Communications for Statistical Applications and Methods
2021, Vol. 28, No. 1, 21–37

https://doi.org/10.29220/CSAM.2021.28.1.021
Print ISSN 2287-7843 / Online ISSN 2383-4757

High-dimensional linear discriminant analysis with moderately clipped LASSO

Jaeho Chang[a], Haeseong Moon[a], Sunghoon Kwon[1,a]

[a]Department of Applied Statistics, Konkuk University, Korea

Abstract

There is a direct connection between linear discriminant analysis (LDA) and linear regression since the direction vector of the LDA can be obtained by the least square estimation. The connection motivates the penalized LDA when the model is high-dimensional where the number of predictive variables is larger than the sample size. In this paper, we study the penalized LDA for a class of penalties, called the moderately clipped LASSO (MCL), which interpolates between the least absolute shrinkage and selection operator (LASSO) and minimax concave penalty. We prove that the MCL penalized LDA correctly identifies the sparsity of the Bayes direction vector with probability tending to one, which is supported by better finite sample performance than LASSO based on concrete numerical studies.

Keywords: high-dimensional LDA, LASSO, MCP, moderately clipped LASSO

1. Introduction

Linear discriminant analysis (LDA) requires an estimation of the inverse conditional covariance matrix where the pooled sample covariance matrix is a manageable solution. However, the pooled sample covariance matrix is singular when the model is high-dimensional where the number of predictive variables exceeds the sample size. The performance of the LDA is undermined (Krzanowski *et al.*, 1995) and asymptotically no better than the random guessing (Bickel and Levina, 2004) without resolving the singularity. Many works of literature have focused on this issue and suggested other alternatives by directly modifying the singular pooled sample covariance matrix or constructing relevant shrunken centroid means (Guo *et al.*, 2006; Fan and Fan, 2008; Wu *et al.*, 2009; Cai and Liu, 2011; Witten and Tibshirani, 2011; Clemmensen *et al.*, 2011; Shao *et al.*, 2011).

Meanwhile, there is a direct connection between the LDA and linear regression (Hastie *et al.*, 2009) since the direction vector of the LDA can be obtained by using the least square estimation (LSE) (Hastie *et al.*, 2009). However, this connection is lost when the pooled sample covariance matrix is singular. This motivated Mai *et al.* (2012) to study penalized LDA with the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996). Mai *et al.* (2012) proved that LASSO penalized LDA finds all the features that significantly contribute to the classification when the model is high-dimensional and sparse, which is a unique theoretical work for the high-dimensional penalized LDA. Mai *et al.* (2012) provided various numerical studies to confirm that LASSO penalized LDA performs better than other competitors.

This paper was supported by Konkuk University in 2019.

[1] Corresponding author: Department of Applied Statistics, Konkuk University, Korea, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea. E-mail: shkwon0522@gmail.com

Published 31 January 2021 / journal homepage: http://csam.or.kr

In this paper, we propose the penalized LDA with the moderately clipped LASSO (MCL) (Kwon *et al.*, 2015) as an alternative to LASSO. The LASSO has been known to have higher prediction accuracy than minimax concave penalty (MCP) (Zhang, 2010) for finite samples since the shrinkage effect can increase prediction accuracy (Efron and Morris, 1975; Casella, 1985; Zhang and Huang, 2008). However, LASSO has proved to select unnecessary predictive variables even for low-dimensional linear regression (Zou, 2006). Meanwhile, MCL was designed to recover the ideal performance of LASSO by indexing a class of non-convex penalties from LASSO to MCP. Therefore MCL can select relevant predictive variables as MCP while keeping the prediction accuracy of LASSO.

We proved that, with probability tending to one, MCL penalized LDA is the same as the oracle LASSO that is a theoretically optimal LASSO obtained by using relevant predictive variables only. The equivalence can be obtained using LASSO penalized LDA as proved by Mai *et al.* (2012). However, MCL does not require the Strong Irrepresentable condition (Zhao and Yu, 2006) on the marginal covariance matrix, and represents a potential theoretical advantage of MCL over LASSO. We provided various numerical studies to show how MCL performs with finite samples as a better alternative to LASSO for the high-dimensional penalized LDA.

The rest of the paper is organized as follows. Section 2 introduces the penalized LDA. Section 3 introduces MCL and related statistical properties. Section 4 shows the results of numerical studies. Relevant proofs are given in the Appendix.

2. Penalized linear discriminant analysis

2.1. Linear discriminant analysis and least square estimation

Let $\mathbf{X} \in \mathbb{R}^p$ be a p-dimensional random vector of predictive variables and $C \in \{1, 2\}$ a random class label to be identified. Given $\mathbf{X} = \mathbf{x}$, the Bayes classifier becomes

$$\phi^{\text{Bayes}}(\mathbf{x}) = \arg\max_{c\in\{1,2\}} \mathbf{P}(C = c|\mathbf{X} = \mathbf{x}),$$

which minimizes the misclassification error, $\mathbf{P}(C \neq \phi(\mathbf{X}))$, over the set of classifiers, $\phi : \mathbb{R}^p \to \{1, 2\}$. The LDA (Fisher, 1936) assumes that

$$\mathbf{X}|C = c \sim N_p(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}), \quad c \in \{1, 2\}, \tag{2.1}$$

where $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}$ denote the mean vector and covariance matrix of the p-dimensional normal distribution. Then the Bayes classifier is equivalent to the Bayes discriminant rule: the class of C given $\mathbf{X} = \mathbf{x}$ becomes 2 if $\mathbf{x}$ satisfies

$$\left(\mathbf{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right)^T \boldsymbol{\beta}^{\text{Bayes}} + \log\left(\frac{\pi_2}{\pi_1}\right) > 0,$$

where $\pi_c = \mathbf{P}(C = c)$, $c \in \{1, 2\}$, are the class probabilities and

$$\boldsymbol{\beta}^{\text{Bayes}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}$$

is the Bayes direction vector with $\boldsymbol{\theta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$.

Let $c_i \in \{1, 2\}$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p, i \leq n$, be n samples of the class label and predictive variables. For each class label $c \in \{1, 2\}$, let $\hat{\boldsymbol{\Sigma}}_c = \sum_{c_i=c}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)^T/(n_c - 1)$ be the sample covariance matrix, $\hat{\boldsymbol{\mu}}_c = \sum_{c_i=c} \mathbf{x}_i/n_c$ the sample mean vector, and $n_c = \sum_{i=1}^n I(c_i = c)$ the number of

samples with class label c. The LDA estimates the Bayes direction vector with the LDA direction vector:

$$\hat{\boldsymbol{\beta}}^{\mathrm{LDA}} = \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\theta}}, \tag{2.2}$$

where $\hat{\boldsymbol{\Sigma}} = \sum_{c\in\{1,2\}}(n_c - 1)\hat{\boldsymbol{\Sigma}}_c/(n-2)$ is the pooled sample covariance matrix and $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1$. These arguments lead to the linear discriminant rule: the class of C given $\mathbf{X} = \mathbf{x}$ becomes 2 if $\mathbf{x}$ satisfies

$$\left(\mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}\right)^T \hat{\boldsymbol{\beta}}^{\mathrm{LDA}} + \log\left(\frac{n_2}{n_1}\right) > 0, \tag{2.3}$$

where the class probabilities are estimated by the sample class proportions, $\hat{\pi}_c = n_c/n$, $c \in \{1, 2\}$.

There is an intimate connection (Hastie *et al.*, 2009) between the LDA and LSE when $p \le n$:

$$\hat{\boldsymbol{\beta}}^{\mathrm{LSE}} = c\hat{\boldsymbol{\beta}}^{\mathrm{LDA}} \tag{2.4}$$

for some constant $c > 0$, where

$$ary\left(\hat{\alpha}^{\mathrm{LSE}}, \hat{\boldsymbol{\beta}}^{\mathrm{LSE}}\right) = \underset{\alpha,\boldsymbol{\beta}}{\arg\min} \sum_{i=1}^{n} \frac{\left(y_i - \alpha - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2}{2n} \tag{2.5}$$

ary and $y_i = (-1)^{c_i} n/n_{c_i}, i \le n$. Hence the linear discriminant rule in (2.3) is the same as:

$$\left(\mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}\right)^T \hat{\boldsymbol{\beta}}^{\mathrm{LSE}} + c\log\left(\frac{n_2}{n_1}\right) > 0, \tag{2.6}$$

which implies that we can cast the LDA into the framework of the LSE.

2.2. Penalized linear discriminant analysis and least square estimation

The equations in (2.2) and (2.4) fail to hold when $p > n$ since the pooled sample covariance matrix is singular, which raises a challenging problem of estimating the Bayes direction vector. For the problem, Mai *et al.* (2012) proposed to estimate the Bayes direction vector by using LASSO:

$$\left(\hat{\alpha}^{\lambda}, \hat{\boldsymbol{\beta}}^{\lambda}\right) = \underset{\alpha,\boldsymbol{\beta}}{\arg\min} \left\{ \sum_{i=1}^{n} \frac{\left(y_i - \alpha - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2}{2n} + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

for some $\lambda > 0$, where the solution can be defined even when $p > n$. Let $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_p)$ be the centered design matrix for the LSE in (2.5), where $\mathbf{Z}_j = (\mathbf{I} - \boldsymbol{\Pi}_1)\mathbf{X}_j$ with $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})^T$, $j \le p$, and $\boldsymbol{\Pi}_1$ is the projection matrix onto $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$. Then we can see that the solution can be obtained by

$$\hat{\boldsymbol{\beta}}^{\lambda} = \underset{\boldsymbol{\beta}}{\arg\min} \left\{ \frac{\boldsymbol{\beta}^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\beta}}{2n} - \hat{\boldsymbol{\theta}}^T\boldsymbol{\beta} + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{2.7}$$

In addition, Mai *et al.* (2012) constructed an optimal discriminant rule for the penalized estimation: the class of C given $\mathbf{X} = \mathbf{x}$ becomes 2 if $\mathbf{x}$ satisfies

$$\left(\mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}\right)^T \hat{\boldsymbol{\beta}}^{\lambda} + \frac{\hat{\boldsymbol{\beta}}^{\lambda^T}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\beta}}^{\lambda}}{\hat{\boldsymbol{\theta}}^T\hat{\boldsymbol{\beta}}^{\lambda}} \log\left(\frac{n_2}{n_1}\right) > 0, \tag{2.8}$$

whenever $\hat{\boldsymbol{\theta}}^T\hat{\boldsymbol{\beta}}^\lambda > 0$. Note that the classification rule in (2.8) can be used for any linear classifier, see Proposition 2 in Mai *et al.* (2012) for some details. For example, the optimal discriminant rule in (2.8) reduces to the linear discriminant rule in (2.6) when $p \le n$ and $\lambda = 0$ since $\hat{\boldsymbol{\beta}}^\lambda = \hat{\boldsymbol{\beta}}^{\text{LSE}} = c\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\theta}}$ implies $\hat{\boldsymbol{\beta}}^{\lambda^T}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\beta}}^\lambda/\hat{\boldsymbol{\theta}}^T\hat{\boldsymbol{\beta}}^\lambda = c$.

3. Penalized LDA with moderately clipped LASSO

3.1. Definition

A natural extension for LASSO in (2.7) is the non-convex penalized estimation. For example, we can use the MCP (Zhang, 2010) that has been preferred for variable selection than LASSO due to the oracle property (Fan and Li, 2001; Kim *et al.*, 2008; Zhang, 2010). However, LASSO has been preferred for prediction than MCP since shrinkage effect often produces better prediction accuracy (Efron and Morris, 1975). As an alternative, we propose to use MCL (Kwon *et al.*, 2015) for the penalized LDA:

$$\hat{\boldsymbol{\beta}}^{\lambda,\gamma} = \arg\min_{\boldsymbol{\beta}} Q_{\lambda,\gamma}(\boldsymbol{\beta}), \tag{3.1}$$

where

$$Q_{\lambda,\gamma}(\boldsymbol{\beta}) = \frac{\boldsymbol{\beta}^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\beta}}{2n} - \hat{\boldsymbol{\theta}}^T\boldsymbol{\beta} + \sum_{j=1}^{p} J_{\lambda,\gamma}(|\beta_j|)$$

and $J_{\lambda,\gamma}$ is MCL that satisfies $J_{\lambda,\gamma}(0) = 0$ and

$$\frac{dJ_{\lambda,\gamma}(t)}{dt} = \nabla J_{\lambda,\gamma}(t) = \max\left\{\lambda - \frac{t}{a}, \gamma\right\}, \quad t > 0$$

for some $a > 1$ and $\lambda \ge \gamma \ge 0$.

Note that $\nabla J_{\lambda,\gamma}(t) = \lambda - t/a$ for $t < a(\lambda - \gamma)$ and $\nabla J_{\lambda,\gamma}(t) = \gamma$ for $t \ge a(\lambda - \gamma)$. Hence, MCL becomes a smooth interpolation between MCP and LASSO with two tuning parameters $\lambda \ge 0$ and $0 \le \gamma \le \lambda$. The two tuning parameters of MCL play different roles. First, λ controls the concavity of MCL near the origin for a fixed γ as it does in MCP. In the right panel of Figure 1, we can see that the concavity of MCL increases as λ does. Second, γ regularizes the amount of shrinkage in large non-zero regression coefficients for a fixed λ as it does in LASSO, which is illustrated in the left panel of Figure 1. These imply that MCL can control the sparsity and shrinkage effect simultaneously. Hence, we can expect an estimator that balances between MCP and LASSO for given finite samples as studied in Kwon *et al.* (2015) for high dimensional linear regression.

3.2. Asymptotic properties

In this subsection, we provide some asymptotic properties for MCL. We assume that there is a non-empty subset $\mathcal{A} \subset \{1, \ldots, p\}$ such that $\beta_j^{\text{Bayes}} \neq 0$ for $j \in \mathcal{A}$ and $\beta_j^{\text{Bayes}} = 0$ for $j \in \mathcal{N} := \mathcal{A}^c$. The main results imply that MCL is asymptotically equivalent to a theoretical estimator, the oracle LASSO:

$$\hat{\boldsymbol{\beta}}^{oL,\gamma} = \arg\min_{\beta_j=0, j\in\mathcal{A}} \left\{ \frac{\boldsymbol{\beta}^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\beta}}{2n} - \hat{\boldsymbol{\theta}}^T\boldsymbol{\beta} + \gamma\sum_{j=1}^{p}|\beta_j| \right\}, \tag{3.2}$$

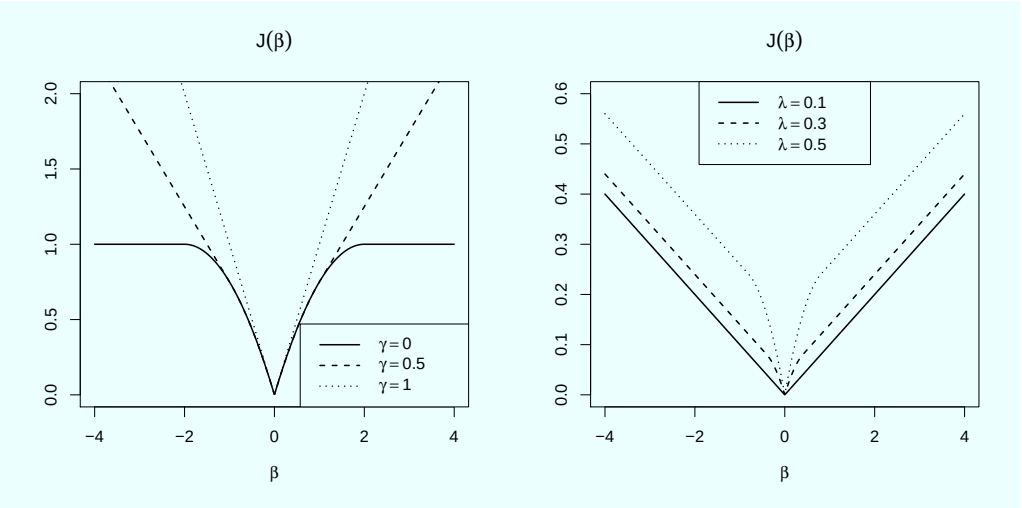

Figure 1: *Various shapes of MCL with $a = 2$: $\lambda = 1$ for the left panel and $\gamma = 0.1$ for the right panel.*

for some $\gamma \geq 0$. Note that the oracle LASSO is simply the oracle LSE (Fan and Li, 2001) if $\gamma = 0$. The oracle LASSO is not available in practice since $\mathcal{A}$ is unknown. However, the oracle LASSO plays an important role in developing asymptotic properties of MCL as studied in Mai *et al.* (2012) and Kwon *et al.* (2015). We can observe similar frameworks in other studies on the non-convex penalized estimation (Kim *et al.*, 2008; Zhang, 2010; Kim and Kwon, 2012).

Before proceeding, we define some notations. For any vector $\mathbf{a} = (a_k, k \leq s)$ and subset $\mathcal{S} \subset \{1, \ldots, s\}$, let $\|\mathbf{a}\|_1 = \sum_{k=1}^{s} |a_k|$, $\|\mathbf{a}\|_2 = (\sum_{k=1}^{s} a_k^2)^{1/2}$, $\|\mathbf{a}\|_\infty = \max_{k \leq s} |a_k|$, $\text{supp}(\mathbf{a}) = \{k : a_k \neq 0\}$ and $\mathbf{a}_\mathcal{S} = (a_k, k \in \mathcal{S})$. For any matrix $\mathbf{A} = (A_{kj}, k \leq s, j \leq t)$ and subsets $\mathcal{S} \subset \{1, \ldots, s\}$ and $\mathcal{S}' \subset \{1, \ldots, t\}$, let $\lambda_{\min}(\mathbf{A})$ be the minimum eigenvalues of symmetric $\mathbf{A}$, $\|\mathbf{A}\|_\infty = \max_{k \leq s} \sum_{j=1}^{t} |A_{kj}|$, $\mathbf{A}_\mathcal{S} = (A_{kj}, k \leq s, j \in \mathcal{S})$, $\mathbf{A}_{\mathcal{S}\mathcal{S}'} = (A_{kj}, k \in \mathcal{S}, j \in \mathcal{S}')$, and $|\mathcal{S}|$ the cardinality of $\mathcal{S}$.

We first introduce a lemma that gives sufficient conditions for the uniqueness of a minimizer of $Q_{\lambda,\gamma}$ when $p \leq n$. Let $\Xi_{\lambda,\gamma}$ be the set of all local minimizers of $Q_{\lambda,\gamma}$ and $\rho = \lambda_{\min}(\mathbf{Z}^T\mathbf{Z}/n)$.

Lemma 1. *If $\hat{\boldsymbol{\beta}}$ satisfies*

$$\hat{\theta}_j - \frac{\mathbf{Z}_j^T \mathbf{Z}\hat{\boldsymbol{\beta}}}{n} = \text{sign}\left(\hat{\beta}_j\right) \nabla J_{\lambda,\gamma}\left(|\hat{\beta}_j|\right), \quad j \in \mathcal{S},$$
$$\left|\hat{\theta}_j - \frac{\mathbf{Z}_j^T \mathbf{Z}\hat{\boldsymbol{\beta}}}{n}\right| \leq \lambda, \quad j \in \mathcal{S}^c,$$

then $\{\hat{\boldsymbol{\beta}}\} = \Xi_{\lambda,\gamma}$, provided that $\rho > 1/a$, where $\mathcal{S} = \text{supp}(\hat{\boldsymbol{\beta}})$.

Remark 1. Lemma 1 is a slight modification of the second order Karush-Kuhn-Tucker sufficient conditions for $Q_{\lambda,\gamma}$ whose proof can be found in other literature (Fan *et al.*, 2014; Kwon *et al.*, 2015). Let $\tilde{\nabla}^2 J_\lambda(t), t > 0$ be the maximum concavity (Zhang, 2010) of the penalty $J_{\lambda,\gamma}$, that is,

$$\tilde{\nabla}^2 J_{\lambda,\gamma}(t) = \lim_{\varepsilon \to 0+} \inf_{t-\varepsilon < t_1 < t_2 < t+\varepsilon} \frac{\nabla J_{\lambda,\gamma}(t_2) - \nabla J_{\lambda,\gamma}(t_1)}{t_2 - t_1}, \quad t > 0.$$

Note that $Q_{\lambda,\gamma}$ is globally convex if $\rho + \inf_{t>0} \tilde{\nabla}^2 J_{\lambda,\gamma}(t) = \rho - 1/a > 0$, which implies $\hat{\boldsymbol{\beta}}$ is a unique minimizer of $Q_{\lambda,\gamma}$.

Lemma 1 implies that $Q_{\lambda,\gamma}$ has a unique minimizer when $p \leq n$, although $J_{\lambda,\gamma}$ is not convex. However, when $p > n$, Lemma 1 fails to hold for any $a > 1$ since $\mathbf{Z}$ does not have full rank. Therefore, we present a slightly weaker result under the Sparse Riesz condition (Zhang, 2010; Kim and Kwon, 2012; Kim *et al.*, 2016). The next lemma shows that there exists a unique minimizer in a restricted parameter space, which is a direct application of Theorem 3 in Kim and Kwon (2012). Let $\Xi^{\kappa}_{\lambda,\gamma} = \{\boldsymbol{\beta} \in \Xi_{\lambda,\gamma} : |\text{supp}(\boldsymbol{\beta})| \leq \kappa\}$ and $\hat{\rho}^{\text{src}}_{\kappa} = \min_{|\mathcal{D}|\leq 2\kappa} \lambda_{\min}(\mathbf{Z}^T_{\mathcal{D}}\mathbf{Z}_{\mathcal{D}}/n)$ for some $\kappa > 0$.

Lemma 2. *If $\hat{\boldsymbol{\beta}}$ satisfies*

$$\hat{\theta}_j - \frac{\mathbf{Z}_j^T\mathbf{Z}\hat{\boldsymbol{\beta}}}{n} = \text{sign}\left(\hat{\beta}_j\right)\nabla J_{\lambda,\gamma}\left(\left|\hat{\beta}_j\right|\right), \quad j \in \mathcal{S},$$
$$\left|\hat{\theta}_j - \frac{\mathbf{Z}_j^T\mathbf{Z}\hat{\boldsymbol{\beta}}}{n}\right| \leq \lambda, \quad j \in \mathcal{S}^c,$$

then $\{\hat{\boldsymbol{\beta}}\} = \Xi^{\kappa}_{\lambda,\gamma}$, provided that $\hat{\rho}^{\text{src}}_{\kappa} > 1/a$ and $|\mathcal{S}| \leq \kappa \leq n/2$, where $\mathcal{S} = \text{supp}(\hat{\boldsymbol{\beta}})$.

Lemma 2 gives sufficient conditions for a minimizer to be unique in $\Xi^{\kappa}_{\lambda,\gamma}$ under the Sparse Riesz condition, and we can construct similar conditions for the oracle LASSO in (3.2) by adding one more condition.

Lemma 3. *If $\hat{\boldsymbol{\beta}}^{oL,\gamma}$ satisfies*

$$\left|\hat{\beta}_j^{oL,\gamma}\right| > a(\lambda - \gamma), \quad j \in \mathcal{A},$$
$$\hat{\theta}_j - \frac{\mathbf{Z}_j^T\mathbf{Z}\hat{\boldsymbol{\beta}}^{oL,\gamma}}{n} = \gamma\text{sign}\left(\hat{\beta}_j^{oL,\gamma}\right), \quad j \in \mathcal{A},$$
$$\left|\hat{\theta}_j - \frac{\mathbf{Z}_j^T\mathbf{Z}\hat{\boldsymbol{\beta}}^{oL,\gamma}}{n}\right| \leq \lambda, \quad j \in \mathcal{A}^c,$$

then $\{\hat{\boldsymbol{\beta}}^{oL,\gamma}\} = \Xi^{\kappa}_{\lambda,\gamma}$, provided that $\hat{\rho}^{\text{src}}_{\kappa} > 1/a$ and $|\mathcal{A}| \leq \kappa \leq n/2$.

Remark 2. Note that the second condition in Lemma 3 is equivalent to the first condition in Lemma 2 under the first condition in Lemma 3 so that Lemma 3 is a corollary of Lemma 2.

We now present the main results that show the oracle LASSO satisfies the conditions in Lemma 3 asymptotically so that the oracle LASSO is equivalent to MCL defined in (3.1). Let $\boldsymbol{\Omega} = \text{Cov}(\mathbf{X})$ be the marginal covariance matrix of $\mathbf{X}$. Let $\boldsymbol{\beta}^*$ be a vector that satisfies $\boldsymbol{\beta}^*_{\mathcal{A}} = \boldsymbol{\Omega}^{-1}_{\mathcal{A}\mathcal{A}}\boldsymbol{\theta}_{\mathcal{A}}$ and $\boldsymbol{\beta}^*_{\mathcal{N}} = \mathbf{0}$. For the study, we need regularity conditions below:

(C1) There are positive constants, $b_i, i \leq 3$, such that

$$\left\|\boldsymbol{\Omega}_{\mathcal{N}\mathcal{A}}\boldsymbol{\Omega}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_{\infty} < b_1, \quad \left\|\boldsymbol{\Omega}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_{\infty} < b_2, \quad \text{and} \quad \|\boldsymbol{\theta}_A\|_{\infty} < b_3.$$

(C2) The model and tuning parameters satisfy

$$q = o\left(nm_{\mathcal{A}}^2\right), \quad \log p = o\left(\frac{nm_{\mathcal{A}}^2}{q^2}\right), \quad \lambda = o\left(m_{\mathcal{A}}\right), \quad \gamma = o(\lambda), \quad \text{and} \quad nm_{\mathcal{A}}^2 \to \infty$$

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

(C3) There exists a sequence $q \le \kappa \le n/2$ such that

$$\liminf_{n\to\infty} \rho_\kappa^{\text{src}} > \frac{1}{a}, \quad \log p = o\left(\frac{n}{\kappa^3}\right), \quad \text{and} \quad \frac{n}{\kappa^2} \to \infty$$

as $n \to \infty$, where $\rho_\kappa^{\text{src}} = \min_{|\mathcal{D}|\le 2\kappa} \lambda_{\min}(\mathbf{\Omega}_{\mathcal{D}\mathcal{D}})$.

Remark 3. Condition (C1) is given by Mai *et al.* (2012) which requires $b_1 = 1$ for LASSO penalized LDA to be selection consistent. The restriction corresponds to the Strong Irrepresentable condition given by Zhao and Yu (2006) for LASSO penalized linear regression. Condition (C2) includes technical assumptions that are standard for the non-convex penalized regression, see Remark 3.2. Note that $\boldsymbol{\beta}^*$ is a scaled version of $\boldsymbol{\beta}^{\text{Bayes}}$ for the LSE framework that satisfies $\boldsymbol{\beta}^* = c\boldsymbol{\beta}^{\text{Bayes}}$ for some constant $c > 0$, see Proposition 3 in Mai *et al.* (2012) for some details. Condition (C3) assumes that any principal sub-matrix $\mathbf{\Omega}_{\mathcal{D}\mathcal{D}}$ of $\mathbf{\Omega}$ is non-singular whenever $|\mathcal{D}| \le 2\kappa$ which corresponds to the sparse Riesz condition in Zhang (2010) imposed on the design matrix in the linear regression.

Theorem 1. *Under (C1)–(C3), the oracle LASSO is unique minimizer of $Q_{\lambda,\gamma}$ with probability tending to one, in the sense that*

$$\lim_{n\to\infty} \mathbf{P}\left(\left\{\hat{\boldsymbol{\beta}}^{oL,\gamma}\right\} = \Xi_{\lambda,\gamma}^{\kappa}\right) = 1.$$

Remark 4. Theorem 1 holds for $\gamma = 0$ which proves that the oracle LSE becomes the unique minimizer of $Q_{\lambda,\gamma}$ when the penalty is MCP.

Remark 5. The conditions (C2) and (C3) can be simplified as

$$m_{\mathcal{A}} \gg \lambda \gg q\sqrt{\frac{\log p}{n}}$$

if $n \gg \kappa^3 \log p$, where $a \gg b$ implies $a/b \to \infty$ as $n \to \infty$. However, we need

$$m_{\mathcal{A}} \gg \lambda \gg \sqrt{\frac{\log p}{n}},$$

for the high-dimensional linear regression (Kwon *et al.*, 2015). The difference between the minimum signal sizes, up to a factor q, happens since the design matrix is high-dimensional and random for the LDA.

4. Numerical studies

In this section, we present the results of numerical studies including simulations and real data analysis. We obtained all the penalized estimators using R package ncpen (Kim *et al.*, 2020) that was developed based on the convex-concave procedure (Yuille and Rangarajan, 2002) and coordinate descent algorithm (Mazumder *et al.*, 2011).

Table 1: Averages of the four measures when $\mathbf{\Sigma} = \mathbf{\Sigma}^{(1)}$

		$p = 1000, q = 5$						$p = 2000, q = 10$					
		$\pi_1 = 0.50$			$\pi_1 = 0.67$			$\pi_1 = 0.50$			$\pi_1 = 0.67$		
	n	300	600	900	300	600	900	300	600	900	300	600	900
TPS	Bayes	5	5	5	5	5	5	10	10	10	10	10	10
	Lasso	5.000	5.000	5.000	4.990	5.000	5.000	9.760	10.000	10.000	9.140	9.985	10.000
	MCL_1	4.995	5.000	5.000	4.945	5.000	5.000	9.025	9.970	10.000	8.110	9.790	9.995
	MCL_2	4.905	5.000	5.000	4.610	4.995	5.000	7.785	9.855	9.985	5.615	9.285	9.915
	MCL_3	3.975	4.980	5.000	2.645	4.795	4.995	3.405	8.205	9.640	1.225	5.210	8.255
	MCP_1	4.865	5.000	5.000	4.580	5.000	5.000	6.595	9.580	9.970	5.200	9.050	9.855
	MCP_2	4.965	5.000	5.000	4.900	5.000	5.000	8.450	9.940	9.995	7.440	9.710	9.975
	MCP_3	4.980	5.000	5.000	4.940	5.000	5.000	9.155	9.970	10.000	8.250	9.860	9.995
FPS	Bayes	0	0	0	0	0	0	0	0	0	0	0	0
	Lasso	22.620	22.535	22.770	24.500	24.930	24.565	41.920	44.765	42.535	41.700	45.615	45.305
	MCL_1	1.655	1.555	1.605	2.170	1.895	2.435	5.060	1.590	1.110	9.765	2.670	1.200
	MCL_2	0.065	0.015	0.005	0.100	0.015	0.015	0.425	0.100	0.020	0.430	0.105	0.025
	MCL_3	0.000	0.000	0.000	0.000	0.000	0.000	0.015	0.000	0.000	0.010	0.000	0.000
	MCP_1	0.705	0.385	0.370	0.780	0.410	0.355	1.420	1.050	0.410	1.275	1.320	0.735
	MCP_2	6.720	3.795	1.985	7.180	5.175	2.835	8.560	12.285	9.380	8.885	12.270	10.625
	MCP_3	8.790	7.670	5.885	10.090	9.010	7.335	15.790	16.195	13.510	16.385	18.165	15.685
CMI	Bayes	1	1	1	1	1	1	1	1	1	1	1	1
	Lasso	0	0	0	0	0	0	0	0	0	0	0	0
	MCL_1	0.610	0.550	0.585	0.455	0.550	0.590	0.020	0.435	0.590	0.000	0.300	0.595
	MCL_2	0.855	0.985	0.995	0.620	0.980	0.985	0.060	0.790	0.965	0.000	0.430	0.900
	MCL_3	0.390	0.980	1.000	0.080	0.815	0.995	0.000	0.155	0.715	0.000	0.000	0.170
	MCP_1	0.475	0.725	0.725	0.235	0.695	0.735	0.000	0.230	0.670	0.000	0.100	0.415
	MCP_2	0.025	0.090	0.385	0.010	0.075	0.190	0.000	0.000	0.000	0.000	0.000	0.005
	MCP_3	0.000	0.025	0.065	0.005	0.025	0.065	0.000	0.000	0.000	0.000	0.000	0.000
ERR	Bayes	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	Lasso	0.2169	0.2064	0.2046	0.2204	0.2082	0.2036	0.2442	0.2168	0.2102	0.2571	0.2223	0.2116
	MCL_1	0.2085	0.2030	0.2031	0.2114	0.2043	0.2018	0.2372	0.2098	0.2056	0.2521	0.2153	0.2059
	MCL_2	0.2188	0.2054	0.2037	0.2261	0.2073	0.2027	0.2653	0.2202	0.2101	0.2769	0.2306	0.2132
	MCL_3	0.2576	0.2125	0.2059	0.2739	0.2213	0.2072	0.3466	0.2582	0.2259	0.3235	0.2816	0.2433
	MCP_1	0.2107	0.2020	0.2028	0.2159	0.2029	0.2008	0.2644	0.2127	0.2045	0.2706	0.2196	0.2051
	MCP_2	0.2163	0.2032	0.2027	0.2201	0.2048	0.2015	0.2559	0.2167	0.2078	0.2647	0.2240	0.2095
	MCP_3	0.2159	0.2052	0.2034	0.2189	0.2066	0.2028	0.2498	0.2165	0.2094	0.2600	0.2232	0.2102

TPS = number of true positive selection; FPS = number of false positive selection; CMI = correct model identification; ERR = classification error rate.

4.1. Simulation studies

The simulation studies were based on the conditional distribution in (2.1) with various scenarios. We considered two different class probabilities: $\pi_1 = 0.5$ (balanced case) and $\pi_1 = 0.67$ (unbalanced case). Given the class label, we also considered two different conditional covariance matrices: $\mathbf{\Sigma} = \mathbf{\Sigma}^{(1)}$ and $\mathbf{\Sigma} = \mathbf{\Sigma}^{(2)}$, where $\Sigma^{(1)}_{jk} = I(i = j)$ (no correlation) and $\Sigma^{(2)}_{jk} = 0.5^{|j-k|}$, $j, k \leq p$ (power decaying correlation). For the conditional mean vectors, we set $\boldsymbol{\mu}_1 = \mathbf{0}_p$ and $\boldsymbol{\mu}_2 = \mathbf{\Sigma}\boldsymbol{\beta}^{\text{Bayes}}$, where $\beta_j^{\text{Bayes}} = \alpha(-1)^j I(j \leq q), j \leq p$ and α was set for the Bayes misclassification error rate to be 0.2 for all scenarios.

We set $n \in \{300, 600, 900\}$, $p \in \{1000, 2000\}$, and $q \in \{5, 10\}$ to compare finite sample performance of MCL with LASSO, including MCP as a special case of MCL. We considered three cases for MCL: MCL_k with $a = 2.1$ and $\gamma = k\hat{\lambda}^{\text{opt}}, k \in \{1, 2, 3\}$, where $\hat{\lambda}^{\text{opt}}$ is the best tuning parameter value for LASSO (Kwon *et al.*, 2015). We also considered three cases of MCP: MCP_k with $a = k + 0.1, k \in \{1, 2, 3\}$ and $\gamma = 0$.

Table 2: Averages of the four measures when $\mathbf{\Sigma} = \mathbf{\Sigma}^{(2)}$

		$p = 1000, q = 5$						$p = 2000, q = 10$					
		$\pi_1 = 0.50$			$\pi_1 = 0.67$			$\pi_1 = 0.50$			$\pi_1 = 0.67$		
	n	300	600	900	300	600	900	300	600	900	300	600	900
TPS	Bayes	5	5	5	5	5	5	10	10	10	10	10	10
	Lasso	4.960	5.000	5.000	4.750	5.000	5.000	7.945	9.970	10.000	5.310	9.785	9.995
	MCL_1	4.860	5.000	5.000	4.430	4.990	5.000	6.145	9.815	9.985	4.100	8.915	9.940
	MCL_2	3.590	4.930	5.000	2.875	4.585	4.985	3.615	8.305	9.825	1.690	5.960	8.700
	MCL_3	2.225	3.125	4.435	1.785	2.585	3.395	1.620	3.460	5.385	0.355	2.410	3.425
	MCP_1	4.875	5.000	5.000	4.480	4.990	5.000	5.200	9.715	9.995	3.195	8.810	9.935
	MCP_2	4.960	5.000	5.000	4.850	5.000	5.000	6.675	9.960	10.000	4.570	9.635	9.995
	MCP_3	4.965	5.000	5.000	4.870	5.000	5.000	7.220	9.930	10.000	5.215	9.675	9.995
FPS	Bayes	0	0	0	0	0	0	0	0	0	0	0	0
	Lasso	42.995	43.620	44.890	43.260	49.450	47.895	64.935	89.360	90.535	39.150	92.305	97.670
	MCL_1	1.655	0.595	1.015	3.860	0.860	0.780	10.120	2.315	0.705	10.955	6.495	1.555
	MCL_2	0.780	0.165	0.150	0.675	0.390	0.140	2.360	2.040	0.595	0.860	2.585	1.680
	MCL_3	0.055	0.005	0.005	0.045	0.000	0.005	0.215	0.050	0.015	0.040	0.010	0.005
	MCP_1	0.790	0.400	0.320	1.310	0.435	0.375	1.570	1.265	0.440	1.340	2.010	0.830
	MCP_2	4.160	1.475	1.805	6.835	2.210	1.445	12.175	7.875	3.140	10.275	14.315	4.255
	MCP_3	13.330	4.360	2.815	17.530	6.785	3.180	27.175	26.075	11.665	22.630	36.215	19.085
CMI	Bayes	1	1	1	1	1	1	1	1	1	1	1	1
	Lasso	0	0	0	0	0	0	0	0	0	0	0	0
	MCL_1	0.340	0.665	0.630	0.110	0.605	0.620	0.000	0.205	0.605	0.000	0.010	0.340
	MCL_2	0.060	0.800	0.895	0.010	0.435	0.890	0.000	0.025	0.505	0.000	0.000	0.090
	MCL_3	0.000	0.080	0.605	0.000	0.015	0.145	0.000	0.000	0.015	0.000	0.000	0.000
	MCP_1	0.475	0.665	0.755	0.210	0.690	0.735	0.000	0.295	0.700	0.000	0.080	0.475
	MCP_2	0.100	0.480	0.595	0.050	0.400	0.555	0.000	0.030	0.255	0.000	0.000	0.095
	MCP_3	0.000	0.105	0.410	0.005	0.045	0.215	0.000	0.000	0.000	0.000	0.000	0.000
ERR	Bayes	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	Lasso	0.2600	0.2283	0.2215	0.2662	0.2326	0.2199	0.3353	0.2536	0.2317	0.3346	0.2681	0.2387
	MCL_1	0.2364	0.2176	0.2157	0.2465	0.2171	0.2125	0.3191	0.2283	0.2153	0.3234	0.2480	0.2176
	MCL_2	0.2896	0.2455	0.2302	0.2800	0.2503	0.2306	0.3570	0.2881	0.2504	0.3267	0.2909	0.2631
	MCL_3	0.3163	0.2988	0.2754	0.2965	0.2795	0.2717	0.4002	0.3480	0.3313	0.3320	0.3127	0.3048
	MCP_1	0.2214	0.2125	0.2121	0.2298	0.2105	0.2085	0.3044	0.2175	0.2095	0.3081	0.2311	0.2097
	MCP_2	0.2192	0.2127	0.2121	0.2226	0.2107	0.2085	0.2843	0.2134	0.2094	0.3030	0.2228	0.2081
	MCP_3	0.2272	0.2135	0.2120	0.2314	0.2122	0.2090	0.2903	0.2199	0.2114	0.3095	0.2307	0.2115

TPS = number of true positive selection; FPS = number of false positive selection; CMI = correct model identification; ERR = classification error rate.

For each method, we used n training samples for direction vector estimation and n independent validation samples for selecting the tuning parameter λ. We checked the selection performance by counting the number of true positive selection (TPS), the number of false positive selection (FPS) and indicator of correct model identification (CMI): TPS $= \sum_{j \le q} I(\hat{\beta}_j \neq 0, \beta_j^{\text{Bayes}} \neq 0)$, FPS $= \sum_{j > q} I(\hat{\beta}_j \neq 0, \beta_j^{\text{Bayes}} = 0)$, and CMI $= I(\text{TPS} = q, \text{FPS} = p - q)$. Furthermore, we compared the misclassification error rate (ERR) obtained from $2n$ independent test samples by using the classification rule in (2.8). We repeated each simulation 200 times and summarized the averages of the four measures in Table 1 and Table 2, including graphical illustrations in Figure 2 and Figure 3.

Table 1 shows the results when $\mathbf{\Sigma} = \mathbf{\Sigma}^{(1)}$. First, TPS converged to q as n increased for all methods while LASSO performed best for each fixed n. Second, FPS decreased as n increased only for MCL and MCP, leading to an increase in CMI. Hence, our theoretical results were supported by the results since MCL and MCP correctly fitted the model as n increased while LASSO overfitted the model regardless of sample size. These trends continued even under the class imbalance and dimension increase. Third, ERR tended to decrease as n increased for all methods. Especially MCL_1 attained the

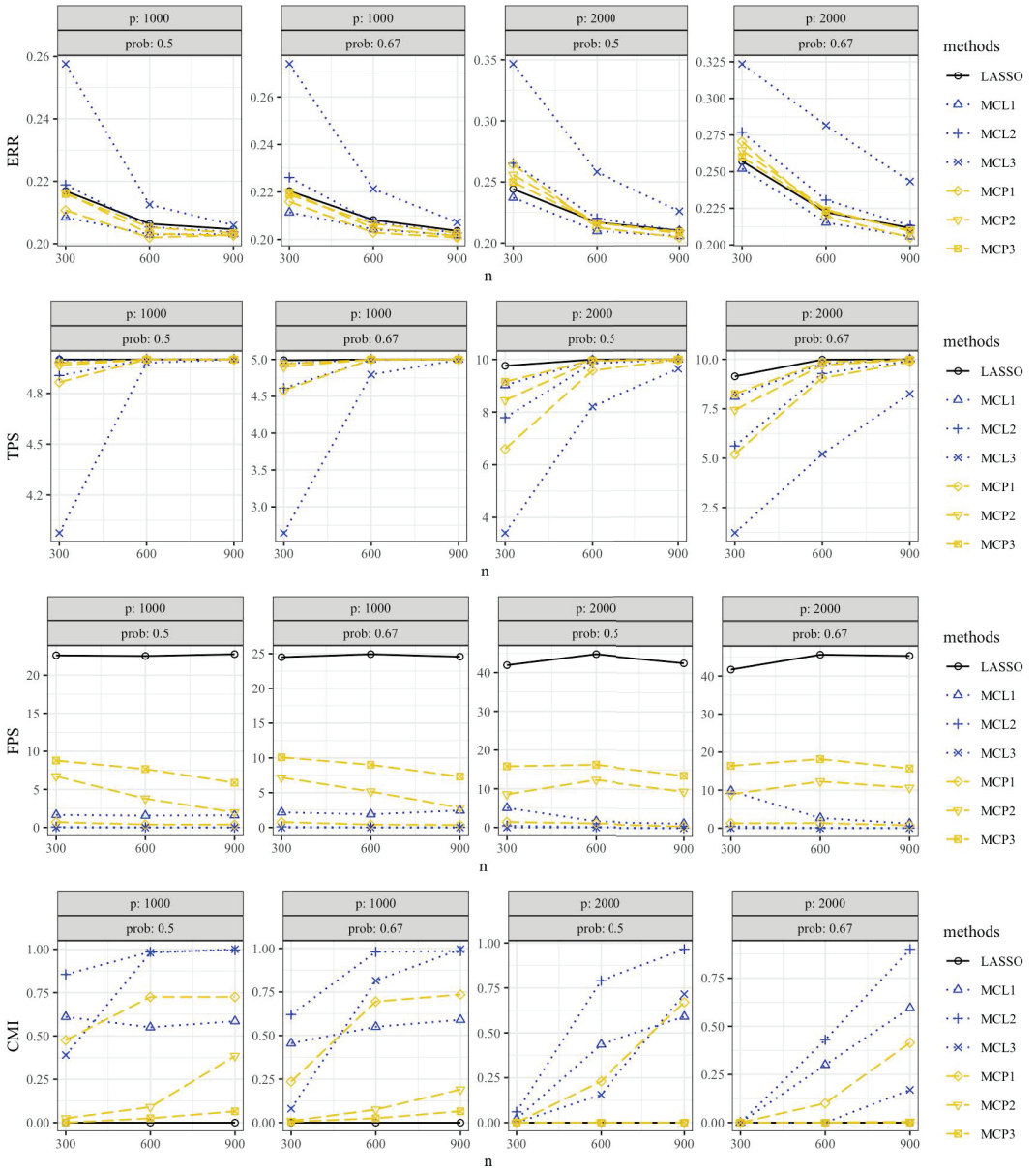

Figure 2: *Averages of the four measures when* $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(1)}$.

lowest ERR for each fixed n and was followed by MCL_2 and LASSO.

Table 2 shows the results when $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(2)}$, where MCPs performed the best for all measures. All selection performance measures tended to be undermined but the patterns were similar in that MCL and MCP fitted the model correctly while LASSO overfitted. For ERR, MCPs performed the best and MCL_1 ranked the second for all cases.

We conclude that MCL can be a nice alternative to LASSO for the high-dimensional penalized

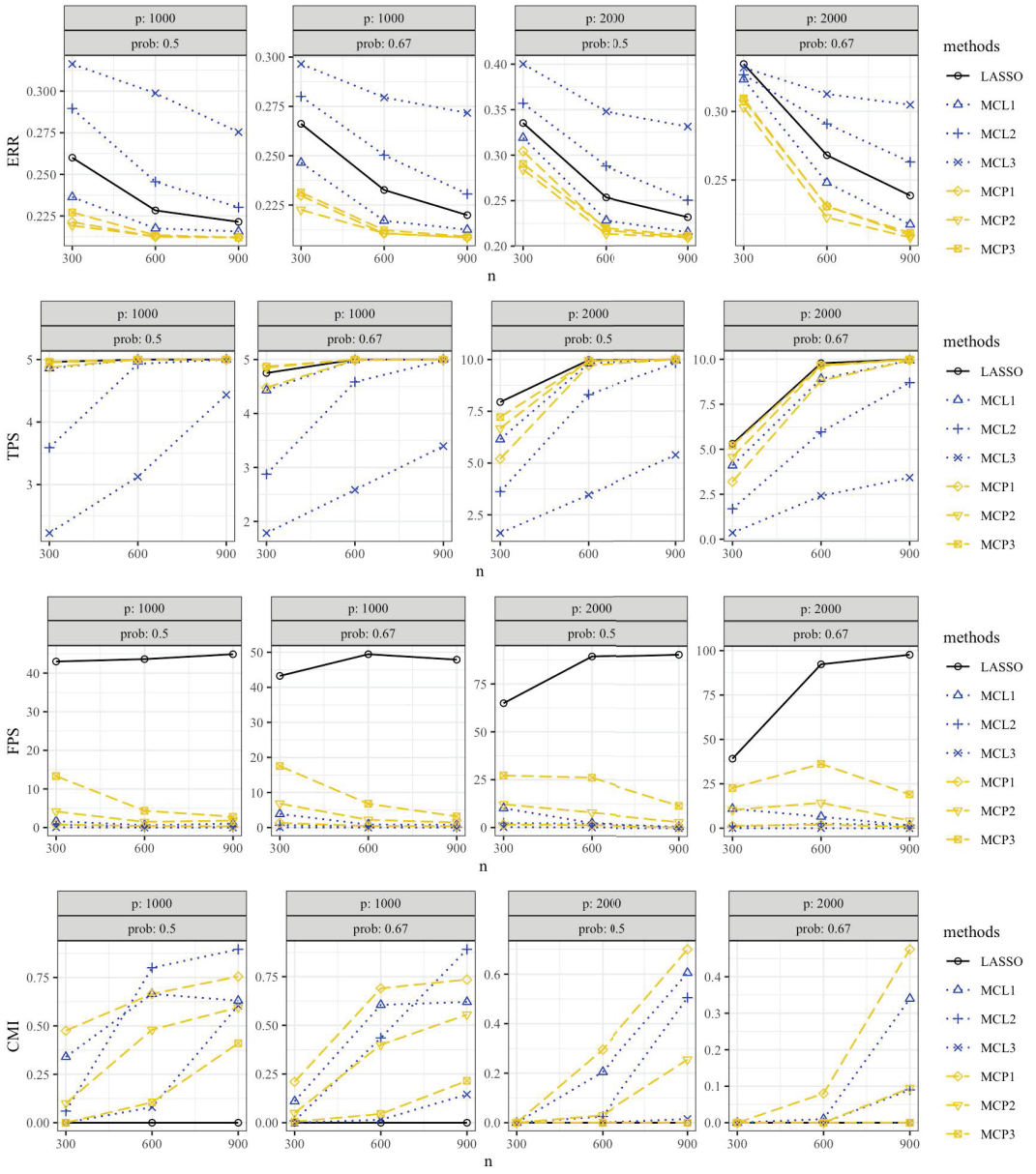

Figure 3: *Averages of the four measures when* $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(2)}$.

LDA. The MCL can correctly identify the sparse Bayes direction vector, keeping almost the same prediction accuracy as LASSO. The MCL_1 performed quite well regardless of the simulation designs considered in this paper, which aligns with the recommendation for the heuristic choice of $\gamma = \hat{\lambda}^{\text{opt}}$ in the linear regression (Kwon *et al.*, 2015).

Table 3: number of incorrectly classified samples and average of the model sizes

		d	LASSO	MCL_1	MCL_2	MCL_3	MCP_1	MCP_2	MCP_3
Chowdary *et al.* (2006), $n = 104$, $p = 22283$	Errors	400	1	1	2	2	4	4	5
		800	2	2	2	4	3	6	3
		1600	3	4	2	2	7	7	6
	Sizes	400	36.28	28.50	24.78	21.84	11.44	11.60	12.33
		800	35.20	25.46	23.10	20.17	10.10	10.78	13.38
		1600	37.29	27.85	24.75	21.06	9.31	10.63	10.53
Gordon *et al.* (2002), $n = 181$, $p = 12533$	Errors	400	2	2	2	1	3	3	3
		800	2	2	2	2	2	2	2
		1600	2	2	3	3	3	3	2
	Sizes	400	41.28	37.33	24.41	16.71	7.17	6.79	7.03
		800	44.31	37.72	22.46	15.33	11.71	12.22	13.08
		1600	49.06	42.38	28.46	25.50	19.31	16.50	15.75
Burczynski *et al.* (2006), $n = 127$, $p = 22283$	Errors	400	6	6	7	9	15	13	15
		800	8	9	13	16	13	14	17
		1600	8	9	10	13	17	16	16
	Sizes	400	47.67	41.50	26.19	24.06	13.70	13.81	13.13
		800	50.04	43.92	25.45	19.13	12.02	11.81	10.75
		1600	49.35	43.69	22.50	18.59	17.14	17.83	14.77
Chin *et al.* (2006), $n = 118$, $p = 22215$	Errors	400	12	11	13	15	21	22	22
		800	15	16	16	18	21	21	20
		1600	14	18	15	14	13	13	13
	Sizes	400	33.62	27.28	21.53	15.16	9.74	9.69	9.61
		800	41.48	31.98	23.98	16.36	5.22	5.04	4.97
		1600	33.22	18.94	13.81	12.80	4.00	4.11	4.80

4.2. Analysis of micro-array samples

The R package datamicroarray (John, 2016) provides a collection of high-dimensional microarray data sets. We chose four data sets (Burczynski *et al.*, 2006; Chin *et al.*, 2006; Chowdary *et al.*, 2006; Gordon *et al.*, 2002) to illustrate how MCL performs with real samples. We first applied the sure independence screening procedure (Fan and Song, 2010) so that the first top $d \in \{400, 800, 1600\}$ predictive variables with the largest marginal regression coefficients were used for the penalized LDA. For comparison, we applied the leave-one-out cross-validation procedure for each data set and calculated the number of incorrectly classified samples (error) as well as the average of the numbers of non-zero regression coefficients (sizes), where the tuning parameters were selected by the 10-fold cross-validation procedure.

Table 3 summarizes the results. In most cases, LASSO had the best prediction accuracy selecting the most variables. The MCPs had the worst prediction accuracy but selected the least variables. The best prediction accuracy occurred when $d = 400$ for all cases where MCL_1 had the same prediction accuracy as LASSO while selecting fewer variables than LASSO. In addition, MCL_2 selected fewer variables than LASSO without losing prediction accuracy much.

5. Concluding remarks

In this paper, we studied the high dimensional penalized LDA with MCL. The nature of MCL produces the same shrinkage effect as LASSO and makes the selection process the same as MCP. Therefore MCL shows similar or better prediction accuracy than LASSO correctly recovering the sparsity of the direction vector. We proved that MCL is selection consistent under reasonable regularity conditions, which was supported by various numerical experiments. One disadvantage of MCL compared

with LASSO may be the additional tuning parameter γ. But the heuristic choice of $\gamma = \hat{\gamma}^{\text{opt}}$ or $\gamma = 2\hat{\gamma}^{\text{opt}}$ performed well as shown in the numerical studies. Further research on MCL could focus on the theoretical justification of the choice of γ, which is not addressed in this paper.

Acknowledgement

We gratefully acknowledge the helpful comments of the Associate Editor and referees that substantially improved the paper. This paper was supported by Konkuk University in 2019.

Appendix:

Proof of Lemma 2: Assume that there exists another local minimizer $\tilde{\boldsymbol{\beta}} \in \Xi^{\kappa}_{\lambda,\gamma}$ with $\tilde{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}$. Let $\mathcal{D} = \mathcal{S} \cup \tilde{\mathcal{S}}$ where $\tilde{\mathcal{S}} = \text{supp}(\tilde{\boldsymbol{\beta}})$. Since $|\mathcal{D}| \leq 2\kappa$, $\lambda_{\min}(\mathbf{Z}_{\mathcal{D}}^T \mathbf{Z}_{\mathcal{D}}/n) > 1/a$ by the sparse Riesz condition. By replacing $\mathbf{Z}$ with $\mathbf{Z}_{\mathcal{D}}$ in Lemma 1, we can see that $\tilde{\boldsymbol{\beta}}$ cannot be a local minimizer. □

We give technical details for the proof of Theorem 1. For any subset $\mathcal{S} \subset \{1, \ldots, p\}$, let $\hat{\boldsymbol{\Omega}} = \mathbf{Z}^T\mathbf{Z}/n$, $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$, $\hat{\boldsymbol{\Delta}}_{\mathcal{S}\mathcal{S}} = \hat{\boldsymbol{\Omega}}_{\mathcal{S}\mathcal{S}} - \boldsymbol{\Omega}_{\mathcal{S}\mathcal{S}}$, $\hat{\boldsymbol{\Gamma}}_{\mathcal{S}\mathcal{S}} = \hat{\boldsymbol{\Omega}}^{-1}_{\mathcal{S}\mathcal{S}} - \boldsymbol{\Omega}^{-1}_{\mathcal{S}\mathcal{S}}$, and $\hat{\boldsymbol{\Lambda}}_{\mathcal{S}^c\mathcal{S}} = \hat{\boldsymbol{\Omega}}_{\mathcal{S}^c\mathcal{S}}\hat{\boldsymbol{\Omega}}^{-1}_{\mathcal{S}\mathcal{S}} - \boldsymbol{\Omega}_{\mathcal{S}^c\mathcal{S}}\boldsymbol{\Omega}^{-1}_{\mathcal{S}\mathcal{S}}$.

Lemma 4. *(Mai et al., 2012) There exist positive constants c_0 and $c_i, i \leq 4$ such that*

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

for any $\epsilon \leq c_0$ and subset $\mathcal{S} \subset \{1, \ldots, p\}$, where $r = |\mathcal{S}|$.

Lemma 5. *Under (C1) and (C2),*

$$\lim_{n\to\infty} \mathbf{P}\left(\hat{\rho}^{\text{src}}_{\kappa} \leq \frac{1}{a}\right) = 0. \tag{A.1}$$

Proof of Lemma 5: For any subset $\mathcal{S} \subset \{1, \ldots, p\}$,

$$\lambda_{\min}\left(\hat{\boldsymbol{\Omega}}_{\mathcal{S}\mathcal{S}}\right) = \inf_{\|\mathbf{u}\|=1} \left\{\mathbf{u}^T\boldsymbol{\Omega}_{\mathcal{S}\mathcal{S}}\mathbf{u} - \mathbf{u}^T\left(\boldsymbol{\Omega}_{\mathcal{S}\mathcal{S}} - \hat{\boldsymbol{\Omega}}_{\mathcal{S}\mathcal{S}}\right)\mathbf{u}\right\} \geq \lambda_{\min}(\boldsymbol{\Omega}_{\mathcal{S}\mathcal{S}}) - \left\|\boldsymbol{\Omega}_{\mathcal{S}\mathcal{S}} - \hat{\boldsymbol{\Omega}}_{\mathcal{S}\mathcal{S}}\right\|_2,$$

where $\|\mathbf{A}\|_2$ denotes the spectral norm of a matrix $\mathbf{A}$. By Lemma 4,

$$\begin{aligned}
\mathbf{P}\left(\hat{\rho}^{\text{src}}_{\kappa} \leq \frac{1}{a}\right) &\leq \mathbf{P}\left(\max_{|\mathcal{S}|\leq 2\kappa} \left\|\hat{\Delta}_{\mathcal{S}\mathcal{S}}\right\|_2 \geq \rho^{\text{src}}_{\kappa} - \frac{1}{a}\right) \\
&\leq \sum_{|\mathcal{S}|\leq 2\kappa} \mathbf{P}\left(\left\|\hat{\Delta}_{\mathcal{S}\mathcal{S}}\right\|_\infty \geq \rho^{\text{src}}_{\kappa} - \frac{1}{a}\right) \\
&\leq (2\kappa)p^{2\kappa}2(2\kappa)^2 \exp\left\{-c_2 n(2\kappa)^{-2}\left(\rho^{\text{src}}_{\kappa} - \frac{1}{a}\right)^2\right\}.
\end{aligned}$$

Hence (A.1) follows if $\rho^{\text{src}}_{\kappa} > 1/a$, $n/\kappa^2 \to \infty$, and $n/\kappa^2 \gg \kappa \log p$. □

Lemma 6. *Under (C1) and (C2),*

$$\lim_{n\to\infty} \mathbf{P}\left(\min_{j\in\mathcal{A}} \left|\hat{\beta}_j^{oL,\gamma}\right| \le a(\lambda-\gamma)\right) = 0. \tag{A.2}$$

Proof of Lemma 6: From the first order necessary conditions of the oracle LASSO,

$$\begin{cases} \hat{\boldsymbol{\theta}}_j - \dfrac{\mathbf{Z}_j^T \mathbf{Z}\hat{\boldsymbol{\beta}}^{oL,\gamma}}{n} = \gamma \text{sign}\left(\hat{\beta}_j^{oL,\gamma}\right), & j \in \left\{j; \hat{\beta}_j^{oL,\gamma} \neq 0\right\}, \\ \left|\hat{\boldsymbol{\theta}}_j - \dfrac{\mathbf{Z}_j^T \mathbf{Z}\hat{\boldsymbol{\beta}}^{oL,\gamma}}{n}\right| \le \gamma, & j \in \left\{j; \hat{\beta}_j^{oL,\gamma} = 0\right\}, \end{cases} \tag{A.3}$$

which implies that

$$\left\|\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \hat{\boldsymbol{\Omega}}_{\mathcal{A}\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma}\right\|_\infty \le \gamma. \tag{A.4}$$

By the triangular inequality,

$$\left|\hat{\beta}_j^{oL,\gamma}\right| \ge \left|\tilde{\beta}_j^{\text{Bayes}}\right| - \left|\hat{\beta}_j^{oL,\gamma} - \tilde{\beta}_j^{\text{Bayes}}\right| \ge m_{\mathcal{A}} - \left\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma} - \tilde{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{Bayes}}\right\|_\infty,$$

for $j \in \mathcal{A}$. Since $\boldsymbol{\beta}_{\mathcal{A}}^* = \boldsymbol{\Omega}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{\theta}_{\mathcal{A}}$,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma} - \tilde{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{Bayes}} &= \hat{\boldsymbol{\Omega}}_{\mathcal{A}\mathcal{A}}^{-1}\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\Omega}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{\theta}_{\mathcal{A}} - \hat{\boldsymbol{\Omega}}_{\mathcal{A}\mathcal{A}}^{-1}\left(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \hat{\boldsymbol{\Omega}}_{\mathcal{A}\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma}\right) \\ &= \hat{\boldsymbol{\Gamma}}_{\mathcal{A}\mathcal{A}}\hat{\boldsymbol{\delta}}_{\mathcal{A}} + \boldsymbol{\Omega}_{\mathcal{A}\mathcal{A}}^{-1}\hat{\boldsymbol{\delta}}_{\mathcal{A}} + \hat{\boldsymbol{\Gamma}}_{\mathcal{A}\mathcal{A}}\boldsymbol{\theta}_{\mathcal{A}} - \hat{\boldsymbol{\Omega}}_{\mathcal{A}\mathcal{A}}^{-1}\left(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \hat{\boldsymbol{\Omega}}_{\mathcal{A}\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma}\right). \end{aligned}$$

Hence (A.4) implies that there exist positive constants $d_i, i \le 5$ such that

$$\begin{aligned} \left\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma} - \tilde{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{Bayes}}\right\|_\infty &\le \left\|\hat{\boldsymbol{\Gamma}}_{\mathcal{A}\mathcal{A}}\right\|_\infty \left\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\right\|_\infty + \left\|\boldsymbol{\Omega}_{\mathcal{A}\mathcal{A}}^{-1}\right\|_\infty \left\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\right\|_\infty + \left\|\hat{\boldsymbol{\Gamma}}_{\mathcal{A}\mathcal{A}}\right\|_\infty \|\boldsymbol{\theta}_{\mathcal{A}}\|_\infty \\ &\quad + \left\|\hat{\boldsymbol{\Gamma}}_{\mathcal{A}\mathcal{A}}\right\|_\infty \left\|\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \hat{\boldsymbol{\Omega}}_{\mathcal{A}\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma}\right\|_\infty + \left\|\boldsymbol{\Omega}_{\mathcal{A}\mathcal{A}}^{-1}\right\|_\infty \left\|\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \hat{\boldsymbol{\Omega}}_{\mathcal{A}\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma}\right\|_\infty \\ &\le d_1 \left\|\hat{\boldsymbol{\Gamma}}_{\mathcal{A}\mathcal{A}}\right\|_\infty + d_2 \left\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\right\|_\infty + d_3\gamma, \end{aligned}$$

for any $\|\hat{\boldsymbol{\Gamma}}_{\mathcal{A}\mathcal{A}}\|_\infty \le d_4$ and $\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_\infty \le d_5$. Under the events $\|\hat{\boldsymbol{\Gamma}}_{\mathcal{A}\mathcal{A}}\|_\infty \le d_4$ and $\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_\infty \le d_5$, there exist positive constants d_6 and d_7 such that

$$\begin{aligned} \mathbf{P}\left(\min_{j\in\mathcal{A}} \left|\hat{\beta}_j^{oL,\gamma}\right| \le a(\lambda-\gamma)\right) &\le \mathbf{P}\left(m_{\mathcal{A}} - \left\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma} - \tilde{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{Bayes}}\right\|_\infty \le a(\lambda-\gamma)\right) \\ &\le \mathbf{P}\left(d_1 \left\|\hat{\boldsymbol{\Gamma}}_{\mathcal{A}\mathcal{A}}\right\|_\infty + d_2 \left\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\right\|_\infty \ge m_{\mathcal{A}} - a(\lambda-\gamma) - d_3\gamma\right) \\ &\le \mathbf{P}\left(\left\|\hat{\boldsymbol{\Gamma}}_{\mathcal{A}\mathcal{A}}\right\|_\infty \ge d_6 m_{\mathcal{A}}\right) + \mathbf{P}\left(\left\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\right\|_\infty \ge d_7 m_{\mathcal{A}}\right) \\ &\le 2q^2 \exp\left(-c_4 n q^{-2}(d_6 m_{\mathcal{A}})^2\right) + 2q \exp\left(-c_1 n (d_7 m_{\mathcal{A}})^2\right) \end{aligned}$$

by Lemma 4 for all sufficiently large n since $m_{\mathcal{A}} \gg \lambda \gg \gamma$. Hence (A.2) follows if $nm_{\mathcal{A}}^2/q^2 \to \infty$ and $nm_{\mathcal{A}}^2 \gg \log q$. □

Lemma 7. *Under (C1) and (C2),*

$$\lim_{n\to\infty} \mathbf{P}\left(\left\|\hat{\boldsymbol{\theta}}_{\mathcal{N}} - \hat{\boldsymbol{\Omega}}_{\mathcal{N}\mathcal{A}}\hat{\boldsymbol{\beta}}^{oL,\gamma}_{\mathcal{A}}\right\|_\infty \le \lambda\right) = 0. \tag{A.5}$$

Proof of Lemma 7: From Lemma 5 and Lemma 6, $\hat{\boldsymbol{\Omega}}_{\mathcal{A}\mathcal{A}}$ is invertible and $\mathrm{supp}(\hat{\boldsymbol{\beta}}^{oL,\gamma}) = \mathcal{A}$ with probability tending to one, respectively. Under these two events,

$$\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \mathbf{Z}^T_{\mathcal{A}}\mathbf{Z}_{\mathcal{A}}\hat{\boldsymbol{\beta}}^{oL,\gamma} = \gamma\hat{\mathbf{s}}_{\mathcal{A}}$$

from (A.3) where $\hat{\mathbf{s}}_{\mathcal{A}} = \mathrm{sign}(\hat{\boldsymbol{\beta}}^{oL,\gamma})$. By using $\boldsymbol{\theta}_{\mathcal{N}} = \boldsymbol{\Omega}_{\mathcal{N}\mathcal{A}}\boldsymbol{\Omega}^{-1}_{\mathcal{A}\mathcal{A}}\boldsymbol{\theta}_{\mathcal{A}}$,

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{\mathcal{N}} - \hat{\boldsymbol{\Omega}}_{\mathcal{N}\mathcal{A}}\hat{\boldsymbol{\beta}}^{oL,\gamma}_{\mathcal{A}} &= \hat{\boldsymbol{\delta}}_{\mathcal{N}} + \boldsymbol{\theta}_{\mathcal{N}} - \hat{\boldsymbol{\Omega}}_{\mathcal{N}\mathcal{A}}\hat{\boldsymbol{\Omega}}^{-1}_{\mathcal{A}\mathcal{A}}\left(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \gamma\hat{\mathbf{s}}_{\mathcal{A}}\right) \\
&= \hat{\boldsymbol{\delta}}_{\mathcal{N}} + \boldsymbol{\Omega}_{\mathcal{N}\mathcal{A}}\boldsymbol{\Omega}^{-1}_{\mathcal{A}\mathcal{A}}\boldsymbol{\theta}_{\mathcal{A}} - \hat{\boldsymbol{\Omega}}_{\mathcal{N}\mathcal{A}}\hat{\boldsymbol{\Omega}}^{-1}_{\mathcal{A}\mathcal{A}}\left(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \gamma\hat{\mathbf{s}}_{\mathcal{A}}\right)
\end{aligned}$$

which implies that there exist positive constants $e_i, i \le 5$ such that

$$\begin{aligned}
\left\|\hat{\boldsymbol{\theta}}_{\mathcal{N}} - \hat{\boldsymbol{\Omega}}_{\mathcal{N}\mathcal{A}}\hat{\boldsymbol{\beta}}^{oL,\gamma}_{\mathcal{A}}\right\|_\infty &\le \left\|\hat{\boldsymbol{\delta}}_{\mathcal{N}}\right\|_\infty + \left\|\boldsymbol{\Omega}_{\mathcal{N}\mathcal{A}}\boldsymbol{\Omega}^{-1}_{\mathcal{A}\mathcal{A}}\boldsymbol{\theta}_{\mathcal{A}} - \hat{\boldsymbol{\Omega}}_{\mathcal{N}\mathcal{A}}\hat{\boldsymbol{\Omega}}^{-1}_{\mathcal{A}\mathcal{A}}\left(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \gamma\hat{\mathbf{s}}_{\mathcal{A}}\right)\right\|_\infty \\
&\le \left\|\hat{\boldsymbol{\delta}}_{\mathcal{N}}\right\|_\infty + \left\|\hat{\boldsymbol{\Lambda}}_{\mathcal{N}\mathcal{A}}\right\|_\infty \left\|\hat{\boldsymbol{\delta}}_{\mathcal{A}} + \gamma\hat{\mathbf{s}}_{\mathcal{A}}\right\|_\infty + \left\|\boldsymbol{\Omega}_{\mathcal{N}\mathcal{A}}\boldsymbol{\Omega}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_\infty \left\|\hat{\boldsymbol{\delta}}_{\mathcal{A}} + \gamma\hat{\mathbf{s}}_{\mathcal{A}}\right\|_\infty + \|\boldsymbol{\theta}_{\mathcal{A}}\|_\infty \left\|\hat{\boldsymbol{\Lambda}}_{\mathcal{N}\mathcal{A}}\right\|_\infty \\
&\le e_1\left\|\hat{\boldsymbol{\delta}}\right\|_\infty + e_2\left\|\hat{\boldsymbol{\Lambda}}_{\mathcal{N}\mathcal{A}}\right\|_\infty + e_3\gamma,
\end{aligned}$$

for any $\|\hat{\boldsymbol{\Lambda}}_{\mathcal{N}\mathcal{A}}\|_\infty \le e_4$ and $\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_\infty \le e_5$. Under the events $\|\hat{\boldsymbol{\Lambda}}_{\mathcal{N}\mathcal{A}}\|_\infty \le e_4$ and $\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_\infty \le e_5$, there exist positive constants e_6 and e_7 such that

$$\begin{aligned}
\mathbf{P}\left(\left\|\hat{\boldsymbol{\theta}}_{\mathcal{N}} - \hat{\boldsymbol{\Omega}}_{\mathcal{N}\mathcal{A}}\hat{\boldsymbol{\beta}}^{oL,\gamma}_{\mathcal{A}}\right\|_\infty > \lambda\right) &\le \mathbf{P}\left(e_1\left\|\hat{\boldsymbol{\delta}}\right\|_\infty + e_2\left\|\hat{\boldsymbol{\Lambda}}_{\mathcal{N}\mathcal{A}}\right\|_\infty > \lambda - e_3\gamma\right) \\
&\le \mathbf{P}\left(\left\|\hat{\boldsymbol{\delta}}\right\|_\infty > e_6\lambda\right) + \mathbf{P}\left(\left\|\hat{\boldsymbol{\Lambda}}_{\mathcal{N}\mathcal{A}}\right\|_\infty > e_7\lambda\right) \\
&\le 2p\exp\left(-c_1 n(e_6\lambda)^2\right) + 2(p-q)q\exp\left(-c_3 n q^{-2}(e_7\lambda)^2\right)
\end{aligned}$$

by Lemma 4 for all sufficiently large n since $\lambda \gg \gamma$. Hence (A.5) follows if $n\lambda^2/q^2 \to \infty$ and $n\lambda^2 \gg q^2 \log p$. □

Proof of Theorem 1: It suffices to show that the sufficient conditions hold with probability tending to one, which follows from Lemmas 5, 6, and 7. □

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.

Burczynski ME, Peterson RL, Twine NC, *et al.* (2006). Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells, *The Journal of Molecular Diagnostics*, **8**,51–61.

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

Casella G (1985). An introduction to empirical bayes data analysis, *The American Statistician*, **39**, 83–87.

Chin K, DeVries S, Fridlyand J, *et al.* (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies, *Cancer Cell*, **10**, 529–541.

Chowdary D, Lathrop J, Skelton J, *et al.* (2006). Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative, *The Journal of Molecular Diagnostics*, **8**, 31–39.

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

Efron B and Morris C (1975). Data analysis using stein's estimator and its generalizations, *Journal of the American Statistical Association*, **70**, 311–319.

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

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 Song R (2010). Sure independence screening in generalized linear models with np-dimensionality, *The Annals of Statistics*, **38**, 3567–3604.

Fan J, Xue L, and Zou H (2014). Strong oracle optimality of folded concave penalized estimation, *Annals of Statistics*, **42**, 819.

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.

Guo Y, Hastie T, and Tibshirani R (2006). Regularized linear discriminant analysis and its application in microarrays, *Biostatistics*, **8**, 86–100.

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

John AR (2016). datamicroarray: Collection of Data Sets for Classification. Available from: https://github.com/ramhiser/datamicroarray.

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

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

Kim Y, Jeon JJ, 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.

Krzanowski W, Jonathan P, McCarthy W, and Thomas M (1995). Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **44**, 101–115.

Kwon S, Lee S, and Kim Y (2015). Moderately clipped lasso, *Computational Statistics & Data Analysis*, **92**, 53–67.

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

Mazumder R, Friedman JH, and Hastie T (2011). Sparsenet: Coordinate descent with nonconvex

penalties, *Journal of the American Statistical Association*, **106**, 1125–1138.

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.

Tibshirani R (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267–288.

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.

Wu MC, Zhang L, Wang Z, Christiani DC, and Lin X (2009). Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection, *Bioinformatics*, **25**, 1145–1151.

Yuille AL and Rangarajan A (2002). The concave-convex procedure (cccp). In *Advances in Neural Information Processing Systems*, pages 1033–1040.

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

Zhang CH and Huang J (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression, *The Annals of Statistics*, **36**, 1567–1594.

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

Zou H (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418–1429.

Received August 18, 2020; Revised October 29, 2020; Accepted November 19, 2020