mathcal{S}\mathcal{S}}\|_\infty\|\hat{\delta}_{\mathcal{S}}\|_\infty + \|\Omega^{-1}_{\mathcal{S}\mathcal{S}}\|_\infty\|\hat{\delta}_{\mathcal{S}}\|_\infty + \|\hat{\Gamma}_{\mathcal{S}\mathcal{S}}\|_\infty\|\theta_{\mathcal{S}}\|_\infty \\ &\leq d_1\|\hat{\Gamma}_{\mathcal{S}\mathcal{S}}\|_\infty + d_2\|\hat{\delta}_{\mathcal{S}}\|_\infty, \end{aligned}$$

for all sufficiently large n. From Lemma 4 again, there exist positive constants d_3 and d_4 such that

$$\begin{aligned}\mathbf{P}\left(\min_{j\in\mathcal{S}}|\hat\beta_j^{\mathrm{OR}}|\le\lambda(a+2/\rho)\right)&\le\mathbf{P}\left(m_{\mathcal{S}}-\|\hat\beta_{\mathcal{S}}^{\mathrm{OR}}-\beta_{\mathcal{S}}^{\mathrm{OR}}\|_\infty\le\lambda(a+2/\rho)\right)\\&\le\mathbf{P}\left(d_1\|\hat\Gamma_{\mathcal{SS}}\|_\infty+d_2\|\hat\delta_{\mathcal{S}}\|_\infty\ge m_{\mathcal{S}}-\lambda(a+2/\rho)\right)\\&\le\mathbf{P}\left(\|\hat\Gamma_{\mathcal{SS}}\|_\infty\ge d_3m_{\mathcal{S}}\right)+\mathbf{P}\left(\|\hat\delta_{\mathcal{S}}\|_\infty\ge d_4m_{\mathcal{S}}\right)\\&\le 2q^2\exp\left(-c_4nq^{-2}(d_3m_{\mathcal{S}})^2\right)+2q\exp\left(-c_1n(d_4m_{\mathcal{S}})^2\right),\end{aligned}$$

for all sufficiently large n since $m_{\mathcal{S}}\gg\lambda$. Hence the first inequality in (3.3) holds with probability tending to 1, provided that $nm_{\mathcal{S}}^2/q^2\gg n\lambda^2/q^2\to\infty$ and $n\lambda^2/q^2\gg\log p$ as $n\to\infty$.

Third, we will show that the third inequality in the uniqueness condition holds with probability tending to 1. By using $\theta_{\mathcal{N}}=\Omega_{\mathcal{NS}}\Omega_{\mathcal{SS}}^{-1}\theta_{\mathcal{S}}$,

$$\begin{aligned}\hat\theta_{\mathcal{N}}-\hat\Omega_{\mathcal{NS}}\hat\beta_{\mathcal{S}}^{\mathrm{OR}}&=\hat\delta_{\mathcal{N}}+\theta_{\mathcal{N}}-\hat\Omega_{\mathcal{NS}}\hat\Omega_{\mathcal{SS}}^{-1}\hat\theta_{\mathcal{S}}\\&=\hat\delta_{\mathcal{N}}+\Omega_{\mathcal{NS}}\Omega_{\mathcal{SS}}^{-1}\theta_{\mathcal{S}}-\hat\Omega_{\mathcal{NS}}\hat\Omega_{\mathcal{SS}}^{-1}\hat\theta_{\mathcal{S}}.\end{aligned}$$

From Lemma 4, there exist positive constants e_1 and e_2 such that

$$\begin{aligned}\|\hat\theta_{\mathcal{N}}-\hat\Omega_{\mathcal{NS}}\hat\beta_{\mathcal{S}}^{\mathrm{OR}}\|_\infty&\le\|\hat\delta_{\mathcal{N}}\|_\infty+\|\Omega_{\mathcal{NS}}\Omega_{\mathcal{SS}}^{-1}\theta_{\mathcal{S}}-\hat\Omega_{\mathcal{NS}}\hat\Omega_{\mathcal{SS}}^{-1}\hat\theta_{\mathcal{S}}\|_\infty\\&\le\|\hat\delta_{\mathcal{N}}\|_\infty+\|\hat\Lambda_{\mathcal{NS}}\|_\infty\|\hat\delta_{\mathcal{S}}\|_\infty+\|\Omega_{\mathcal{NS}}\Omega_{\mathcal{SS}}^{-1}\|_\infty\|\hat\delta_{\mathcal{S}}\|_\infty+\|\theta_{\mathcal{S}}\|_\infty\|\hat\Lambda_{\mathcal{NS}}\|_\infty\\&\le e_1\|\hat\delta\|_\infty+e_2\|\hat\Lambda_{\mathcal{NS}}\|_\infty,\end{aligned}$$

for all sufficiently large n. From Lemma 4 again, there exist positive constants e_3 and e_4 such that

$$\begin{aligned}\mathbf{P}\left(\|\hat\theta_{\mathcal{N}}-\hat\Omega_{\mathcal{NS}}\hat\beta_{\mathcal{S}}^{\mathrm{OR}}\|_\infty>\lambda\min\{1,a\rho/2\}\right)&\le\mathbf{P}\left(e_1\|\hat\delta\|_\infty+e_2\|\hat\Lambda_{\mathcal{NS}}\|_\infty>\lambda\min\{1,a\rho/2\}\right)\\&\le\mathbf{P}\left(\|\hat\delta\|_\infty>e_3\lambda\right)+\mathbf{P}\left(\|\hat\Lambda_{\mathcal{NS}}\|_\infty>e_4\lambda\right)\\&\le 2p\exp\left(-c_1n(e_3\lambda)^2\right)+2(p-q)q\exp\left(-c_3nq^{-2}(e_4\lambda)^2\right),\end{aligned}$$

for all sufficiently large n. Hence the third inequality in (3.3) holds with probability tending to 1, provided that $n\lambda^2/q^2\to\infty$ and $n\lambda^2/q^2\gg\log p$ as $n\to\infty$. □

Acknowledgement

This paper was supported by Konkuk University in 2021.

References

Bickel PJ and 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**, 989–1010.

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

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

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

Fan J, Feng Y, and Tong X (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **74**, 745–771.

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 1348–1360.

Fan J and Peng H (2004). Nonconcave penalized likelihood with a diverging number of parameters, *The Annals of Statistics*, **32**, 928–961.

Fan Y and Tang CY (2012). Tuning parameter selection in high dimensional penalized likelihood, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **75**, 531–552.

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

Gordon GJ, Jensen RV, Hsiao LL *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.

Hastie T, Tibshirani R, and Buja A (1994). Flexible discriminant analysis by optimal scoring, *Journal of the American Statistical Association*, **89**, 1255–1270.

Hastie T, Tibshirani R, and Friedman J (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science & Business Media, Berlin.

James GM, Radchenko P, and Lv J (2009). Dasso: Connections between the dantzig selector and lasso, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **71**, 127–142.

Kim D, Lee S, and Kwon S (2020). A unified algorithm for the non-convex penalized estimation: The ncpen package, *The R Journal*, **12**, 120–133.

Kim Y, Choi H, and Oh H-S (2008). Smoothly clipped absolute deviation on high dimensions, *Journal of the American Statistical Association*, **103**, 1665–1673.

Kim Y, Jeon J-J, and Han S (2016). A necessary condition for the strong oracle property, *Scandinavian Journal of Statistics*, **43**, 610–624.

Kim Y and Kwon S (2012). Global optimality of nonconvex penalized estimators, *Biometrika*, **99**, 315–325.

Kwon S and Kim Y (2012). Large sample properties of the scad-penalized maximum likelihood estimation on high dimensions, *Statistica Sinica*, **22**, 629–653.

Kwon S, Moon H, Chang J, and Lee S (2021). Sufficient conditions for the oracle property in penalized linear regression, *The Korean Journal of Applied Statistics*, **34**, 279–293.

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

Na O and Kwon S (2018). Non-convex penalized estimation for the ar process, *Communications for Statistical Applications and Methods*, **25**, 453–470.

Shen X, Pan W, Zhu Y, and Zhou H (2013). On constrained and regularized high-dimensional regression, *Annals of the Institute of Statistical Mathematics*, **65**, 807–832.

Singh D, Febbo PG, Ross K *et al.* (2002). Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, **1**, 203–209.

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

Trendafilov NT and Jolliffe IT (2007). Dalass: Variable selection in discriminant analysis via the lasso, *Computational Statistics & Data Analysis*, **51**, 3718–3736.

Witten DM and Tibshirani R (2011). Penalized classification using fisher's linear discriminant, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 753–772.

Zhang C-H (2010). Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, **38**, 894–942.

Zhang C-H and Zhang T (2012). A general theory of concave regularization for high-dimensional sparse estimation problems, *Statistical Science*, **27**, 576–593.

Zhao P and Yu B (2006). On model selection consistency of lasso, *Journal of Machine Learning Research*, **7**, 2541–2563.

Received December 10, 2023; Revised March 20, 2024; Accepted April 09, 2024