Journal of Machine Learning Research 24 (2023) 1-45 Submitted 9/22; Revised 6/23; Published 9/23

Pivotal Estimation of Linear Discriminant Analysis in High Dimensions

Ethan X. Fang[†] ETHAN.FANG@DUKE.EDU
Yajun Mei[‡] YMEI3@GATECH.EDU
Yuyang Shi[‡] YYSHI@GATECH.EDU
Qunzhi Xu[‡] XUQUNZHI@GATECH.EDU
Tuo Zhao[‡] TOURZHAO@GATECH.EDU
† *Department of Biostatistics and Bioinformatics, Duke University*
‡ *School of Industrial and Systems Engineering, Georgia Tech*

Editor: Mladen Kolar

Abstract

We consider the linear discriminant analysis problem in the high-dimensional settings. In this work, we propose PANDA(PivotAl liNear Discriminant Analysis), a tuning-insensitive method in the sense that it requires very little effort to tune the parameters. Moreover, we prove that PANDA achieves the optimal convergence rate in terms of both the estimation error and misclassification rate. Our theoretical results are backed up by thorough numerical studies using both simulated and real datasets. In comparison with the existing methods, we observe that our proposed PANDA yields equal or better performance, and requires substantially less effort in parameter tuning.

Keywords: Linear classification; Sparsity; Tuning-insensitive; Convex optimization.

1. Introduction

We consider the linear discriminant analysis problem with n_0 samples $(X_i^{(0)})_{i=1}^{n_0}$ from class 0 and n_1 samples $(X_i^{(1)})_{i=1}^{n_1}$ from class 1. In particular, consider the Gaussian case where $X_i^{(\ell)} \sim N(\mu^{(\ell)}, \Sigma), \ell = 0, 1$. Under the ideal setting where all parameters $\mu^{(0)}, \mu^{(1)}, \Sigma$ are pre-specified, the Bayes rule classifies a new sample Z by

$$f^*(Z) = \mathbb{1}\big\{(Z - \mu_m)^\top \Sigma^{-1} \mu_d > 0\big\},$$

where $\mu_m = (\mu^{(0)} + \mu^{(1)})/2$ and $\mu_d = (\mu^{(1)} - \mu^{(0)})$, and is proved to be optimal in terms of misclassification rate, see Anderson (2003). However, the Bayes rule is often not practical, as in reality the parameters are always unknown and need to be estimated.

Under the classical low-dimensional setting $p < n$, we estimate $\mu^{(0)}$, $\mu^{(1)}$ and Σ^{-1} by their sample versions, and use the plug-in Bayes rule to classify the new sample. In particular, let $\widehat{\mu}^{(\ell)}$'s and $\widehat{\Sigma}$ be the the sample means and the pooled sample covariance matrix, and let $\widehat{\mu}_m = (\widehat{\mu}^{(0)} + \widehat{\mu}^{(1)})/2$, $\widehat{\mu}_d = (\widehat{\mu}^{(1)} - \widehat{\mu}^{(0)})$. Given a new sample Z, the following rule

$$\widehat{f}(Z) = \mathbb{1}\left\{\widehat{\mu}_d^\top \widehat{\Sigma}^{-1} (Z - \widehat{\mu}_m) > 0\right\},$$

asymptotically achieves the optimal Bayesian risk. Unfortunately, this method is inapplicable to high-dimensional settings where $p \gg n$ because it is difficult to estimate Σ^{-1} due to the singularity of $\widehat{\Sigma}$. Such high dimensionality issues exist unavoidably in many critical modern scenarios such as genomics, and it is important to develop efficient methods for LDA in high dimensions.

Several methods have been developed in the literature for high-dimensional LDA with sparsity assumptions imposed, which are common in many real-world applications such as the fMRI decoding and biomarker identification (Yamashita et al., 2008; Shi et al., 2009). The existing methods can be further divided into two tracks based on the different sparsity assumptions. The first track is to assume that Σ is sparse and estimate $\mu_d = \mu^{(1)} - \mu^{(0)}$ and Σ separately. A simple approach is the naive Bayes rule or independence rule discussed in Bickel et al. (2004). Tibshirani et al. (2002), and Fan and Fan (2008) proposed the nearest shrunken centroid method and the Features Annealed Independence Rules (FAIR) respectively for selecting significant features. Also see the sparse linear discriminant analysis (SLDA) proposed in Shao et al. (2011).

Another track of work assumes the sparsity of the discriminant direction $\beta^* = \Sigma^{-1}\mu_d$ and directly estimates β^* from the samples. Witten and Tibshirani (2011) and Clemmensen et al. (2011) proposed the sparse discriminant analysis method with multiple classes by imposing fused LASSO penalty and elastic net penalty respectively. Mai et al. (2012) proposed to estimate β^* by minimizing an ℓ_1-penalized least square loss, and Fan et al. (2012) proposed the regularized optimal affine discriminant (ROAD) method.

Existing theoretical results in the literature of high-dimensional LDA often require the knowledge of unknown population. For the better understanding, here we present the linear programming discriminant (LPD) rule in Cai and Liu (2011) with more details. The LPD rule provides an estimator $\widehat{\beta}$ for β^* by solving the following linear optimization problem

$$\widehat{\beta} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \|\beta\|_1, \quad \text{subject to} \quad \|\widehat{\Sigma}\beta - \widehat{\mu}_d\|_\infty \le \lambda\widehat{\sigma}_{\max},$$

with $\widehat{\sigma}_{\max} = \sqrt{\max_j \widehat{\Sigma}_{jj}}$ and tuning parameter λ. The authors show that to ensure the fast convergence rate of $\widehat{\beta}$, a reasonable choice of λ would be

$$\lambda = O\left(\Delta\sqrt{\frac{\log p}{n}}\right),$$

where $\Delta = \sqrt{\beta^{*\top}\Sigma\beta^*}$. In practice, this choice of λ heavily relies on the unknown population quantity Δ, which takes substantial effort to tune. To reduce the tuning effort, Cai and Zhang (2019) propose the adaptive linear discriminant analysis (AdaLDA) rule, which is a two-stage method that achieves the minimax optimal convergence rate in both the estimation error and misclassification rate. Specifically, the AdaLDA rule solves a two-stage problem: in the first stage it constructs an estimator $\widehat{\Delta}$ for Δ and in the second stage the estimator is plugging into the LPD framework to obtain the estimator for β^*.

In this paper, we propose a novel one-stage method for high-dimensional linear discriminant analysis named **PANDA** (<u>P</u>ivot<u>A</u>l li<u>N</u>ear <u>D</u>iscriminant <u>A</u>nalysis). Our method is tuning-insensitive, in the sense that it automatically adapts to the population pattern and requires less effort to tune. Motivated by Gautier et al. (2011) for high-dimensional linear regression, the proposed PANDA method simultaneously estimates β^* and Δ by solving a single convex optimization problem, and is shown to attain the same minimax optimal convergence rate as the AdaLDA method. Moreover, our detailed numerical results show that the PANDA method achieves similar or more competitive performance than the LPD and AdaLDA methods in terms of β^* estimation error and misclassification rate, with less cost of computational time.

It is worth mentioning that the topic of variable selection has also been investigated in high-dimensional LDA. For example, Kolar and Liu (2015) established the optimal results of variable selection for sparse discriminant analysis in Mai et al. (2012) and the ROAD estimator in Fan et al. (2012), and Gaynanova and Kolar (2015) further extended the result to the multi-group sparse discriminant analysis. We also include some numerical studies investigating the variable selection properties of PANDA in Section 5.

Paper Organization. The rest of this paper is organized as follows. In Section 2, we briefly review the LDA problem and the AdaLDA rule. In Section 3, we propose the PANDA method. In Section 4, we provide theoretical justifications of PANDA. In Section 5, we present the numerical studies. In Section 6, we discuss the extension of our PANDA method to the multiple-class LDA problem. In Section 7, we provide proofs of our main results. We conclude the paper in Section 8.

Notations. Let $v = (v_1, \cdots, v_p)^\top \in \mathbb{R}^p$ be a $p-$dimensional real vector. We define the following vector norms: $\|v\|_1 = \sum_{j=1}^p |v_j|$, $\|v\|_2^2 = \sum_{j=1}^p v_j^2$, and $\|v\|_\infty = \max_{1\leq j\leq p} |v_j|$. For $p \in \mathbb{N}$, we denote by $[p]$ the set $\{1, 2, \cdots, p\}$. For $j \in [p]$, let e_j be the $j-$th canonical basis in $\mathbb{R}^p$. For $S \subseteq [p]$, let v_S denote the the subvector of v confined to S, and S^c denotes the complement of S. For a matrix $\Sigma \in \mathbb{R}^{p\times p}$, $\Sigma \succ 0$ denotes that Σ is symmetric and positive definite, and $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ denote the smallest and the largest eigenvalue of Σ, respectively. We let $\mathbf{0}$ and $\mathbf{1}$ denote vectors with all the entries equal to 0 and 1, respectively. We use $\mathbb{1}\{\cdot\}$ to denote the indicator function.

2. Background

In this section, we provide necessary mathematical background. For better presentation, we split this section into two subsections. We review the problem setup of LDA in Section 2.1, and the AdaLDA method in Section 2.2.

2.1 Problem Setup

We consider the linear discriminant analysis problem with n_0 samples $(X_i^{(0)})_{i=1}^{n_0}$ from class 0 and n_1 samples $(X_i^{(1)})_{i=1}^{n_1}$ from class 1. In particular, consider the Gaussian case where $X_i^{(\ell)} \sim N(\mu^{(\ell)}, \Sigma), \ell = 0, 1$. Our goal is to find a linear discriminant rule $f_{\alpha,\beta}(\cdot)$ such that given a new sample Z, we predict the class label of Z by

$$f_{\alpha,\beta}(Z) = \mathbb{1}\left\{\beta^\top (Z - \alpha) > 0\right\},$$

with some $\alpha, \beta \in \mathbb{R}^p$. For simplicity, we assume the two classes have equal prior weights, i.e., $\mathbb{P}(Z \text{ is from Class } 0) = \mathbb{P}(Z \text{ is from Class } 1) = 1/2$. Then the misclassification rate of

$f_{\alpha,\beta}(\cdot)$ can be written as

$$\begin{aligned}\mathcal{R}(f_{\alpha,\beta}) &= \frac{1}{2}\mathbb{P}_{Z\sim N(\mu^{(0)},\Sigma)}(f_{\alpha,\beta}(Z)=1) + \frac{1}{2}\mathbb{P}_{Z\sim N(\mu^{(1)},\Sigma)}(f_{\alpha,\beta}(Z)=0)\\ &= \frac{1}{2}\Phi\left(-\frac{\beta^\top\left(\alpha-\mu^{(0)}\right)}{\sqrt{\beta^\top\Sigma\beta}}\right) + \frac{1}{2}\Phi\left(-\frac{\beta^\top\left(\mu^{(1)}-\alpha\right)}{\sqrt{\beta^\top\Sigma\beta}}\right),\end{aligned}$$

where Φ is the CDF of the standard Gaussian distribution.

The optimal misclassification rate (also known as the Bayes error) is achieved by the Fisher's discriminant rule $f_{\alpha^*,\beta^*}(\cdot)$ with $\alpha^* = (\mu^{(0)}+\mu^{(1)})/2$ and $\beta^* = \Sigma^{-1}\left(\mu^{(1)}-\mu^{(0)}\right)$. Accordingly, the optimal misclassification rate is $\mathcal{R}^* = \Phi(-\Delta/2)$, where $\Delta = \sqrt{\beta^{*\top}\Sigma\beta^*} = \sqrt{\mu_d^\top\Sigma^{-1}\mu_d}$ is the signal-noise ratio of the classification problem.

2.2 The AdaLDA method

In this subsection, we review the AdaLDA method proposed in Cai and Zhang (2019), which is tuning-insensitive and serves as a good comparison to our method. Let the sample means and the pooled covariance matrix be

$$\widehat{\mu}^{(\ell)} = \frac{1}{n_\ell}\sum_{i=1}^{n_\ell} X_i^{(\ell)} \quad \text{and} \quad \widehat{\Sigma} = \frac{1}{n_0+n_1}\sum_{\ell=0,1}\sum_{i=1}^{n_\ell}(X_i^{(\ell)}-\widehat{\mu}^{(\ell)})(X_i^{(\ell)}-\widehat{\mu}^{(\ell)})^\top.$$

The AdaLDA method estimates β^* through two stages. In the first stage, AdaLDA solves the following linear optimization problem to obtain an initial estimator $\widetilde{\beta}$,

$$\begin{aligned}&\widetilde{\beta} \in \arg\min_{\beta} \quad \|\beta\|_1,\\ &\quad \text{subject to} \quad \|\widehat{\Sigma}\beta-\widehat{\mu}_d\|_\infty \le 4\widehat{\sigma}_{\max}\cdot\sqrt{\frac{\log p}{n}}\cdot\left(\lambda\beta^\top\widehat{\mu}_d\ +1\right),\end{aligned} \tag{1}$$

where $n=\min(n_0,n_1)$, λ is a tuning parameter, $\widehat{\mu}_d = \widehat{\mu}^{(1)}-\widehat{\mu}^{(0)}$ is the difference of the sample means, and $\widehat{\sigma}_{\max} = \sqrt{\max_j\widehat{\Sigma}_{jj}}$. The initial esstimator $\widetilde{\beta}$ is used to construct an estimator $\widehat{\Delta} = \sqrt{|\widetilde{\beta}^\top\widehat{\mu}_d|}$ for Δ. In the second stage, AdaLDA solves another linear optimization problem to obtain the final estimator $\widehat{\beta}$

$$\begin{aligned}&\widehat{\beta} \in \arg\min_{\beta} \quad \|\beta\|_1,\\ &\quad \text{subject to} \quad |e_j^\top(\widehat{\Sigma}\beta-\widehat{\mu}_d)| \le 4\widehat{\sigma}_{\max}\cdot\sqrt{\frac{\log p}{n}}\cdot\sqrt{\lambda\widehat{\Delta}^2+1}, \quad \text{for all } j\in[p].\end{aligned}$$

With $\widehat{\beta}$ and $\widehat{\mu}_m = \left(\widehat{\mu}^{(0)}+\widehat{\mu}^{(1)}\right)/2$, AdaLDA constructs the linear discriminant rule $f_{\widehat{\mu}_m,\widehat{\beta}}$.

With a slight abuse of the notation, we let $\mathcal{R}(\widehat{\beta}) = \mathcal{R}(f_{\widehat{\mu}_m,\widehat{\beta}})$. Since the tuning parameters in the two steps do not depend on any unknown population quantities, the AdaLDA method is tuning-insensitive. Assuming β^* contains at most s nonzero entries, Cai and

Zhang (2019) prove that under some mild assumptions, by choosing λ as a proper constant, both $\widehat{\beta}$ and $\mathcal{R}(\widehat{\beta})$ achieve the minimax optimal rates of convergence that

$$\|\widehat{\beta} - \beta^*\|_2 = \mathcal{O}_P\left(\Delta\sqrt{\frac{s\log p}{n}}\right) \quad \text{and} \quad \mathcal{R}(\widehat{\beta}) - \mathcal{R}^* = \mathcal{O}_P\left(\exp\left(-\frac{\Delta^2}{8}\right)\Delta\frac{s\log p}{n}\right).$$

3. The PANDA Method

In this section, we propose PANDA, a one-stage and tuning-insensitive method for linear discriminant analysis in high dimensions. To begin with, we would like to first recall the LPD method Cai and Liu (2011), which motivates our formulation. Specifically, the LPD method estimates β^* by solving the following linear optimization problem that

$$\textbf{LPD}: \quad \widehat{\beta} \in \mathop{\arg\min}_{\beta\in\mathbb{R}^p} \|\beta\|_1, \quad \text{subject to} \quad \|\widehat{\Sigma}\beta - \widehat{\mu}_d\|_\infty \le \lambda\widehat{\sigma}_{\max}. \tag{2}$$

As discussed earlier, the tuning parameter λ in (2) depends on the unknown population quantity $\Delta = \sqrt{\beta^{*\top}\Sigma\beta^*}$, which is difficult to tune in practice.

To address this issue, we introduce τ as an estimator of Δ, and plug it into (2), as inspired by the pivotal method for high-dimensional linear regression in Gautier et al. (2011). This leads to the following optimization problem

$$(\widehat{\beta}, \widehat{\tau}) \in \mathop{\arg\min}_{\beta\in\mathbb{R}^p, \tau\in\mathbb{R}} \|\beta\|_1, \quad \text{subject to} \quad \|\widehat{\Sigma}\beta - \widehat{\mu}_d\|_\infty \le \lambda\widehat{\sigma}_{\max}(\tau+1), \quad \sqrt{\beta^\top\widehat{\Sigma}\beta} = \tau. \tag{3}$$

The optimization problem in (3) is nonconvex due to the quadratic equality constraint $\sqrt{\beta^\top\widehat{\Sigma}\beta} = \tau$. Thus, we propose to relax the equality constraint into an inequality constraint, and obtain

$$(\widehat{\beta}, \widehat{\tau}) \in \mathop{\arg\min}_{\beta\in\mathbb{R}^p, \tau\in\mathbb{R}} \|\beta\|_1, \quad \text{subject to} \quad \|\widehat{\Sigma}\beta - \widehat{\mu}_d\|_\infty \le \lambda\widehat{\sigma}_{\max}(\tau+1), \quad \sqrt{\beta^\top\widehat{\Sigma}\beta} \le \tau. \tag{4}$$

However, as the objective function in (4) is free of τ, τ can be arbitrarily large. In fact, (4) admits a trivial solution $\widehat{\beta} = 0$ when $\widehat{\tau}$ is larger than $\lambda^{-1}\|\widehat{\mu}_d\|_\infty - 1$, which makes (4) inapplicable.

To solve this problem, we introduce an additional penalty term $c\tau^2$ to the objective in (4), which leads to the following PANDA's formulation:

$$\begin{aligned}\textbf{PANDA}: \quad (\widehat{\beta}, \widehat{\tau}) \in \mathop{\arg\min}_{\beta\in\mathbb{R}^p, \tau\in\mathbb{R}} \quad & \|\beta\|_1 + c\tau^2, \\ \text{subject to} \quad & \|\widehat{\Sigma}\beta - \widehat{\mu}_d\|_\infty \le \lambda\widehat{\sigma}_{\max}(\tau+1), \quad \sqrt{\beta^\top\widehat{\Sigma}\beta} \le \tau,\end{aligned} \tag{5}$$

where $c > 0$ and $\lambda > 0$ are two tuning parameters. Note that different from the linear penalty term used in Gautier's pivotal method, our penalty term is quadratic in τ. In fact, we can show that to guarantee the tuning-insensitivity of our PANDA method, the penalty term on τ must be quadratic. We provide more detailed discussion in Section F of the supplementary material.

Note that both our proposed PANDA method and the AdaLDA method adopt the similar idea of plugging in an estimator of the unknown quantity Δ to the tuning parameter λ in the LPD method to achieve tuning-insensitivity. The main difference is that AdaLDA constructs the estimator for Δ in a separate linear program (1), while PANDA estimates β^* and Δ in a single convex program.

We point out that the problem in (5) is a second order conic optimization problem. By introducing auxiliary variables $w \in \mathbb{R}^p$ and $u \in \mathbb{R}$, the problem in (5) is equivalent to the following optimization problem

$$\begin{aligned} &\min_{\beta,\tau,w,u} \quad \sum_{j=1}^{p} w_j + cu, && (6)\\ &\text{subject to} \quad -w_j \le \beta_j \le w_j, \quad -\lambda\widehat{\sigma}_{\max}(\tau+1)\mathbf{1} \le \widehat{\Sigma}\beta - \widehat{\mu}_d \le \lambda\widehat{\sigma}_{\max}(\tau+1)\mathbf{1}, \\ &\qquad \|\widehat{\Sigma}^{1/2}\beta\|_2 \le \tau, \quad \sqrt{\tau^2 + \frac{1}{4}(1-u)^2} \le \frac{1}{2}(1+u). \end{aligned}$$

Such a second order conic optimization problem is convex, and can be solved in a polynomial time using the interior point method (Nesterov and Nemirovskii, 1994). Computationally, we also provide an efficient scheme in Algorithm 1 using the alternating direction method of multipliers (ADMM) following Boyd et al. (2011) to solve (6). We provide more details on the derivation of the algorithm in Section A of the supplementary material.

Algorithm 1: ADMM with proximal method for solving problem (6)

```
Input: Sample mean difference $\widehat{\mu} = \widehat{\mu}^{(1)} - \widehat{\mu}^{(0)}$; Pooled sample covariance matrix $\widehat{\Sigma}$;
Tuning parameter $c$, $\lambda$; Initialization $\beta^0$, $\tau^0$ $u^0$, $v^0$, $w^0$, $s^0$; Penalty parameter $\rho > 0$;
Primal step size $\eta > 0$; Number of iterations $T$.
for $t = 1, 2, \cdots, T$ do
    $\beta^t \leftarrow \beta^{t-1} - \eta\nabla_\beta L_\rho(\beta^{t-1}, u^{t-1}, v^{t-1}, w^{t-1}, \tau^{t-1}, s^{t-1})$
    $u^t \leftarrow \Pi_{\mathcal{C}_1}[u^{t-1} - \eta\nabla_u L_\rho(\beta^t, u^{t-1}, v^{t-1}, w^{t-1}, \tau^{t-1}, s^{t-1})]$
    $v^t \leftarrow \Pi_{\mathcal{C}_2}[v^{t-1} - \eta\nabla_v L_\rho(\beta^t, u^t, v^{t-1}, w^{t-1}, \tau^{t-1}, s^{t-1})]$
    $\widetilde{\tau}^t \leftarrow \tau^{t-1} - \eta\nabla_\tau L_\rho(\beta^t, u^t, v^t, w^{t-1}, \tau^{t-1}, s^{t-1})$
    $\widetilde{w}^t \leftarrow w^{t-1} - \eta\nabla_w L_\rho(\beta^t, u^t, v^t, w^{t-1}, \tau^{t-1}, s^{t-1})$
    $(w^t, \tau^t) \leftarrow \Pi_{\mathcal{C}_2}(\widetilde{w}^t, \widetilde{\tau}^t)$
    $s^t \leftarrow s^{t-1} + A_\beta\beta^t + A_u u^t + A_v v^t + A_w w^t + A_\tau\tau^t - b^t$
end for
Output $(\beta^T, \tau^T, w^T, u^T)$
```

4. Statistical Properties

In this section, we establish theoretical guarantees for our proposed PANDA method. For notational simplicity, we denote

$$\mu_m = (\mu^{(0)} + \mu^{(1)})/2, \quad \widehat{\mu}_m = (\widehat{\mu}^{(0)} + \widehat{\mu}^{(1)})/2, \quad \mu_d = \mu^{(1)} - \mu^{(0)},$$
$$\widehat{\mu}_d = \widehat{\mu}^{(1)} - \widehat{\mu}^{(0)}, \quad \sigma_{\max} = \max_j(\Sigma_{jj})^{1/2}, \quad \widehat{\sigma}_{\max} = \max_j(\widehat{\Sigma}_{jj})^{1/2}.$$

Without loss of generality, here we only consider the case where $n_0 = n_1 = n$, and our results can be easily extended to the general case where $n_0 \neq n_1$. We require the following weak sparsity condition on β^* :

$$\beta^* \in \mathbb{B}_q(R) \coloneqq \left\{ \beta \in \mathbb{R}^p : \sum_j |\beta_j|^q \leq R \right\},$$

where $q \in [0, 1)$ and R can scale with n and p. Note that when $q = 0$, $\mathbb{B}_q(R)$ is reduced to the class of R-sparse vectors, i.e., $\mathbb{B}_0(R) \coloneqq \left\{ \beta \in \mathbb{R}^p : \sum_j \mathbb{1}\{\beta_j \neq 0\} \leq R \right\}$. We also need to impose the following two mild assumptions.

Assumption 1 *There exists a constant a such that $\|\mu_d\|_\infty \geq a > 0$.*

Assumption 2 *There exists some M such that $M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M$.*

Essentially, Assumption 1 requires the two classes to be distinguishable, and Assumption 2 requires the covariance matrix Σ to be sufficiently well-conditioned, as its condition number is upper bounded by M^2.

We are now ready to present the theoretical guarantees of the PANDA method in (5). Let us begin with the convergence rates of $\widehat{\beta}$ and $\widehat{\tau}$.

Theorem 1 (Parameter Estimation) *Suppose that Assumption 2 hold, and $\beta^* \in \mathbb{B}_q(R)$ for some $q \in [0, 1)$ and some $R > 0$. Let $\left(\widehat{\beta}, \widehat{\tau}\right)$ be an optimal solution of (5). Given*

$$c = \frac{1}{8\left(\|\widehat{\mu}_d\|_\infty + 4\widehat{\sigma}_{\max}\sqrt{\frac{\log p}{n}}\right)}, \qquad \lambda = 20\sqrt{\frac{\log p}{n}}, \tag{7}$$

for sufficiently large n such that

$$n \geq C^{(1)} \cdot a^{-2}\Delta^2\sigma_{\max}^2 M^{2+\frac{1}{1-q}} R^{\frac{2}{1-q}} \log p \tag{8}$$

where $C^{(1)}$ is an absolute constant, we have, with probability goes to 1,

$$\|\widehat{\beta} - \beta^*\|_1 \leq C_1 \cdot (\Delta + 1)\left(\sigma_{\max} M\right)^{1-q} R \left(\frac{\log p}{n}\right)^{(1-q)/2}, \tag{9a}$$

$$\|\widehat{\beta} - \beta^*\|_2 \leq C_2 \cdot (\Delta + 1)\left(\sigma_{\max} M\right)^{1-q/2} \sqrt{R} \left(\frac{\log p}{n}\right)^{1/2-q/4}, \tag{9b}$$

$$\frac{|\widehat{\tau}^2 - \Delta^2|}{\Delta^2} \leq C_3 \cdot (1 + \Delta^{-1})\sigma_{\max}^{1-q/2} M^{3/2-q} R \Big(\frac{\log p}{n}\Big)^{(1-q)/2}, \tag{9c}$$

where C_1, C_2 and C_3 are positive constants.

Note that our proposed PANDA method is tuning-insensitive, as the chosen tuning parameters c and λ in (7) do not depend on any unknown population quantity. In the next theorem, we show that the sample complexity requirement (8) can be relaxed under some more restrictive conditions.

Theorem 2 *Suppose that Assumption 2 holds, and $\beta^* \in \mathbb{B}_q(R)$ for some $q \in [0,1)$ and some $R > 0$. Let $\left(\widehat{\beta}, \widehat{\tau}\right)$ be an optimal solution to problem (5). When $\widehat{\tau} = \sqrt{\widehat{\beta}^\top \widehat{\Sigma}\widehat{\beta}}$, given*

$$c = \frac{1}{8\left(\|\widehat{\mu}_d\|_\infty + 4\widehat{\sigma}_{\max}\sqrt{\frac{\log p}{n}}\right)}, \quad \lambda = 20\sqrt{\frac{\log p}{n}},$$

for sufficiently large n such that

$$n \geq C^{(2)} \cdot a^{-2}\Delta^2\sigma_{\max}^2 M^{2+\frac{1}{1-q}} R^{\frac{2}{2-q}} \log p, \tag{10}$$

where $C^{(2)}$ is an absolute constant, we have, with probability goes to 1,

$$\|\widehat{\beta} - \beta^*\|_2 \leq C_1 \cdot (\Delta + 1)(\sigma_{\max} M)^{1-q/2}\sqrt{R}\left(\frac{\log p}{n}\right)^{1/2-q/4}, \tag{11a}$$

$$\frac{|\widehat{\tau}^2 - \Delta^2|}{\Delta^2} \leq C_2 \cdot (1 + \Delta^{-1})\sigma_{\max}^{1-q/2} M^{3/2-q}\sqrt{R}\Big(\frac{\log p}{n}\Big)^{(1-q)/2}, \tag{11b}$$

where C_1 and C_2 are positive constants.

Note that in the above theorem, we impose the additional assumption that $\widehat{\tau} = \sqrt{\widehat{\beta}^\top \widehat{\Sigma}\widehat{\beta}}$, i.e. the second inequality constraint of PANDA is active at the optimal solution. We point out that in practice, we can numerically verify if this assumption indeed holds. Also, in our later simulations, we find that this assumption holds when the tuning parameters are properly chosen.

We next compare our results with Cai and Zhang (2019) for $q = 0$. Note that Cai and Zhang (2019) consider the following parameter space of β^* and Σ,

$$\begin{aligned}\Theta_s = \big\{&(\beta^*, \Sigma) : \beta^* \in \mathbb{R}^p,\ \Sigma \in \mathbb{R}^{p\times p},\ |\mathrm{supp}(\beta^*)| \leq s,\\ &M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M, \Delta \geq c_L > 0\big\},\end{aligned} \tag{12}$$

where M and c_L are absolute constants that do not scale with n, p and s. They then establish the following minimax lower bound,

$$\inf_{\widehat{\beta}} \sup_{(\beta^*,\Sigma)\in\Theta_s} \mathbb{E}\left[\|\widehat{\beta} - \beta^*\|_2\right] \geq C_M \cdot \Delta\sqrt{\frac{s\log p}{n}},$$

where the infimum is taken over any estimator $\widehat{\beta}$ based on the samples, and C_M is some constant depending on M. Under such a setting, both AdaLDA and PANDA are minimax optimal in terms of β^* estimation. When M is allowed to scale with n, p and s, the PANDA method still attains the same rates of convergence for parameter estimation as the AdaLDA method. Specifically, we follow the same analysis in Cai and Zhang (2019) and rewrite their results with explicit dependence on M as follows,

$$\|\widehat{\beta} - \beta^*\|_2 = \mathcal{O}_P\left(\sigma_{\max} M\Delta\sqrt{\frac{s\log p}{n}}\right),$$

$$\frac{|\widehat{\Delta}^2 - \Delta^2|}{\Delta^2} = \mathcal{O}_P\left(\sigma_{\max} M^{3/2}\sqrt{\frac{s\log p}{n}}\right).$$

In addition, to ensure the above rates of convergence with high probability, the sample size n needs to satisfy that

$$n = \mathcal{O}_P\left(\sigma_{\max}^2 M^3 \Delta^2 s \log p\right).$$

As can be seen, in Theorem 2, our convergence rates (11a) and (11b) matches the convergence rates in Cai and Zhang (2019) with the same order of sample complexity.

Next, let us establish an upper bound for the misclassification rate of the obtained estimator $\widehat{\beta}$ in the PANDA method.

Theorem 3 (Misclassification Rate) *Under the identical conditions as in Theorem 1 or 2, we have, with probability goes to 1,*

$$\mathcal{R}(\widehat{\beta}) - \mathcal{R}^* \leq C \cdot \exp\left(-\frac{\Delta^2}{8}\right) \sigma_{\max}^{-q} M^{3-q} \Delta R \left(\frac{\log p}{n}\right)^{1-q/2}$$

where C is an absolute positive constants.

When $q = 0$ and $R = s$, Cai and Zhang (2019) consider the parameter space of β^* and Σ defined in (12), where M is a constant, and establish the following minimax lower bound

$$\inf_{\widehat{f}} \sup_{(\beta^*, \Sigma) \in \Theta_s} \mathcal{R}(\widehat{f}) - \mathcal{R}^* \geq C \cdot \exp\left(-\frac{\Delta^2}{8}\right) \Delta^{-1} \frac{s \log p}{n},$$

where the infimum is taken over any linear discriminant rule $\widehat{f}$ based on the samples. Under such a setting, both AdaLDA and PANDA attain the minimax optimal rates of convergence for the misclassification rate that

$$\mathcal{R}(\widehat{\beta}) - \mathcal{R}^* = \mathcal{O}_P\left(\exp\left(-\frac{\Delta^2}{8}\right) M^3 \Delta \frac{s \log p}{n}\right).$$

Remark 4 *The probability of the convergence rates in Theorems 1, 2 and 3 is due to the uncertainty of data, which is addressed in Lemma 8 and Lemma 10 in later analysis. As a summary, the probability of our convergence rates to hold is at least $1 - 4p^{-1} - 2p\exp(-\frac{n-1}{16}) - c_1 \exp(-c_2 n)$. With our sample size condition in* (8)*, the above probability has an order of $1 - \mathcal{O}(p^{-1})$.*

Remark 5 *Note that while the choice of the tuning parameters c and λ in* (7) *guarantees the optimal rates of convergence in both the estimation error and misclassification rate, in practice we recommend to fine-tune these parameters to achieve more appealing performance. In our numerical studies below, we use an independent validation set to tune the parameters in our PANDA method as well as the LPD and AdaLDA method for comparison. We also include the results of our PANDA method with the fine-tuned parameters and with parameters set as in* (7) *for comparison.*

5. Numerical Results

In this section, we thoroughly compare our proposed PANDA method with the LPD method and AdaLDA method through numerical experiments using both simulated and real data.

5.1 Simulation

To make a fair comparison of the three methods' performances, we fine-tune the parameters for each method on a validation dataset independent from the training data, and we provide both the estimation error of β^* (in ℓ_2 norm) and the population risk (2.1) of each method.
Settings: We follow the settings in Cai and Zhang (2019) to generate Σ and β^*.

(a) **AR(1)**. We let $\Omega_{j,k} = 0.9^{|j-k|}$, $\Sigma = \Omega^{-1}$ and $\beta^* = (2/\sqrt{s}, \cdots, 2/\sqrt{s}, 0, \cdots, 0)^\top$, where the first s entries are non-zero and $\|\beta^*\|_2 = 2$.

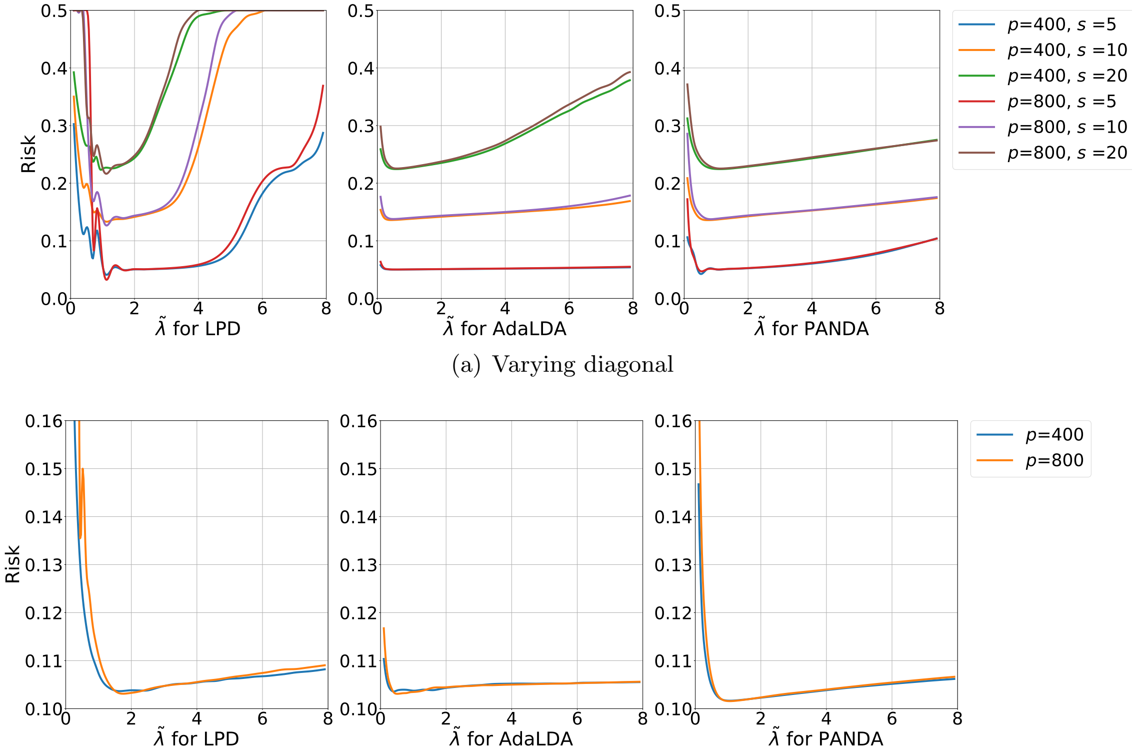
(b) **Varying diagonal**. We let the diagonal entries of Σ as $\Sigma_{j,j} = 11$ for $j = 1, 2, \cdots, 5$, and $\Sigma_{j,j} = 1 + U_j$ for $j = 6, 7, \cdots, p$, where U_i's are independently drawn from the uniform distribution $U(0, 1)$, and we let the off-diagonal entries be $\Sigma_{j,k} = 0.9^{|j-k|}$. We let $\beta^* = (1/\sqrt{s}, \cdots, 1/\sqrt{s}, 0, \cdots, 0)^\top$, where only the first s entries are non-zero and $\|\beta^*\|_2 = 1$.

(c) **Erdös-Rényi random graph**. We let $\widetilde{\Omega}_{j,k} = u_{j,k} v_{j,k}$, where $v_{j,k}$'s are i.i.d. Bernoulli random variables with success rate 0.2, and $u_{j,k}$'s are i.i.d. uniform random variables over $[0.5, 1] \bigcup [-1, -0.5]$, and $v_{j,k}$'s and $u_{j,k}$'s are independent. Then we let $\widetilde{\Omega}_s = (\widetilde{\Omega} + \widetilde{\Omega}^\top)/2$ and $\Omega_0 = \widetilde{\Omega}_s + \left[\max(-\lambda_{\min}(\widetilde{\Omega}_s), 0) + 0.05\right] I_p$. Let D_0 be a diagonal matrix with diagonal elements same as Ω_0's. We let $\Omega = D_0^{-1/2} \Omega_0 D_0^{-1/2}$ and $\Sigma = \Omega^{-1}$, and let $\beta^* = (1/\sqrt{s}, \cdots, 1/\sqrt{s}, 0, \cdots, 0)^\top$ where only the first s entries are non-zero and $\|\beta^*\|_2 = 1$.

(d) **Block sparse model**. We first construct a matrix B of size $p \times p$ as follows. For $1 \leq j \leq p/2$ and $j < k \leq p$, we let $B_{j,k} = B_{k,j} = 10 b_{i,j}$, where $b_{j,k}$'s are i.i.d. Bernoulli variables with success rate 0.5. For $p/2 < j < k \leq p$, we let $B_{j,k} = B_{k,j} = 10$. For the diagonal elements, we let $B_{j,j} = 1$ for $1 \leq j \leq p$. Then we let $w = \max(-\lambda_{\min}(B), 0) + 0.05$ and let $\Omega = (B + wI_p)/(1 + w)$ and $\Sigma = \Omega^{-1}$. We let $\beta^* = (\frac{1}{2\sqrt{s}}, \cdots, \frac{1}{2\sqrt{s}}, 0, \cdots, 0)^\top$, where only the first s entries are non-zero and $\|\beta^*\|_2 = 1/2$.

(e) **Approximately sparse setting**. We let $\Sigma_{j,k} = 0.9^{|j-k|}$ and $\beta^*_j = 0.75^j$, which are approximately sparse. Note that $\|\beta^*\|_2 \approx 3$ when p is large.

Parameter Tuning: While both the AdaLDA method and PANDA method achieve guaranteed theoretical properties with specific tuning parameters, we observe in our experiments that tuning these parameters via a validation set yields better empirical results. In our experiments, under each setting, we randomly sample a validation dataset with $n = 200$ data points from each class. Motivated by the choice of λ in (7), we let $\lambda = \widetilde{\lambda} \cdot \sqrt{\log p/n}$, and we tune the parameter $\widetilde{\lambda}$, as equivalent to tuning λ. For a fair comparison, for all the three methods (LPD, AdaLDA, and PANDA) we tune $\widetilde{\lambda}$ by a grid search over a range from 0.1 to 8.0, with a grid size 0.1. Figures 1 and 2 show the results of the misclassification risks and the estimation errors $\|\widehat{\beta} - \beta^*\|_2$ versus the $\widetilde{\lambda}$ value in the three methods, averaged over 100 replicates under each setting of different p and s. For the parameter c in the PANDA method, we observe that the results are insensitive to the value of c as long as c is not too small, see Table 1 for the result of the misclassification rate with different choices of c under the AR(1) model as an example. Therefore, we set $c = 20$ for all settings.



(a) Varying diagonal

(b) Approximately Sparse

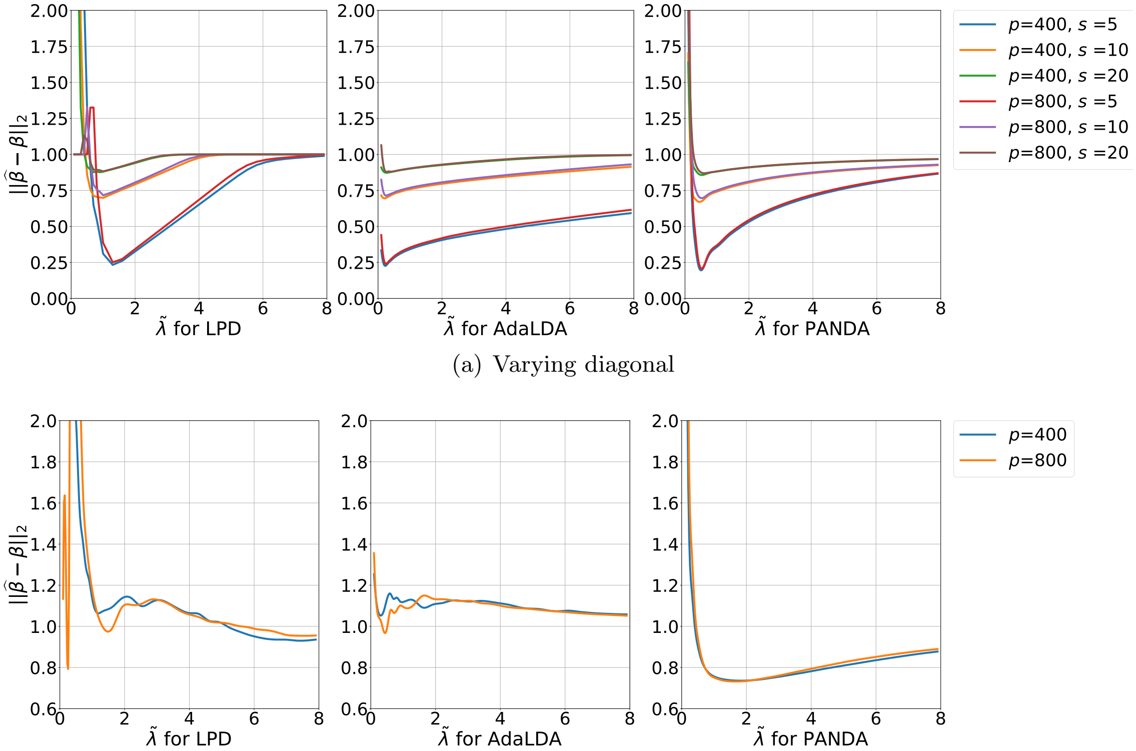
Figure 1: *The misclassification rate v.s. the values of the parameter $\widetilde{\lambda}$ in LPD (left), AdaLDA (middle) and PANDA (right). Results are averaged over 100 replicates.*

Table 1: Misclassification rate of the PANDA method under the AR(1) model with $n = 200$, $p = 400$, $s = 5$ and different c, averaged over 100 replicates. The standard deviations are provided in brackets.

c	1e-3	1e-2	0.1
Misclassification rate	0.3729 (0.1489)	0.2155 (0.0035)	0.2106 (0.0050)
c	1	10	100
Misclassification rate	0.2044 (0.0049)	0.2036 (0.0053)	0.2035 (0.0054)

Tuning Sensitivity: We thoroughly investigate the sensitivity of the tuning parameters under different settings. Since the choice of λ in the LPD method relies on the unknown population quantity Δ, so does the optimal value of λ (or $\widetilde{\lambda}$, equivalently) in practice. We consider following settings to see how the population distribution, especially the scale of Δ, changes the empirically optimal tuning parameters of the LPD, AdaLDA and PANDA methods. For the varying diagonal model, we set $p = 400, 800$, $s = 5$, and $\beta^* = \eta \cdot (1/\sqrt{s}, \cdots, 1/\sqrt{s}, 0, \cdots, 0)^\top$ for $\eta = 1, 2, 4$, where the first s entries are non-zero. For the approximately sparse β model, we set $p = 400$, 800, and $\beta^*_j = \eta \cdot 0.75^j$ for $\eta = 1, 2, 4$.

During the tuning process, we observe that the empirically optimal tuning parameter $\widetilde{\lambda}$ for the PANDA method is less sensitive to the change of unknown population quantities among different settings, in comparison with the LPD method and AdaLDA method. In



(a) Varying diagonal

(b) Approximately Sparse

Figure 2: *ℓ_2 estimation error v.s. values of tuning parameter $\widetilde{\lambda}$ in LPD (left), AdaLDA (middle) and PANDA (right). Results are averaged over 100 replicates.*

particular, Figure 3 shows the distribution of the empirically optimal tuning parameter over 100 replicates under each setting as specified above. The results show that for the PANDA method, the optimal tuning parameter is always close to 1, and does not change much across the different settings.

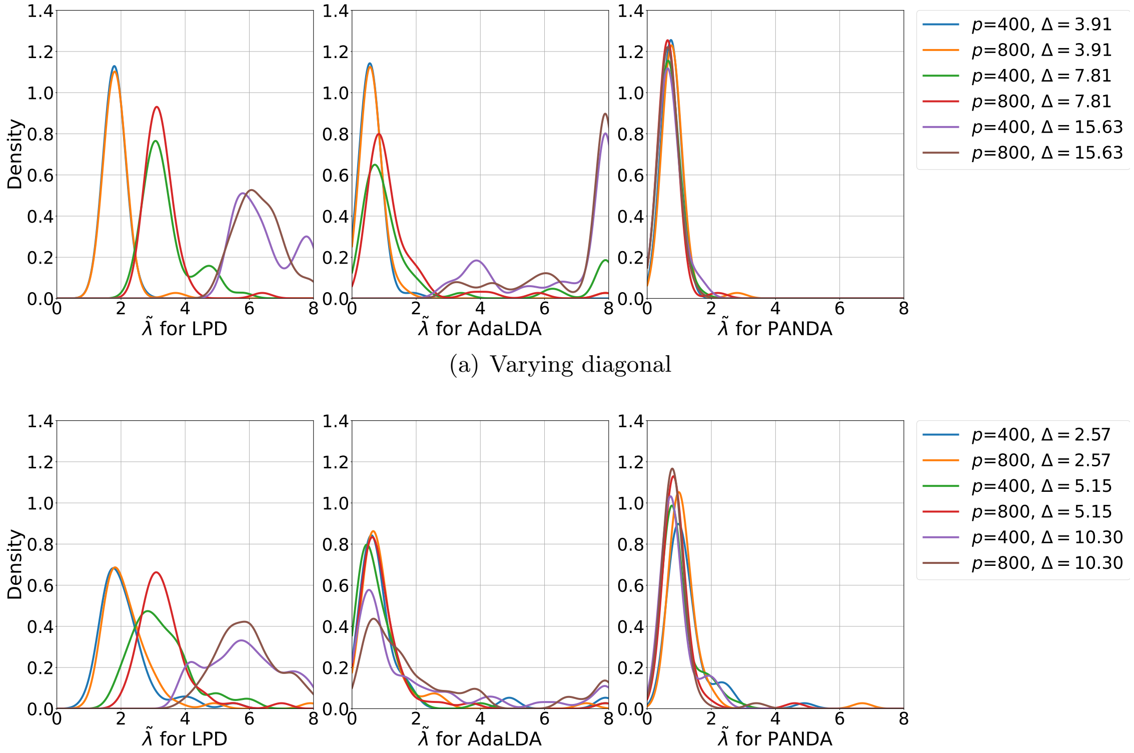
Parameter Estimation: Table 2 summarizes the estimation error of β^*, $\|\widehat{\beta}-\beta^*\|_2$, averaged over 100 random replicates under each setting. It is seen that our proposed PANDA method achieves equal or better performance compared with the LPD and AdaLDA methods in most settings.

Risk Evaluation: Table 3 summarizes the misclassification rate under each setting averaged over 100 random replicates. It is seen that our proposed PANDA method achieves similar or better performances than the LPD method and AdaLDA method in most settings.

Running Time: Table 4 summarizes the running time of our PANDA method and the AdaLDA method under the Varying Diagonal model on a regular computer (Intel Core i5, 2.3GHz). For both methods we use Gurobi, a commercial software that provides state-of-the-art solver for linear programming and second order cone programming, to solve the optimization problems. As can be seen, our PANDA method requires less running time than the AdaLDA method.

Table 2: *The ℓ_2 estimation errors under each setting, averaged over 100 replicates. The standard deviations are given in brackets. The lower value at the significance level* 0.05 *between the AdaLDA and the PANDA method are marked in bold.*

Model	**Specification**						
AR(1) $\lVert\beta^*\rVert = 2$	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
	LPD	1.8875	1.9607	1.9846	1.8960	1.9669	1.9868
		(0.0494)	(0.0313)	(0.0101)	(0.0416)	(0.0199)	(0.0094)
	AdaLDA	1.8854	1.9545	1.9821	1.8952	1.9593	1.9850
		(0.0495)	(0.0200)	(0.0098)	(0.0412)	(0.0184)	(0.0084)
	PANDA	**1.8673**	**1.9521**	1.9814	**1.8856**	1.9571	**1.9830**
		(0.0542)	(0.0229)	(0.0112)	(0.0460)	(0.0190)	(0.0104)
Varying Diagonal $\lVert\beta^*\rVert = 1$	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
	LPD	0.3135	0.7273	0.8841	0.3158	0.7346	0.8949
		(0.1088)	(0.0488)	(0.0178)	(0.1128)	(0.0393)	(0.0190)
	AdaLDA	**0.2753**	0.7198	0.8837	**0.2942**	0.7371	0.8935
		(0.0712)	(0.0387)	(0.0172)	(0.0764)	(0.0374)	(0.0146)
	PANDA	0.3113	0.7177	**0.8797**	0.3197	**0.7305**	**0.8901**
		(0.1110)	(0.0478)	(0.0171)	(0.1166)	(0.0381)	(0.0176)
Erdös-Rényi Random Graph $\lVert\beta^*\rVert = 1$	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
	LPD	0.5715	0.7071	1.0416	0.5933	0.7677	0.9344
		(0.1108)	(0.0965)	(0.1608)	(0.1168)	(0.0855)	(0.0867)
	AdaLDA	0.5688	0.6895	1.0055	0.5949	0.7642	0.9308
		(0.1136)	(0.0761)	(0.0637)	(0.0980)	(0.0914)	(0.1126)
	PANDA	**0.5366**	0.7078	**0.9477**	**0.5753**	**0.7326**	0.9114
		(0.1162)	(0.2120)	(0.0895)	(0.0966)	(0.1054)	(0.2358)
Block Sparse $\lVert\beta^*\rVert = 1$	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
	LPD	0.5066	0.5636	0.6571	0.4653	0.5490	0.5475
		(0.1184)	(0.1400)	(0.2355)	(0.0908)	(0.0653)	(0.0987)
	AdaLDA	0.5145	0.5480	0.5790	0.4798	0.5391	**0.5036**
		(0.0321)	(0.0082)	(0.0110)	(0.0235)	(0.0143)	(0.0044)
	PANDA	**0.4332**	**0.4986**	**0.5409**	0.4789	**0.5229**	0.5425
		(0.0511)	(0.0272)	(0.0278)	(0.1241)	(0.0665)	(0.1058)
Approximately Sparse $\lVert\beta^*\rVert \approx 3$	p	400	800	1200			
	LPD	1.0152	0.9900	0.9750			
		(0.2968)	(0.2897)	(0.3112)			
	AdaLDA	1.0117	1.0273	1.0013			
		(0.2877)	(0.2998)	(0.3192)			
	PANDA	**0.8205**	**0.8547**	**0.8514**			
		(0.2328)	(0.2701)	(0.2380)			



(a) Varying diagonal

(b) Approximately Sparse

Figure 3: *The distribution of the empirically optimal tuning parameter $\widetilde{\lambda}$ for LPD (left), AdaLDA (middle) and PANDA (right) over 100 replicates, approximated with kernel smoothing. The optimal choice of the parameter $\widetilde{\lambda}$ in our PANDA method relies less on the population.*

Variable Selection: We expect our PANDA method is capable for variable selection, as similar to the LPD and AdaLDA method. Here we report the performance of the three methods in the accuracy of finding the sparse signal, under the AR(1) and Varying Diagonal model as described above. To be more specific, we compute the average of True Positive and True Negative, together with the Precision and Recall for identifying the non-zero entries in β^*, after applying a threshold at 0.01 for entries in $\widehat{\beta}$. The results under the two models are summarized in Tables 5 and 6, respectively. We see that PANDA achieves comparable performance with LPD and AdaLDA in the sense of accuracy of variable selection.

Table 3: *The misclassification rate under each setting averaged over 100 replicates. The standard deviations are given in brackets. The lower value at the significance level* 0.05 *between the AdaLDA and the PANDA method are marked in bold.*

Model	**Specification**						
AR(1)	(s,p)	$(5,400)$	$(10,400)$	$(20,400)$	$(5,800)$	$(10,800)$	$(20,800)$
	LPD	0.2086 (0.0074)	0.2900 (0.0109)	0.3535 (0.0099)	0.2112 (0.0074)	0.2908 (0.0066)	0.3532 (0.0080)
	AdaLDA	0.2082 (0.0068)	0.2890 (0.0080)	0.3522 (0.0075)	0.2120 (0.0088)	0.2913 (0.0072)	**0.3525** (0.0082)
	PANDA	**0.2068** (0.0069)	0.2886 (0.0087)	0.3542 (0.0104)	0.2114 (0.0084)	0.2910 (0.0079)	0.3571 (0.01206)
Varying Diagonal	(s,p)	$(5,400)$	$(10,400)$	$(20,400)$	$(5,800)$	$(10,800)$	$(20,800)$
	LPD	0.0515 (0.0028)	0.1382 (0.0054)	0.2269 (0.0065)	0.0520 (0.0038)	0.1390 (0.0056)	0.2289 (0.0087)
	AdaLDA	0.0508 (0.0018)	0.1376 (0.0046)	0.2266 (0.0063)	0.0513 (0.0032)	0.1386 (0.0054)	0.2284 (0.0081)
	PANDA	0.0512 (0.0026)	0.1374 (0.0040)	0.2266 (0.0064)	0.0514 (0.0025)	0.1384 (0.0048)	0.2292 (0.0088)
Erdös-Rényi Random Graph	(s,p)	$(5,400)$	$(10,400)$	$(20,400)$	$(5,800)$	$(10,800)$	$(20,800)$
	LPD	0.2857 (0.0138)	0.2424 (0.0099)	0.1150 (0.0054)	0.2757 (0.0148)	0.3256 (0.0182)	0.3289 (0.0145)
	AdaLDA	0.2849 (0.0129)	0.2414 (0.090)	0.1162 (0.0058)	0.2758 (0.0138)	0.3246 (0.0185)	0.3281 (0.0152)
	PANDA	**0.2823** (0.0117)	0.2403 (0.0106)	**0.1114** (0.0044)	**0.2721** (0.0129)	**0.3183** (0.0166)	**0.3209** (0.0161)
Block Sparse	(s,p)	$(5,400)$	$(10,400)$	$(20,400)$	$(5,800)$	$(10,800)$	$(20,800)$
	LPD	0.1643 (0.0056)	0.0954 (0.0038)	0.0451 (0.0028)	0.4184 (0.0170)	0.1724 (0.0029)	0.3776 (0.0077)
	AdaLDA	0.1745 (0.0061)	0.1002 (0.0009)	0.0451 (0.0003)	0.4378 (0.0156)	0.1739 (0.0007)	0.3811 (0.0020)
	PANDA	**0.1614** (0.0047)	**0.0938** (0.0018)	**0.0437** (0.0007)	**0.4168** (0.0159)	**0.1706** (0.0026)	**0.3753** (0.0072)
Approximately Sparse	p	400	800	1200			
	LPD	0.1054 (0.0046)	0.1047 (0.0030)	0.1053 (0.0040)			
	AdaLDA	0.1042 (0.0029)	0.1043 (0.0035)	0.1042 (0.0038)			
	PANDA	0.1034 (0.0033)	0.1039 (0.0038)	0.1040 (0.0045)			

Table 4: *Running time (in seconds) of the PANDA and AdaLDA methods under the Varying Diagonal model using Gurobi, over 100 replicates. The standard deviations are given in brackets.*

(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$
AdaLDA	106.739 (2.653)	107.743 (2.588)	107.017 (2.782)
PANDA	70.202 (4.751)	71.312 (4.389)	72.112 (4.965)
(s, p)	$(5, 800)$	$(10, 800)$	$(20, 800)$
AdaLDA	413.262 (13.209)	413.876 (12.708)	416.793 (12.383)
PANDA	325.486 (16.372)	326.125 (16.504)	333.427 (13.554)

Table 5: *The results on variable selection over 100 replicates under the AR(1) model. The standard deviations are given in brackets.*

Criteria	**Specification**						
	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
True Positive	LPD	1.95 (0.59)	1.51 (0.76)	1.27 (0.65)	1.76 (0.62)	1.17 (0.49)	1.02 (0.45)
	AdaLDA	1.97 (0.56)	1.60 (0.70)	1.34 (0.65)	1.77 (0.55)	1.38 (0.56)	1.03 (0.33)
	PANDA	2.20 (0.68)	1.75 (0.84)	1.58 (0.96)	1.96 (0.65)	1.51 (0.69)	1.29 (0.57)
	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
True Negative	LPD	386.87 (10.20)	385.1 (11.46)	375.95 (7.91)	787.72 (8.01)	786.57 (6.60)	778.23 (4.04)
	AdaLDA	387.55 (6.92)	385.31 (7.48)	376.67 (4.80)	786.91 (9.79)	784.42 (7.72)	777.92 (4.04)
	PANDA	386.69 (8.39)	384.04 (10.17)	375.45 (7.10)	785.80 (11.24)	783.44 (11.04)	775.16 (7.32)
	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
Precision	LPD	0.3741 (0.2992)	0.5668 (0.3666)	0.5955 (0.3778)	0.4039 (0.3207)	0.6202 (0.3726)	0.7003 (0.3520)
	AdaLDA	0.3624 (0.2834)	0.5126 (0.3432)	0.5465 (0.3514)	0.3880 (0.3166)	0.4406 (0.3388)	0.6379 (0.3537)
	PANDA	0.3713 (0.2858)	.4824 (0.3373)	0.4797 (0.3164)	0.4055 (0.3266)	0.4511 (0.3427)	0.4411 (0.3132)
	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
Recall	LPD	0.3900 (0.1185)	0.1510 (0.07588)	0.0635 (0.0324)	0.3520 (0.1243)	0.1170 (0.0493)	0.0510 (0.0224)
	AdaLDA	0.3940 (0.1118)	0.1600 (0.0696)	0.0670 (0.0327)	0.3540 (0.1096)	0.1380 (0.0565)	0.0515 (0.0166)
	PANDA	0.4400 (0.1363)	0.1750 (0.0845)	0.0790 (0.0478)	0.3920 (0.1300)	0.1510 (0.0689)	0.0645 (0.0287)

5.2 Leukemia data

We investigate the performance of the PANDA, LPD, and AdaLDA methods on a Leukemia dataset from high-density oligonucleotide microarrays. This dataset was first analyzed by Golub et al. (1999), and it contains 72 samples of two categories: 47 of acute lymphoblas-

Table 6: *The results on variable selection over 100 replicates under the Varying Diagonal model. The standard deviations are given in brackets.*

Criteria	**Specification**						
True Positive	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
	LPD	5.00	5.14	4.33	5.00	5.10	3.81
		(0)	(0.85)	(1.14)	(0)	(0.77)	(1.35)
	AdaLDA	5.00	5.25	4.34	5.00	5.08	3.89
		(0)	(0.54)	(1.12)	(0)	(0.60)	(1.27)
	PANDA	5.00	5.30	4.44	5.00	5.20	4.02
		(0)	(0.69)	(1.00)	(0)	(0.64)	(1.31)
True Negative	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
	LPD	394.94	389.97	380.00	794.90	790.0	780.0
		(0.31)	(0.30)	(0)	(0.48)	(0)	(0)
	AdaLDA	394.95	389.99	380.0	794.94	789.99	780.0
		(0.26)	(0.10)	(0)	(0.31)	(0.10)	(0)
	PANDA	394.86	390.0	379.98	794.95	790.0	779.99
		(0.75)	(0)	(0.20)	(0.26)	(0)	(0.10)
Precision	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
	LPD	0.9910	0.9970	1	0.9863	0.9939	1
		(0.0461)	(0.0302)	(0)	(0.0635)	(0.0010)	(0)
	AdaLDA	0.9921	0.9986	0.9608	0.9910	0.9983	1
		(0.0400)	(0.0143)	(0.0028)	(0.0461)	(0.0167)	(0)
	PANDA	0.9830	1	0.9975	0.9921	1	0.9985
		(0.0733)	(0)	(0.0251)	(0.0400)	(0)	(0.0145)
Recall	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
	LPD	1	0.514	0.2165	1	0.51	0.1905
		(0)	(0.0853)	(0.0569)	(0)	(0.0772)	(0.0673)
	AdaLDA	1	0.525	0.217	1	0.508	0.1945
		(0)	(0.0539)	(0.0560)	(0)	(0.0598)	(0.0635)
	PANDA	1	0.53	0.222	1	0.52	0.2010
		(0)	(0.0689)	(0.0499)	(0)	(0.0636)	(0.0655)

tic leukemia (ALL), and 25 of acute myeloid leukemia (AML). Each sample contains the quantitative expression levels of 7129 genes.

Preprocessing: We follow the preprocessing steps in Cai and Zhang (2019). First, we combine the data from both categories and compute the sample variance of each gene. Then, we drop the genes with sample variance beyond the lower and upper 6-quantiles of the total 7129 genes.

Result: To provide a fair comparison among the LPD, AdaLDA, and PANDA methods, we tune the parameters using a validation set. After preprocessing the raw data, we randomly split the data into training, validation, and testing sets. Specifically, the training set contains 29 ALL and 15 AML samples, the validation set contains 9 ALL and 5 AML samples, and the testing set contains 9 ALL and 5 AML samples. For the computational efficiency, we only use 2000 genes with the largest absolute values of the two-sample t-test in the training set, as suggested by Cai and Zhang (2019). We repeat the process 100 times, and provide the three methods' average misclassification rates on the testing set (testing error) and their standard deviations in Table 7. As can be seen, the PANDA method achieves a lower misclassification rate than both the LPD and AdaLDA methods.

Table 7: *The performance of PANDA, AdaLDA and LPD on the Leukemia dataset. The testing errors are averaged over 100 replicates. The standard deviation of the testing errors are given in brackets. The difference between PANDA and the other two methods is significant by pair-wise t-test with a p-value less than 0.001.*

	LPD	AdaLDA	PANDA
Testing Error	9.28% (6.87%)	10.64% (7.92%)	6.93% (6.74%)

6. Extension to multiple-class LDA

In this section, we discuss the extension of PANDA method to K-class LDA in high dimensions. To be more specific, we consider the following data setting. Suppose we have samples $\left\{X_i^{(k)} : k = 1, 2, \cdots, K,\ i = 1, 2, \cdots, n_k\right\}$ from K classes denoted by $k = 1, 2, \cdots, K$, such that $X_i^{(k)}$'s are i.i.d. from $N(\mu^{(k)}, \Sigma)$. Also, we suppose that the prior probabilities $\pi_1, \pi_2, \cdots, \pi_K$ for the K classes are known. Then the oracle classification rule for future data Z is given by $f(Z) = \text{argmax}_k D_k$, where $D_1 = 0$, $D_k = \left(Z - \frac{\mu^{(1)}+\mu^{(k)}}{2}\right)^\top \beta^{(k)} + \log\left(\frac{\pi_k}{\pi_1}\right)$, with $\beta^{(k)} = \Sigma^{-1}(\mu^{(k)} - \mu^{(1)})$. In addition, we define $\Delta_k = \sqrt{\beta^{(k)\top}\Sigma\beta^{(k)}}$. Let $\widehat{\mu}^{(k)}$ be the sample mean of data in class k, and let $\widehat{\Sigma}$ be the pooled sample covariance matrix over the K classes. Then, one can construct the classifier by using the K-class PANDA method, which simultaneously estimate $\beta^{(k)}$'s and Δ_k's via the following optimization problems.

$$(\widehat{\beta}^{(k)}, \widehat{\tau}^k) \in \mathop{\arg\min}_{\beta,\tau} \quad \|\beta\|_1 + c_k\tau^2, \tag{13}$$
$$\text{subject to} \quad \|\widehat{\Sigma}\beta - (\widehat{\mu}^{(k)} - \widehat{\mu}^{(1)})\|_\infty \le \lambda\widehat{\sigma}_{\max}(\tau+1), \quad \sqrt{\beta^\top\widehat{\Sigma}\beta} \le \tau.$$

Based on $\widehat{\beta}^{(k)}$'s, one can construct the classifier by $\widehat{f}(Z) = \arg\max_k \widehat{D}_k$ with $\widehat{D}_1 = 0$ and $\widehat{D}_k = (Z - \frac{\widehat{\mu}^{(1)}+\widehat{\mu}^{(k)}}{2})^\top\widehat{\beta}^{(k)}$.

Following the similar technical argument as for Theorems 1, 2 and 3, we can establish the following theoretical properties for K-class PANDA method.

Theorem 6 *Suppose that Assumption 2 hold, and $\beta^{(k)*} \in \mathbb{B}_q(R)$ for some $q \in [0, 1)$ and some $R > 0$ for all $k = 2, 3, \cdots, K$. Let $\left(\widehat{\beta}^{(k)}, \widehat{\tau}_k\right)$ be an optimal solution of* (13)*. Given*

$$c_k = \frac{1}{8\left(\|\widehat{\mu}^{(k)} - \widehat{\mu}^{(1)}\|_\infty + 4\widehat{\sigma}_{\max}\sqrt{\frac{\log p}{n}}\right)}, \qquad \lambda = 20\sqrt{\frac{\log p}{n}},$$

for sufficiently large n such that

$$n \ge C \cdot a^{-2}\Delta_k^2\sigma_{\max}^2 M^{2+\frac{1}{1-q}} R^{\frac{2}{1-q}} \log p$$

where C is an absolute constant, we have, with probability goes to 1,

$$\begin{aligned}
\|\widehat{\beta}^{(k)} - \beta^{(k)*}\|_1 &\le C_1 \cdot (\Delta_k + 1)\,(\sigma_{\max} M)^{1-q}\, R \left(\frac{\log p}{n}\right)^{(1-q)/2}, \\
\|\widehat{\beta}^{(k)} - \beta^{(k)*}\|_2 &\le C_2 \cdot (\Delta_k + 1)\,(\sigma_{\max} M)^{1-q/2}\, \sqrt{R} \left(\frac{\log p}{n}\right)^{1/2-q/4}, \\
\frac{|\widehat{\tau}_k^2 - \Delta_k^2|}{\Delta_k^2} &\le C_3 \cdot (1 + \Delta_k^{-1}) \sigma_{\max}^{1-q/2} M^{3/2-q} R \Big(\frac{\log p}{n}\Big)^{(1-q)/2},
\end{aligned}$$

where C_1, C_2 and C_3 are positive constants.

Theorem 7 *Let $\Delta_{\min} = \min\{(\mu^{(j)} - \mu^{(i)})^\top \Sigma^{-1} (\mu^{(j)} - \mu^{(i)}) : \ 1 \le i < j \le K\}$. Under the identical conditions as in Theorem 6, we have, with probability goes to 1,*

$$\mathcal{R}(\widehat{f}) - \mathcal{R}^* \le C \cdot \exp\left(-\frac{\Delta_{\min}^2}{8}\right) \sigma_{\max}^{-q} M^{3-q} \Delta_{\min} R \left(\frac{\log p}{n}\right)^{1-q/2},$$

where C is an absolute positive constants.

7. Proofs of the Main Results

In this section, we provide the proof for Theorem 1 in Section 7.1 and Theorem 3 in Section 7.3. The proofs of lemmas can be found in the supplementary material.

7.1 Proof of Theorem 1

Proof We denote by $\delta = \widehat{\beta} - \beta^*$ and $\tau^* = \sqrt{\beta^{*\top} \widehat{\Sigma} \beta^*}$. We first derive the upper bound for $\|\delta\|_1$. Based on this upper bound, we then derive the upper bounds for $\|\delta\|_2$ and $\widehat{\tau}$.

For ease of presentation, we first define the following events,

$$\begin{aligned}
\mathcal{E}_\tau &= \left\{ |\beta^{*\top} (\widehat{\Sigma} - \Sigma) \beta^*| \le \frac{1}{2} \beta^{*\top} \Sigma \beta^* \right\} = \left\{ \frac{1}{2} \Delta^2 \le \tau^{*2} \le \frac{3}{2} \Delta^2 \right\}, \\
\mathcal{E}_{\sigma_{\max}} &= \left\{ |\widehat{\sigma}_{\max}^2 - \sigma_{\max}^2| \le \frac{1}{2} \sigma_{\max}^2 \right\}, \\
\mathcal{E}_{\mu_d} &= \left\{ \|\mu_d\|_\infty - 2\sqrt{2}\sigma_{\max} \sqrt{\frac{\log p}{n}} \le \|\widehat{\mu}_d\|_\infty \le \|\mu_d\|_\infty + 2\sqrt{2}\sigma_{\max} \sqrt{\frac{\log p}{n}} \right\}, \\
\mathcal{E}_1 &= \left\{ \|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty \le 10 \sigma_{\max} \Delta \sqrt{\frac{\log p}{n}} \right\}, \\
\mathcal{E}_2 &= \left\{ \|\widehat{\Sigma}\beta^* - \widehat{\mu}\|_\infty \le 20 \widehat{\sigma}_{\max} \sqrt{\frac{\log p}{n}} (\tau^* + 1) \right\}.
\end{aligned}$$

Before we proceed, we introduce the following lemma.

Lemma 8 *For any $\beta^* \in \mathbb{R}^p$, we have*

$$\mathbb{P}\left(\mathcal{E}_\tau\right) \geq 1 - 2\exp\left(-\frac{n-1}{16}\right), \quad \mathbb{P}\left(\mathcal{E}_{\sigma_{\max}}\right) \geq 1 - 2p\exp\left(-\frac{n-1}{16}\right),$$
$$\mathbb{P}\left(\mathcal{E}_{\mu_d}\right) \geq 1 - 2p^{-1}, \quad \mathbb{P}(\mathcal{E}_1) \geq 1 - 2p^{-1}.$$

Moreover, we have

$$\mathcal{E}_2 \supseteq \left(\mathcal{E}_\tau \bigcap \mathcal{E}_{\sigma_{\max}} \bigcap \mathcal{E}_1\right).$$

Upper bound for $\|\delta\|_1$. We first provide an upper bound for $\delta^\top \widehat{\Sigma}\delta$ in terms of $\|\delta\|_1$, which is essential for deriving an upper bound of $\|\delta\|_1$.

Lemma 9 *Suppose that the events $\mathcal{E}_\tau$, $\mathcal{E}_{\sigma_{\max}}$, $\mathcal{E}_1$ and $\mathcal{E}_2$ hold. Then we have*

$$\delta^\top \widehat{\Sigma}\delta \leq 2\lambda\sigma_{\max}\|\delta\|_1 \left(3\Delta + 2 + \sqrt{\frac{\|\delta\|_1}{c}}\right). \tag{15}$$

Our next step is to derive a lower bound for $\delta^\top \widehat{\Sigma}\delta$ in terms of $\|\delta\|_1$, based on the restricted eigenvalue condition of $\widehat{\Sigma}$ on certain restricted subset of $\mathbb{R}^p$. We first introduce the eigenvalue condition of $\widehat{\Sigma}$ that holds with high probability.

Lemma 10 *Suppose that Assumption 2 holds, and $n \geq 2$. There exist absolute positive constants c_1 and c_2 such that*

$$\delta^\top \widehat{\Sigma}\delta \geq \frac{1}{32M}\|\delta\|_2^2 - 81\sigma_{\max}^2 \frac{\log p}{n}\|\delta\|_1^2 \quad \textit{for all } \delta \in \mathbb{R}^p, \tag{16}$$

with probability at least $1 - c_1\exp(-c_2 n)$.

Based on the above result, we derive the restricted eigenvalue condition of $\widehat{\Sigma}$ over a restricted subset. In particular, for $S \subseteq [p]$ and $\beta^* \in \mathbb{R}^p$, we let

$$\mathcal{C}_{S,\beta^*} := \left\{\delta \in \mathbb{R}^p : \ \|\delta_{S^c}\|_1 \leq 3\|\delta_S\|_1 + 4\|\beta^*_{S^c}\|_1\right\}.$$

The next lemma shows that $\delta \in \mathcal{C}_{S,\beta^*}$ for any $S \subseteq [p]$.

Lemma 11 *Suppose that Assumption 1 and events $\mathcal{E}_{\mu_d}$, $\mathcal{E}_{\sigma_{\max}}$, $\mathcal{E}_1$ and $\mathcal{E}_2$, hold. Let $S \subseteq [p]$. Given c and λ in (7), we have $\delta \in \mathcal{C}_{S,\beta^*}$ when n satisfies*

$$n \geq 100a^{-2}\sigma_{\max}^2\Delta^2 \log p.$$

Now, we choose a subset S_η that

$$S_\eta = \left\{j \in [p] : \ |\beta_j^*| \geq \eta\right\},$$
$$\text{where} \quad \eta = \sigma_{\max} M\sqrt{\frac{\log p}{n}}. \tag{17}$$

We further show the upper bounds for $|S_\eta|$ and $\|\beta^*_{S_\eta^c}\|_1$ in the next lemma.

Lemma 12 *When $\beta^* \in \mathbb{B}_q(R)$, we have that*

$$|S_\eta| \le \eta^{-q} R, \tag{18}$$

$$\|\beta^*_{S_\eta^c}\| \le \eta^{1-q} R. \tag{19}$$

Note that if S_η is empty, we immediately have that

$$\|\delta\|_1 \le 4\|\beta^*_{S_\eta^c}\| \le 4\eta^{1-q} R = 4(\sigma_{\max} M)^{1-q} R \left(\frac{\log p}{n}\right)^{\frac{1-q}{2}},$$

which matches the upper bound in (9a).

When S_η is non-empty and $\delta \in \mathcal{C}_{S_\eta,\beta^*}$, we have that

$$\|\delta\|_1 \le 4\|\delta_{S_\eta}\|_1 + 4\|\beta^*_{S_\eta^c}\|_1 \le 4\sqrt{|S_\eta|}\|\delta\|_2 + 4\|\beta^*_{S_\eta^c}\|_1. \tag{20}$$

Plugging the above inequality into (16) yields that

$$\begin{aligned}\delta^\top \widehat{\Sigma} \delta &\ge \frac{1}{512M|S_\eta|}\left(\|\delta\|_1 - 4\|\beta^*_{S_\eta^c}\|\right)^2 - 81\sigma_{\max}^2 \frac{\log p}{n}\|\delta\|_1^2 \\ &\ge \left(\frac{1}{512M|S_\eta|} - 81\sigma_{\max}^2\frac{\log p}{n}\right)\|\delta\|_1^2 - \frac{\|\beta^*_{S_\eta^c}\|_1}{64M|S_\eta|}\|\delta\|_1.\end{aligned}$$

When n satisfies that

$$n \ge C \cdot \sigma_{\max}^2 M |S_\eta| \log p$$

for some constant C, we have that

$$\delta^\top \widehat{\Sigma}\delta \ge \frac{1}{1024M|S_\eta|}\|\delta\|_1^2 - \frac{\|\beta^*_{S_\eta^c}\|_1}{64M|S_\eta|}\|\delta\|_1. \tag{21}$$

Combining (15) with (21), we have that

$$\frac{1}{1024M|S_\eta|}\|\delta\|_1^2 - \frac{\|\beta^*_{S_\eta^c}\|_1}{64M|S_\eta|}\|\delta\|_1 \le 2\lambda\sigma_{\max}\|\delta\|_1\left(3\Delta + 2 + \sqrt{\frac{\|\delta\|_1}{c}}\right).$$

Solving the above inequality with our chosen c, λ and η as in (7) and (17), and using the upper bounds (18) and (19), we have the upper bound for $\|\delta\|_1$ that

$$\|\delta\|_1 \le C \cdot (\sigma_{\max} M)^{1-q}(\Delta + 1) R \left(\frac{\log p}{n}\right)^{\frac{1-q}{2}} \tag{22}$$

for some constant C, given n satisfies that

$$n \ge C \cdot a^{-2}\Delta^2 \sigma_{\max}^2 M^{2+\frac{1}{1-q}} R^{\frac{2}{1-q}} \log p$$

for some constant C.

Upper bound for $\|\delta\|_2$. We prove (9b) based on the previous upper bound for $\|\delta\|_1$. Following Lemma 10, there exist some absolute positive constants c_1 and c_2 such that, with probability at least $1 - c_1 \exp(-c_2 n)$, we have

$$\delta^\top \widehat{\Sigma} \delta \geq \frac{1}{32M}\|\delta\|_2^2 - 81\sigma_{\max}^2 \frac{\log p}{n}\|\delta\|_1^2.$$

The above inequality gives an upper bound of $\|\delta\|_2^2$ in terms of $\delta^\top \widehat{\Sigma}\delta$ and $\|\delta\|_1$, whereas the latter two terms can be further upper bounded using Lemma 9 and (22), respectively.

To bound $\delta^\top \widehat{\Sigma}\delta$, following Lemma 9, we have that

$$\delta^\top \widehat{\Sigma} \delta \leq \lambda \widehat{\sigma}_{\max}\|\delta\|_1 \left(\sqrt{\frac{\|\delta\|_1}{c}} + 3\Delta + 2\right).$$

Note that $\Delta = \sqrt{\mu_d^\top \Sigma^{-1}\mu_d} \geq M^{-1/2}\|\mu_d\|_\infty$, and thus $\|\mu_d\|_\infty/\Delta \leq M^{1/2}$. Hence (7.1) and (7.1) together imply that

$$\begin{aligned}\|\delta\|_2^2 &\leq C \cdot \left[M\delta^\top \widehat{\Sigma}\delta + \sigma_{\max}^2 M \frac{\log p}{n}\|\delta\|_1^2\right] \\ &\leq C \cdot \left[\lambda\widehat{\sigma}_{\max} M(\Delta+1)\|\delta\|_1 + \frac{\lambda\widehat{\sigma}_{\max}}{\sqrt{c}} M\|\delta\|_1^{3/2} + \sigma_{\max}^2 M \frac{\log p}{n}\|\delta\|_1^2\right]\end{aligned}$$

for some constant C. By our choice of c and λ in (7) and the upper bound of $\|\delta\|_1$ in (9a), when n satisfies that

$$n \geq C \cdot \sigma_{\max}^2 M^{2+\frac{1}{1-q}} R^{\frac{2}{1-q}} \log p$$

for some absolute constant C, (7.1) reduces to

$$\|\delta\|_2^2 \leq C \cdot \lambda\widehat{\sigma}_{\max} M(\Delta+1)\|\delta\|_1 \leq C \cdot (\sigma_{\max} M)^{2-q}(\Delta+1)^2 R \left(\frac{\log p}{n}\right)^{1-q/2},$$

which shows (9b) holds.

Upper bound of $|\widehat{\tau}^2 - \Delta^2|/\Delta^2$. Note that $|\widehat{\tau}^2 - \Delta^2| \leq |\widehat{\tau}^2 - \tau^{*2}| + |\tau^{*2} - \Delta^2|$. We upper bound the two terms on the right-hand side respectively in the next lemma.

Lemma 13 *Suppose that Assumption 2, events $\mathcal{E}_\tau, \mathcal{E}_{\sigma_{\max}}, \mathcal{E}_{\mu_d}, \mathcal{E}_1$ and (9a) hold. When n satisfies* (8) *for some absolute constant C, we have that*

$$|\widehat{\tau}^2 - \tau^{*2}| \leq C \cdot \Delta(\Delta+1)\sigma_{\max}^{1-q/2} M^{(3-q)/2} R \left(\frac{\log p}{n}\right)^{(1-q)/2} \tag{23a}$$

$$|\tau^{*2} - \Delta^2| \leq C \cdot \Delta^2 \sigma_{\max}^{1-q/2} M^{(1-q)/2} \sqrt{R}\left(\frac{\log p}{n}\right)^{(2-q)/4} \tag{23b}$$

for some absolute constant C.

Combining (23a) and (23b), we obtain that

$$\frac{|\widehat{\tau}^2 - \Delta^2|}{\Delta^2} \leq C \cdot (1+\Delta^{-1})\sigma_{\max}^{1-q/2} M^{\frac{3-q}{2}} R \left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}$$

for some absolute constant C, and our claim (9c) follows as desired. ■

7.2 Proof of Theorem 2

Proof We first introduce the following lemma that gives a different upper bound of $\delta^\top\widehat{\Sigma}\delta$ as in Lemma 9, with the additional condition that $\widehat{\tau} = \sqrt{\widehat{\beta}^\top\widehat{\Sigma}\widehat{\beta}}$.

Lemma 14 *Suppose that the events $\mathcal{E}_\tau$, $\mathcal{E}_1$ and $\mathcal{E}_2$ hold, and $\widehat{\tau} = \sqrt{\widehat{\beta}^\top\widehat{\Sigma}\widehat{\beta}}$. Then we have*

$$\begin{aligned}\delta^\top\widehat{\Sigma}\delta \leq & C \cdot \lambda\sigma_{\max}\|\delta\|_1\Bigg\{\lambda\sigma_{\max}\|\delta\|_1 + \tau^* + 1 + \left(20\sigma_{\max}\Delta\sqrt{\frac{\log p}{n}}\|\delta\|_1\right)^{1/2} \\ & + \left(2\|\mu_d\|_2\|\delta\|_2\right)^{1/2}\Bigg\},\end{aligned} \tag{24}$$

for some constant C.

Upper bound of $\|\delta\|_2$. Based on Lemma 10 in the previous part, with probability goes to 1 we have that

$$\delta^\top\widehat{\Sigma}\delta \geq \frac{1}{32M}\|\delta\|_2^2 - 81\sigma_{\max}^2\frac{\log p}{n}\|\delta\|_1^2 \quad \text{for all } \delta \in \mathbb{R}^p.$$

When $\delta \in \mathcal{C}_{S_\eta,\beta^*}$, combining the above equation with (24), and using (20), we have that

$$\frac{1}{M}\|\delta\|_2^2 \leq C \cdot \left[\sigma_{\max}^2\frac{\log p}{n}\|\beta^*_{S_\eta^c}\|_1^2 + \lambda\sigma_{\max}\left(\sqrt{|S_\eta|}\|\delta\|_2 + \|\beta^*_{S_\eta^c}\|_1\right)\left(\Delta + 1 + \sqrt{\|\mu_d\|_2\|\delta\|_2}\right)\right]$$

for some constant C, when n satisfies that

$$n \geq C \cdot \sigma_{\max}^2 M^{\frac{2-2q}{2-q}} R^{\frac{2}{2-q}} \log p$$

for some constant C. By setting η as in (17), and using (18) and (19), we finally obtain

$$\|\delta\|_2 \leq C \cdot (\sigma_{\max}M)^{1-q/2}(\Delta+1)\sqrt{R}\left(\frac{\log p}{n}\right)^{1/2-q/4} \tag{25}$$

for some constant C.

Upper bound of $|\widehat{\tau}^2 - \Delta^2|/\Delta^2$. Note that $|\widehat{\tau}^2 - \Delta^2| \leq |\widehat{\tau}^2 - \tau^{*2}| + |\tau^{*2} - \Delta^2|$. In Lemma 13, we have already shown the upper bound for the term $|\tau^{*2} - \Delta^2|$ as (23b), which we also adopt here. With the additional condition that $\sqrt{\widehat{\beta}^\top\widehat{\Sigma}\widehat{\beta}} = \widehat{\tau}$, the upper bound of the term $|\widehat{\tau}^2 - \tau^{*2}|$ can be tighter than (23a), as shown in the following lemma.

Lemma 15 *Suppose that Assumption 2, events $\mathcal{E}_\tau, \mathcal{E}_{\sigma_{\max}}, \mathcal{E}_{\mu_d}, \mathcal{E}_1$ and (9a) hold. Also, suppose that $\sqrt{\widehat{\beta}^\top\widehat{\Sigma}\widehat{\beta}^\top} = \widehat{\tau}$. When n satisfies* (10), *we have that*

$$|\widehat{\tau}^2 - \tau^{*2}| \leq C \cdot \Delta(\Delta+1)\sigma_{\max}^{1-q/2}M^{(3-q)/2}\sqrt{R}\left(\frac{\log p}{n}\right)^{1/2-q/4}. \tag{26}$$

for some absolute constant C.

Combining (23b) and (26), we have that

$$\frac{|\widehat{\tau}^2 - \Delta^2|}{\Delta^2} \leq C \cdot (1 + \Delta^{-1}) \sigma_{\max}^{1-q/2} M^{(3-q)/2} \sqrt{R} \left(\frac{\log p}{n} \right)^{1/2-q/4}.$$

for some constant C. ■

7.3 Proof of Theorem 3

Proof Let $\widehat{\Delta} = \sqrt{\widehat{\beta}^\top \Sigma \widehat{\beta}}$. The misclassification rate of $\widehat{\beta}$ is

$$\mathcal{R}(\widehat{\beta}) = \frac{1}{2}\Phi\left(-\frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) + \frac{1}{2}\Phi\left(\frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right),$$

where $\Phi(\cdot)$ is the CDF of the standard Gaussian distribution. Recall that the optimal risk achieved by Fisher's rule is $\mathcal{R}^* = \Phi(-\frac{\Delta}{2})$. For the first term on the right-hand side of (7.3), its second order Taylor's expansion is

$$\begin{aligned}\Phi\left(-\frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) = \Phi\left(-\frac{\Delta}{2}\right) + \Phi'\left(-\frac{\Delta}{2}\right)\left(\frac{\Delta}{2} - \frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) \\ + \frac{\Phi''(t_1)}{2}\left(\frac{\Delta}{2} - \frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)^2,\end{aligned}$$

where $t_1 \in \left(\frac{-\Delta}{2}, -\frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)$. Similarly, for the second term in (7.3), we have

$$\begin{aligned}\Phi\left(\frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) = \Phi\left(-\frac{\Delta}{2}\right) + \Phi'\left(-\frac{\Delta}{2}\right)\left(\frac{\Delta}{2} + \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) \\ + \frac{\Phi''(t_2)}{2}\left(\frac{\Delta}{2} + \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)^2,\end{aligned}$$

where $t_2 \in (\frac{-\Delta}{2}, \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}})$. Combining (7.3) and (7.3), we have

$$\begin{aligned}\mathcal{R}(\widehat{\beta}) - \mathcal{R}^* =& \Phi'\left(-\frac{\Delta}{2}\right)\left(\frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} \right) + \frac{\Phi''(t_1)}{2}\left(\frac{\Delta}{2} - \frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)^2 \\ &+ \frac{\Phi''(t_2)}{2}\left(\frac{\Delta}{2} + \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)^2.\end{aligned} \tag{27}$$

We now introduce a lemma that upper bounds the first term on the right-hand side of (27).

Lemma 16 *Suppose* (9b) *holds, and n satisfies that*

$$n \geq C \cdot \sigma_{\max}^2 M^{2+2/(2-q)} R^{2/(2-q)} \log p$$

for some constant C. Then we have

$$\frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} \leq \frac{M}{2\Delta}\|\delta\|_2^2,$$

Note that $\Phi'(-\Delta/2) = (2\pi)^{-1/2}\exp(-\Delta^2/8)$. Following Lemma 16, we have

$$\Phi'\left(-\frac{\Delta}{2}\right)\left(\frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}}\right) \leq \frac{M}{2\sqrt{2\pi}\Delta}\exp\left(-\frac{\Delta^2}{8}\right)\|\delta\|_2^2. \tag{28}$$

Now we consider the second-order term in (27). First, using Lemma 16, we have

$$\begin{aligned}\frac{\Delta}{2} - \frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} &= \frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} + \frac{\widehat{\beta}^\top(\mu_m - \widehat{\mu}_m)}{\widehat{\Delta}} \\ &\leq \frac{M}{2\Delta}\|\delta\|_2^2 + \frac{\widehat{\beta}^\top(\mu^{(0)} - \widehat{\mu}^{(0)}) + \widehat{\beta}^\top(\mu^{(1)} - \widehat{\mu}^{(1)})}{2\widehat{\Delta}}.\end{aligned}$$

After taking square, the first term on the right-hand side gives $\frac{M}{4\Delta^2}\|\delta\|_2^4$, which is negligible compared to the first-order term. Hence it suffices to bound the second term on the right-hand side of (7.3). For this aim we introduce the next lemma.

Lemma 17 *Under the identical conditions as in Theorem 1 or 2, with probability at least $1 - 4p^{-1}$ we have*

$$\left(\frac{\widehat{\beta}^\top(\mu^{(0)} - \widehat{\mu}^{(0)}) + \widehat{\beta}^\top(\mu^{(1)} - \widehat{\mu}^{(1)})}{2\widehat{\Delta}}\right)^2 \leq C \cdot \sigma_{\max}^{-q} M^{1-q} R\left(\frac{\log p}{n}\right)^{1-q/2} \tag{29}$$

for some constant C.

Since $t_1 > -\Delta/2$, we have $|\Phi''(t_1)| \leq C \cdot \Delta \exp\left(-\Delta^2/8\right)$. Combining this with (29), we bound the second term in (27) by

$$\frac{|\Phi''(t_1)|}{2}\left(\frac{\Delta}{2} - \frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}}\right)^2 \leq C \cdot \Delta \exp\left(-\frac{\Delta^2}{8}\right)\sigma_{\max}^{-q} M^{1-q} R\left(\frac{\log p}{n}\right)^{1-q/2}$$

for some constant C. Likewise, the third term in (27) is also subject to this bound.

Finally, plugging (28) and (7.3) into (27), and using (9b), we achieve that

$$\mathcal{R}(\widehat{\beta}) - \mathcal{R}(\beta^*) \leq C \cdot \exp\left(-\frac{\Delta^2}{8}\right)\sigma_{\max}^{-q} M^{3-q}\Delta R\left(\frac{\log p}{n}\right)^{1-q/2}$$

for some constant C, which completes the proof. ■

8. Conclusion and Discussion

In this work, we propose PANDA, a novel one-stage and tuning-insensitive method for high-dimensional linear discriminant analysis. We prove that PANDA achieves the optimal convergence rate in both the estimation error and misclassification rate. Our numerical studies show that PANDA achieves equal or better performance compared with existing methods, and requires less effort in parameter tuning.

Below, we discuss some related work in the existing literature. Besides Gautier et al. (2011), there are other pivotal methods for regression and inverse covariance estimation problems. For examples, Belloni et al. (2011) and Sun and Zhang (2012) propose the scaled Lasso method (also known as square-root Lasso) for sparse linear regression, which enjoys a similar tuning-insensitive property to Gautier et al. (2011); Belloni et al. (2014) extend the scaled Lasso to nonparametric regression; Liu et al. (2015) extend the scaled Lasso to sparse multivariate regression with inhomogeneous noise; Bunea et al. (2013) extend the scaled Lasso to sparse linear regression with group structures; Sun and Zhang (2013) and Liu and Wang (2017) extend the scaled Lasso to inverse covariance matrix estimation; Zhao and Liu (2013) extend Gautier et al. (2011) to inverse covariance matrix estimation for heavy tail elliptical distributions; Belloni and Chernozhukov (2011) and Wang (2013) show that the sparse quantile regression and LAD Lasso are also pivotal methods, which enjoy similar tuning-insensitive properties, respectively.

Acknowledgments

The authors thank the action editor and reviewers for their helpful comments, which led to a substantial improvement of the paper. E. X. Fang is partially supported by NSF DMS-2230795 and DMS-2230797. Y. Mei, Y. Shi, and Q. Xu were partially supported by NSF grant DMS-2015405, NIH grant 1R21AI157618-01A1, and the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002378. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix A. An ADMM Algorithm for Solving (5)

This section discusses the implementation of the ADMM algorithm for solving (5). For that purpose, we first re-write the problem (5) as

$$
\begin{aligned}
(\widehat{\beta}, \widehat{\tau}) \in \underset{\beta,u,v,w,\in\mathbb{R}^p,\tau\in\mathbb{R}}{\arg\min} \quad & \|\beta\|_1 + c\tau^2 \\
\text{subject to} \quad & \widehat{\Sigma}\beta - \lambda\widehat{\sigma}_{\max}\tau\mathbf{1} + u = \widehat{\mu}_d + \lambda\widehat{\sigma}_{\max}\mathbf{1}, \\
& \widehat{\Sigma}\beta + \lambda\widehat{\sigma}_{\max}\tau\mathbf{1} - v = \widehat{\mu}_d - \lambda\widehat{\sigma}_{\max}\mathbf{1}, \\
& w - \widehat{\Sigma}^{1/2}\beta = 0, \\
& u \geq 0, \ v \geq 0, \\
& \|w\|_2 \leq \tau.
\end{aligned}
$$

Note that the first three constraints in (A) are linear and the last three constraints are conic.

To simplify the notation, we write the first three linear constraints as

$$A_\beta\beta + A_u u + A_v v + A_w w + A_\tau\tau = b$$

for some real matrices A_β, A_u, A_v, A_w, A_τ and real vector b. We can further write the problem as

$$
\begin{aligned}
(\widehat{\beta}, \widehat{\tau}) \in \underset{\beta,u,v,w\in\mathbb{R}^p,\tau\in\mathbb{R}}{\arg\min} \quad & \|\beta\|_1 + c\tau^2 \\
\text{subject to} \quad & A_\beta\beta + A_u u + A_v v + A_w w + A_\tau\tau = b, \\
& u, v \in \mathcal{C}_1, \\
& (w, \tau) \in \mathcal{C}_2,
\end{aligned}
$$

where

$$
\begin{aligned}
\mathcal{C}_1 &= \left\{x \in \mathbb{R}^p : x_j \geq 0, \ j \in [p]\right\}, \\
\mathcal{C}_2 &= \left\{(x, y) \in \mathbb{R}^p \times \mathbb{R} : y \geq \sqrt{\sum_{j=1}^p x_j^2}\right\}
\end{aligned}
$$

are two convex cones.

The augmented Lagrangian function with scaled dual variables is

$$L_\rho(\beta, u, v, w, \tau, s) = \|\beta\|_1 + c\tau^2 + \frac{\rho}{2}\|A_\beta\beta + A_u u + A_v v + A_w w + A_\tau\tau - b + s\|_2^2 - \frac{\rho}{2}\|s\|_2^2,$$

where s is the scaled dual variable and $\rho > 0$ is the penalty parameter.

Based on the augmented Lagrangian function above, we can derive the ADMM algorithm described in Algorithm 1 in Section 3.

In this appendix we prove the following theorem from Section 6.2:

Theorem *Let u, v, w be discrete variables such that v, w do not co-occur with u (i.e., $u \neq 0 \Rightarrow v = w = 0$ in a given dataset $\mathcal{D}$). Let N_{v0}, N_{w0} be the number of data points for which*

$v = 0, w = 0$ respectively, and let I_{uv}, I_{uw} be the respective empirical mutual information values based on the sample $\mathcal{D}$. Then

$$N_{v0} \; > \; N_{w0} \;\; \Rightarrow \;\; I_{uv} \; \leq \; I_{uw}$$

with equality only if u is identically 0. ■

Appendix B. Additional Numerical Results

In this section, we present additional simulation results as supplement to Section 5. In subsection B.1, we include results of PANDA performance with different choices of tuning parameter c. In subsection B.2, we report the performance of LPD, AdaLDA and PANDA when we vary the sample size n. In subsection B.3, we present the Area Under the Curve (AUC) of the three methods as another performance metric for LDA.

B.1 PANDA performance with c and λ in Theorem 1

In this subsection, we consider the choice of c and λ as in (7) for our PANDA method in our simulations. Tables 8 and 9 summarizes the performance of our PANDA method with c and λ set as in (7), versus $c = 20$ and λ fine-tuned under the AR(1) model, together with the performance of LPD and AdaLDA for reference. From these tables, we can see that with parameter c set as in (7), the PANDA method may not achieve the most desirable empirical performance, and we thus recommend cross-validation in practice.

Table 8: *The ℓ_2 estimation errors of β^* under the AR(1) model, with $n = 200$ and different (s, p), averaged over 100 replicates. The standard deviations are given in brackets. The lower value at the significance level 0.05 between the AdaLDA and the PANDA method are marked in bold.*

Method	(s,p)					
	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
LPD	1.8875 (0.0494)	1.9607 (0.0313)	1.9846 (0.0101)	1.8960 (0.0416)	1.9669 (0.0199)	1.9868 (0.0094)
AdaLDA	1.8854 (0.0495)	1.9545 (0.0200)	1.9821 (0.0098)	1.8952 (0.0412)	1.9593 (0.0184)	1.9850 (0.0084)
PANDA (with $c = 20$)	**1.8673** (0.0542)	**1.9521** (0.0229)	1.9814 (0.0112)	**1.8856** (0.0460)	1.9571 (0.0190)	**1.9830** (0.0104)
PANDA (with c, λ in Thm 1)	1.9997 (0.0019)	2.0000 (0)	2.0000 (0)	2.0000 (0)	2.0000 (0)	2.0000 ()

B.2 Performance of LPD, AdaLDA and PANDA with different n

Here we present results on the performance of LPD, AdaLDA and our PANDA method with varying sample size. Tables 10 and 11 summarize the ℓ_2 error of β^* estimation and the misclassification rate under the AR(1) model, with $n = 100$, 200 and 400. As can be seen, for every setting of n, the three methods achieve comparable performance.

Table 9: *The misclassification rate under the AR(1) model with different s and p, averaged over 100 replicates. The standard deviations are given in brackets.*

Method	(s,p)					
	$(5,400)$	$(10,400)$	$(20,400)$	$(5,800)$	$(10,800)$	$(20,800)$
LPD	0.2086	0.2900	0.3535	0.2112	0.2908	0.3532
	(0.0074)	(0.0109)	(0.0099)	(0.0074)	(0.0066)	(0.0080)
AdaLDA	0.2082	0.2890	0.3522	0.2120	0.2913	**0.3525**
	(0.0068)	(0.0080)	(0.0075)	(0.0088)	(0.0072)	(0.0082)
PANDA (with $c = 20$)	**0.2068**	0.2886	0.3542	0.2114	0.2910	0.3571
	(0.0069)	(0.0087)	(0.0104)	(0.0084)	(0.0079)	(0.01206)
PANDA (with c, λ in Thm 1)	0.2444	0.3112	0.3671	0.2413	0.3156	0.3749
	(0.0162)	(0.0167)	(0.0115)	(0.0165)	(0.0187)	(0.0192)

Table 10: *The ℓ_2 estimation errors of β^* under the AR(1) model, with different n, s and p, averaged over 100 replicates. The standard deviations are given in brackets. The lower value at the significance level* 0.05 *between the AdaLDA and the PANDA method are marked in bold.*

n	**Specification**						
$n = 100$	(s,p)	$(5,400)$	$(10,400)$	$(20,400)$	$(5,800)$	$(10,800)$	$(20,800)$
	LPD	1.9258	1.9640	1.9814	1.9236	1.9695	1.9834
		(0.0408)	(0.0105)	(0.0109)	(0.0396)	(0.0230)	(0.0077)
	AdaLDA	1.9324	1.9709	1.9896	1.9298	1.9641	1.9946
		(0.0292)	(0.0113)	(0.0135)	(0.0326)	(0.0200)	(0.0118)
	PANDA	1.9161	1.9571	1.9920	**1.9112**	1.9734	1.9944
		(0.0344)	(0.0292)	(0.0199)	(0.0388)	(0.0303)	(0.0140)
$n = 200$	(s,p)	$(5,400)$	$(10,400)$	$(20,400)$	$(5,800)$	$(10,800)$	$(20,800)$
	LPD	1.8875	1.9607	1.9846	1.8960	1.9669	1.9868
		(0.0494)	(0.0313)	(0.0101)	(0.0416)	(0.0199)	(0.0094)
	AdaLDA	1.8854	1.9545	1.9821	1.8952	1.9593	1.9850
		(0.0495)	(0.0200)	(0.0098)	(0.0412)	(0.0184)	(0.0084)
	PANDA	**1.8673**	**1.9521**	1.9814	**1.8856**	1.9571	**1.9830**
		(0.0542)	(0.0229)	(0.0112)	(0.0460)	(0.0190)	(0.0104)
$n = 400$	(s,p)	$(5,400)$	$(10,400)$	$(20,400)$	$(5,800)$	$(10,800)$	$(20,800)$
	LPD	1.8265	1.9456	1.9801	1.8695	1.9824	3.9300
		(0.1903)	(0.0247)	(0.0116)	(0.0601)	(0.0182)	(0.0086)
	AdaLDA	1.8498	1.9399	1.9749	1.8711	1.9452	1.9775
		(0.0764)	(0.0203)	(0.0106)	(0.0370)	(0.0176)	(0.0087)
	PANDA	**1.3936**	**1.9319**	**1.9706**	**1.7353**	**1.9416**	**1.9748**
		(0.3866)	(0.0851)	(0.0221)	(0.3031)	(0.0204)	(0.0109)

Table 11: *The misclassification rate under the AR(1) model, with different n, s and p, averaged over 100 replicates. The standard deviations are given in brackets. The lower value at the significance level* 0.05 *between the AdaLDA and the PANDA method are marked in bold.*

n	**Specification**						
$n = 100$	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
	LPD	0.2241	0.3019	0.3611	0.2339	0.3152	0.3801
		(0.0092)	(0.0086)	(0.0126)	(0.0097)	(0.0110)	(0.0256)
	AdaLDA	0.2166	0.2969	0.3714	0.2181	0.3021	**0.3738**
		(0.0053)	(0.0068)	(0.0173)	(0.0064)	(0.0093)	(0.0106)
	PANDA	0.2170	0.3136	0.3875	0.2212	0.3214	0.4049
		(0.0082)	(0.0224)	(0.0152)	(0.0077)	(0.0093)	(0.0206)
$n = 200$	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
	LPD	0.2086	0.2900	0.3535	0.2112	0.2908	0.3532
		(0.0074)	(0.0109)	(0.0099)	(0.0074)	(0.0066)	(0.0080)
	AdaLDA	0.2082	0.2890	0.3522	0.2120	0.2913	**0.3525**
		(0.0068)	(0.0080)	(0.0075)	(0.0088)	(0.0072)	(0.0082)
	PANDA	**0.2068**	0.2886	0.3542	0.2114	0.2910	0.3571
		(0.0069)	(0.0087)	(0.0104)	(0.0084)	(0.0079)	(0.0121)
$n = 400$	(s, p)	$(5, 400)$	$(10, 400)$	$(20, 400)$	$(5, 800)$	$(10, 800)$	$(20, 800)$
	LPD	0.2000	0.2815	0.3466	0.2017	0.2824	0.3468
		(0.0056)	(0.0058)	(0.0043)	(0.0058)	(0.0055)	(0.0044)
	AdaLDA	0.1989	0.2808	0.3452	0.2003	0.2818	**0**.3466
		(0.0042)	(0.0050)	(0.0043)	(0.0042)	(0.0050)	(0.0051)
	PANDA	**0.1913**	0.2803	0.3454	0.2000	0.2814	0.3472
		(0.0067)	(0.0053)	(0.0072)	(0.0055)	(0.0059)	(0.0074)

B.3 AUC of LPD, AdaLDA and PANDA

Area Under the Curve (AUC) is another performance metric for binary classification, which looks at the trade-off between the precision and recall rate. In Table 12 we report the AUC over the testing data with different s and p, averaged over 100 replicates. As can be seen, the three methods also achieve comparable performance in AUC.

Appendix C. Proofs

This section provides the detailed proofs to the lemmas in the main body of the paper, and is split into eight subsections, one subsection for the proof of each lemma.

C.1 Proof of Lemma 8

Proof There are four main statements in Lemma 8, and let us prove them one by one.

Table 12: *The AUC over testing data, averaged over 100 replicates. The standard deviations are given in brackets.*

Model	**Specification**						
AR(1)	(s,p)	$(5,400)$	$(10,400)$	$(20,400)$	$(5,800)$	$(10,800)$	$(20,800)$
	LPD	0.8770	0.7858	0.7034	0.8699	0.7828	0.7051
		(0.0189)	(0.0251)	(0.0297)	(0.0191)	(0.0234)	(0.0295)
	AdaLDA	0.8773	0.7872	0.7048	0.8698	0.7815	**0.7059**
		(0.0188)	(0.0238)	(0.0270)	(0.0205)	(0.0228)	(0.0298)
	PANDA	0.8784	0.7878	0.7028	0.8700	0.7816	0.7001
		(0.0190)	(0.0245)	(0.0306)	(0.0201)	(0.0252)	(0.0321)
Varying Diagonal	(s,p)	$(5,400)$	$(10,400)$	$(20,400)$	$(5,800)$	$(10,800)$	$(20,800)$
	LPD	0.9898	0.9392	0.8565	0.9899	0.9386	0.8563
		(0.0038)	(0.0125)	(0.0197)	(0.0034)	(0.0109)	(0.0192)
	AdaLDA	0.9899	0.9398	0.8566	0.9900	0.9390	0.8566
		(0.0037)	(0.0119)	(0.0193)	(0.0034)	(0.0106)	(0.0192)
	PANDA	0.9898	0.9401	0.8567	0.9899	0.9390	0.8558
		(.0038)	(0.0117)	(0.0195)	(0.0033)	(0.0108)	(0.0188)
Erdös-Rényi Random Graph	(s,p)	$(5,400)$	$(10,400)$	$(20,400)$	$(5,800)$	$(10,800)$	$(20,800)$
	LPD	0.7826	0.8401	0.9563	0.7992	0.7372	0.7337
		(0.0284)	(0.0236)	(0.0101)	(0.0253)	(0.0332)	(0.0257)
	AdaLDA	0.7845	0.8415	0.9558	0.7995	0.7390	0.7353
		(0.0295)	(0.0241)	(0.0100)	(0.0256)	(0.0325)	(0.0270)
	PANDA	**0.7867**	0.8412	**0.9589**	**0.8039**	**0.7464**	**0.7439**
		(0.0272)	(0.0236)	(0.0098)	(0.0238)	(0.0316)	(0.0278)
Block Sparse	(s,p)	$(5,400)$	$(10,400)$	$(20,400)$	$(5,800)$	$(10,800)$	$(20,800)$
	LPD	0.9183	0.9685	0.9920	0.6130	0.9096	0.6688
		(0.0142)	(0.0077)	(0.0034)	(0.0369)	(0.0127)	(0.0280)
	AdaLDA	0.9093	0.9660	0.9921	0.5869	0.9082	0.6653
		(0.0156)	(0.0083)	(0.0031)	(0.0331)	(0.0134)	(0.0248)
	PANDA	**0.9207**	**0.9696**	**0.9925**	**0.6152**	**0.9113**	**0.6717**
		(0.0129)	(0.0075)	(0.0031)	(0.0361)	(0.0127)	(0.0291)
Approximately Sparse	p	400	800	1200			
	LPD	0.9626	0.9621	0.9624			
		(0.0091)	(0.0086)	(0.0075)			
	AdaLDA	0.9625	0.9626	0.9627			
		(0.0098)	(0.0081)	(0.0080)			
	PANDA	0.9628	0.9621	0.9634			
		(0.0098)	(0.0088)	(0.0082)			

(i) It suffices to show that

$$\mathbb{P}\left(\mathcal{E}_\tau\right) = \mathbb{P}\left(|\beta^{*\top}(\widehat{\Sigma} - \Sigma)\beta^*| \le \frac{1}{2}\beta^{*\top}\Sigma\beta^* \right) \ge 1 - 2e^{-(n-1)/16}.$$

Let $\{Y_i\}_{i=1}^{2n-2}$ be i.i.d. random vectors following the multivariate normal distribution $N(\mathbf{0}, \Sigma)$. Then

$$\widehat{\Sigma} \overset{\mathrm{d}}{=} \frac{1}{2n-2}\sum_{i=1}^{2n-2} Y_i Y_i^\top, \qquad \text{and} \qquad \beta^{*\top}\widehat{\Sigma}\beta^* \overset{\mathrm{d}}{=} \frac{1}{2n-2}\sum_{i=1}^{2n-2}(\beta^{*\top}Y_i)^2,$$

where $\overset{\mathrm{d}}{=}$ denotes equal in distribution. Note that $\{\beta^{*\top}Y_i\}$ are i.i.d Gaussian random variables following distribution $N(0, \beta^{*\top}\Sigma\beta^*)$, thus $\{(\beta^{*\top}Y_i)^2\}$ are i.i.d. sub-exponential random variables, so for any $t \in (0, \beta^{*\top}\Sigma\beta^*)$, we have

$$\mathbb{P}\left(\Big| \frac{1}{2n-2}\sum_i (\beta^{*\top}Y_i)^2 - \beta^{*\top}\Sigma\beta^* \Big| \ge t \right) \le 2\exp\left\{ -\frac{(2n-2)t^2}{8(\beta^{*\top}\Sigma\beta^*)^2} \right\}.$$

Relation (i) follows directly by taking $t = \frac{1}{2}\beta^{*\top}\widehat{\Sigma}\beta^*$, and thus part (i) of Lemma 8 holds.

(ii) Now we need to show that

$$\mathbb{P}\left(\mathcal{E}_{\sigma_{\max}}\right) = \mathbb{P}\left(|\widehat{\sigma}_{\max}^2 - \sigma_{\max}^2| \le \frac{1}{2}\sigma_{\max}^2 \right) \ge 1 - 2pe^{-(n-1)/16}.$$

To prove this, we set $\beta^* = e_j$ for $j \in [p]$ and use (i) with a union bound argument to obtain that

$$\mathbb{P}\left(|\widehat{\Sigma}_{j,j} - \Sigma_{j,j}| \le \frac{1}{2}\Sigma_{j,j},\ \forall j \in [p] \right) \ge 1 - 2pe^{-(n-1)/16},$$

where the event on the left-hand side implies that $|\widehat{\sigma}_{\max}^2 - \sigma_{\max}^2| \le \frac{1}{2}\sigma_{\max}^2$.

(iii) Here it suffices to show that

$$\mathbb{P}\left(\|\mu_d\|_\infty - 2\sqrt{2}\sigma_{\max}\sqrt{\frac{\log p}{n}} \le \|\widehat{\mu}_d\|_\infty \le \|\mu_d\|_\infty + 2\sqrt{2}\sigma_{\max}\sqrt{\frac{\log p}{n}} \right) \ge 1 - 2p^{-1}.$$

Notice that $\widehat{\mu}_d \sim N(\mu_d, \frac{2}{n}\Sigma)$. Let $\mu_{d,j}$ and $\widehat{\mu}_{d,j}$ denote the j-th coordinate of μ_d and $\widehat{\mu}_d$, respectively. We have $\widehat{\mu}_{d,j} \sim N(\mu_{d,j}, \frac{2}{n}\Sigma_{j,j})$. Therefore, for any $j \in [p]$ we have that

$$\mathbb{P}\left(|\widehat{\mu}_{d,j} - \mu_{d,j}| > t\right) \le 2\exp\left\{ \frac{-nt^2}{4(\Sigma_{j,j})^2} \right\} \le 2\exp\left\{ -\frac{nt^2}{4\sigma_{\max}^2} \right\}.$$

Taking $t = \sigma_{\max}\sqrt{\frac{8\log p}{n}}$ and applying the union bound for all $j \in [p]$, we have with probability at least $1 - 2p^{-1}$ that

$$|\widehat{\mu}_{d,j} - \mu_{d,j}| \le \sigma_{\max}\sqrt{\frac{8\log p}{n}},\ \forall j \in [p],$$

which implies that $|\|\widehat{\mu}_d\|_\infty - \|\mu_d\|_\infty| \le 2\sqrt{2}\sigma_{\max}\sqrt{\log p/n}$.

(iv) The lower bound of $\mathbb{P}(\mathcal{E}_1)$ follows an argument in Cai and Zhang (2019). Since $\beta^* = \Sigma^{-1}\mu_d$, we have that $\widehat{\Sigma}\beta^* - \widehat{\mu}_d = (\widehat{\Sigma} - \Sigma)\beta^* - (\widehat{\mu}_d - \mu_d)$. By A.5.1 in the supplement of Cai and Zhang (2019), we have that

$$\mathbb{P}\left(|e_j^\top(\widehat{\Sigma} - \Sigma)\beta^*| \le 10\sigma_{\max}\Delta\sqrt{\frac{\log p}{n}},\ \forall j \in [p]\right) \ge 1 - 2p^{-1},$$

where the event on the left-hand side is equivalent to event $\mathcal{E}_1$. Furthermore, recall that $\Delta^2 = \beta^{*\top}\Sigma\beta^*$. Therefore, under events $\mathcal{E}_\tau$ and $\mathcal{E}_{\sigma_{\max}}$, we have $\Delta \le \sqrt{2}\tau^*$ and $\sigma_{\max} \le \sqrt{2}\widehat{\sigma}_{\max}$. These two conditions and event $\mathcal{E}_1$ together imply $\mathcal{E}_2$.

■

C.2 Proof of Lemma 9

Proof When (β^*, τ^*) is feasible to (5), from the first constraint of (5) we have

$$\|\widehat{\Sigma}\delta\|_\infty = \|\widehat{\Sigma}(\widehat{\beta} - \beta^*)\|_\infty \le \|\widehat{\Sigma}\widehat{\beta} - \widehat{\mu}_d\|_\infty + \|\widehat{\Sigma}\beta^* - \widehat{\mu}_d\|_\infty \le \lambda\widehat{\sigma}_{\max}(\widehat{\tau} + \tau^*) + 2\lambda\widehat{\sigma}_{\max}.$$

In addition, due to the optimality of $(\widehat{\beta}, \widehat{\tau})$, we have

$$\|\widehat{\beta}\|_1 + c\widehat{\tau}^2 \le \|\beta^*\|_1 + c\tau^{*2},$$

which implies that

$$\widehat{\tau} \le \tau^* + \sqrt{\frac{\|\delta\|_1}{c}}.$$

Plugging the above inequality into (C.2), we obtain that

$$\|\widehat{\Sigma}\delta\|_\infty \le 2\lambda\widehat{\sigma}_{\max}(\tau^* + 1) + \lambda\widehat{\sigma}_{\max}\sqrt{\frac{\|\delta\|_1}{c}}.$$

Under the events $\mathcal{E}_\tau$ and $\mathcal{E}_{\sigma_{\max}}$, we have $\tau^* \le \sqrt{\frac{3}{2}}\Delta$ and $\widehat{\sigma}_{\max} \le 2\sigma_{\max}$, so we further have that

$$\|\widehat{\Sigma}\delta\|_\infty \le 2\lambda\sigma_{\max}\left(3\Delta + 2 + \sqrt{\frac{\|\delta\|_1}{c}}\right).$$

Finally, applying Hölder's inequality, we obtain that

$$\delta^\top\widehat{\Sigma}\delta \le \|\delta\|_1\|\widehat{\Sigma}\delta\|_\infty \le 2\lambda\sigma_{\max}\|\delta\|_1\left(3\Delta + 2 + \sqrt{\frac{\|\delta\|_1}{c}}\right).$$

Thus Lemma 9 holds. ■

C.3 Proof of Lemma 10

Proof Lemma 10 is an application of a theorem in Raskutti et al. (2010), which is given by the following lemma.

Lemma 18 (Theorem 1 of Raskutti et al. (2010)) *For any Gaussian random design* $Z \in \mathbb{R}^{n\times p}$ *with i.i.d.* $N(\mathbf{0}, \Sigma)$ *raws, there exist absolute positive constants* c_1, c_2 *such that*

$$\frac{\|Z\delta\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Sigma^{1/2}\delta\|_2 - 9\sigma_{\max}\sqrt{\frac{\log p}{n}}\|\delta\|_1, \ \forall \delta \in \mathbb{R}^p,$$

with probability at least $1 - c_1 \exp(-c_2 n)$.

Now we are ready to prove Lemma 10. Suppose $n \geq 2$. Then the pooled covariance matrix $\widehat{\Sigma}$ is obtained by

$$\widehat{\Sigma} = \frac{1}{2n-2}\left[\sum_{i=1}^{n}\left(X_i^{(0)} - \widehat{\mu}^{(0)}\right)\left(X_i^{(0)} - \widehat{\mu}^{(0)}\right)^\top + \sum_{i=1}^{n}\left(X_i^{(1)} - \widehat{\mu}^{(1)}\right)\left(X_i^{(1)} - \widehat{\mu}^{(1)}\right)^\top\right],$$

and $\widehat{\Sigma}$ has the same distribution as

$$\widetilde{\Sigma} = \frac{1}{2n-2}\sum_{i=1}^{2n-2} Z_i Z_i^\top,$$

where Z_j's are i.i.d. samples from $N(0, \Sigma)$. Hence $\widehat{\Sigma}$ can be viewed as the sample covariance matrix of a Gaussian random design with 0 mean.

By Lemma 18 (i.e., Theorem 1 of Raskutti et al. (2010)), there exist absolute positive constants c_1 and c_2 such that with probability at least $1 - c_1 \exp(-c_2 n)$,

$$\|\widehat{\Sigma}^{1/2}\delta\|_2 \geq \frac{1}{4}\|\Sigma^{1/2}\delta\|_2 - 9\sigma_{\max}\sqrt{\frac{\log p}{2n-2}}\|\delta\|_1.$$

When $n \geq 2$ and $\lambda_{\min} \geq M^{-1}$, we have

$$\|\widehat{\Sigma}^{1/2}\delta\|_2 \geq \frac{1}{4\sqrt{M}}\|\delta\|_2 - 9\sigma_{\max}\sqrt{\frac{\log p}{n}}\|\delta\|_1,$$

and thus

$$\begin{aligned}\delta^\top\widehat{\Sigma}\delta &\geq \left(\frac{1}{4\sqrt{M}}\|\delta\|_2 - 9\sigma_{\max}\sqrt{\frac{\log p}{n}}\|\delta\|_1\right)^2 \\ &\geq \frac{1}{32M}\|\delta\|_2^2 - 81\sigma_{\max}^2\frac{\log p}{n}\|\delta\|_1^2.\end{aligned}$$

Here the last inequality follows from the fact that

$$(a-b)^2 = \left(\frac{1}{2}a^2 - 2ab + 2b^2\right) + \frac{1}{2}a^2 - b^2 \geq \frac{1}{2}a^2 - b^2$$

for any number $a, b \geq 0$. Thus Lemma 10 holds. ■

C.4 Proof of Lemma 11

Proof For any $S \subseteq [p]$, we have

$$\|\widehat{\beta}\|_1 = \|\beta^* + \delta\|_1 \geq \|\beta^*_S\|_1 + \|\delta_{S^c}\|_1 - \|\beta^*_{S^c}\|_1 - \|\delta_S\|_1.$$

Combining the above inequality with $\|\beta^*\|_1 \leq \|\beta^*_S\|_1 + \|\beta^*_{S^c}\|_1$, we have

$$\|\widehat{\beta}\|_1 - \|\beta^*\|_1 \geq \|\delta_{S^c}\|_1 - \|\delta_S\|_1 - 2\|\beta^*_{S^c}\|_1. \tag{30}$$

When (β^*, τ^*) is feasible to (5), by optimality we have

$$\|\widehat{\beta}\|_1 + c\widehat{\tau}^2 \leq \|\beta^*\|_1 + c\tau^{*2}. \tag{31}$$

Combining (30) and (31) yields

$$\|\delta_{S^c}\|_1 - \|\delta_S\|_1 - 2\|\beta^*_{S^c}\|_1 \leq \|\widehat{\beta}\|_1 - \|\beta^*\|_1 \leq c(\tau^{*2} - \widehat{\tau}^2).$$

Since $\tau^{*2} = \beta^{*\top}\widehat{\Sigma}\beta^*$ and $\widehat{\tau}^2 \geq \widehat{\beta}^\top\widehat{\Sigma}\widehat{\beta}$, it follows that

$$\begin{aligned} c(\tau^{*2} - \widehat{\tau}^2) &\leq -2c\delta^\top(\widehat{\Sigma}\beta^*) \\ &= -2c\delta^\top(\widehat{\Sigma} - \Sigma)\beta^* - 2c\delta^\top\mu_d \\ &\leq 2c\|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty\|\delta\|_1 + 2c\|\mu_d\|_\infty\|\delta\|_1. \end{aligned}$$

Under event $\mathcal{E}_1$, we have

$$\|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty \leq 10\sigma_{\max}\Delta\sqrt{\frac{\log p}{n}}.$$

When n satisfies that

$$n \geq 100a^{-2}\sigma^2_{\max}\Delta^2\log p,$$

we have

$$c(\tau^{*2} - \widehat{\tau}^2) \leq 4c\|\mu_d\|_\infty\|\delta\|_1.$$

By setting c as in (7), we have that

$$\frac{1}{2}\|\delta_{S^c}\|_1 \leq \frac{3}{2}\|\delta_S\|_1 + 2\|\beta^*_{S^c}\|_1.$$

Thus $\|\delta_{S^c}\|_1 \leq 3\|\delta_S\|_1 + 4\|\beta^*\|_1$, which completes the proof of Lemma 11. ■

C.5 Proof of Lemma 12

Proof From the definitions of $\mathbb{B}_q(R)$ and S_η, we have that

$$R \geq \sum_j |\beta^*_j|^q \geq \eta^q|S_\eta|,$$

and

$$R \geq \sum_j |\beta^*_j|^q = \sum_j |\beta^*_j| \cdot |\beta^*_j|^{q-1} \geq \eta^{q-1}\|\beta^*_{S^c_\eta}\|_1.$$

Lemma 12 follows immediately from these two inequalities, and thus holds. ■

C.6 Proof of Lemma 13

Proof Let us first prove relation (23a). Under the optimality condition, we have $\|\widehat{\beta}\|_1 + c\widehat{\tau}^2 \le \|\beta^*\|_1 + c\tau^{*2}$, and thus

$$\widehat{\tau}^2 - \tau^{*2} \le \frac{1}{c}\|\delta\|_1 \le C \cdot \Delta(\Delta+1)\sigma_{\max}^{1-q} M^{3/2-q} R \left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}$$

for some positive constant C. Here the last inequality uses (9a) and $\|\mu_d\|_\infty \le M^{1/2}\Delta$.

Note that the second constraint in (5) implies that

$$\widehat{\tau}^2 \ge \widehat{\beta}^\top \widehat{\Sigma} \widehat{\beta} = (\beta^* + \delta)^\top \widehat{\Sigma} (\beta^* + \delta) \ge \tau^{*2} + 2\delta^\top \widehat{\Sigma} \beta^*,$$

hence

$$\begin{aligned} \widehat{\tau}^2 - \tau^{*2} &\ge -2|\delta^\top \widehat{\Sigma} \beta^*| \\ &\ge -2\left|\delta^\top \left[(\widehat{\Sigma} - \Sigma)\beta^* + \mu_d\right]\right| \\ &\ge -2\|\delta\|_2 \|\mu_d\|_2 - 2\|\delta\|_1 \|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty. \end{aligned} \tag{32}$$

Note that under the event $\mathcal{E}_1$, we have

$$\|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty \le 10\sigma_{\max}\Delta\sqrt{\frac{\log p}{n}}. \tag{33}$$

Plugging (9a), (9b), (33) and $\|\mu_d\|_2 \le M^{1/2}\Delta$ into (32), we obtain that

$$\widehat{\tau}^2 - \tau^{*2} \ge -C \cdot \Delta(\Delta+1)\sigma_{\max}^{1-q/2} M^{(3-q)/2} \sqrt{R} \left(\frac{\log p}{n}\right)^{1/2-q/4}.$$

Combining the above equation and (C.6) yields (23a).

Next, let us prove the result (23b) in Lemma 13. Note that the gap between τ^{*2} and Δ^2 can be written as $|\tau^{*2} - \Delta^2| = |\beta^{*\top}(\widehat{\Sigma} - \Sigma)\beta^*|$. To bound this gap, we first apply Hölder's inequality that

$$|\beta^{*\top}(\widehat{\Sigma} - \Sigma)\beta^*| \le \|\beta^*\|_1 \|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty.$$

Under event $\mathcal{E}_1$, the term $\|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty$ can be again bounded by (33). To bound the term $\|\beta^*\|_1$, we note that

$$\|\beta^*\|_1 = \|\beta^*_{S_\eta}\|_1 + \|\beta^*_{S_\eta^c}\|_1 \le \sqrt{|S_\eta|}\ \|\beta^*\|_2 + \|\beta^*_{S_\eta^c}\|_1 \le \eta^{-q/2}\sqrt{R}M^{1/2}\Delta + \eta^{1-q}R.$$

The last inequality above uses equations (18) and (19). By our choice of η in (17), when n satisfies that

$$n \ge C \cdot \Delta^2 \sigma_{\max}^2 M R \log p$$

for some absolute constant C, we have that

$$\|\beta^*\|_1 \le C \cdot \eta^{-q/2}\sqrt{R}M^{1/2}\Delta \le C \cdot \Delta^2 \sigma_{\max}^{-q/2} M^{(1-q)/2} \Delta \sqrt{R} \left(\frac{\log p}{n}\right)^{-q/4}.$$

Hence we have

$$|\tau^{*2} - \Delta^2| = |\beta^{*\top}(\widehat{\Sigma} - \Sigma)\beta^*| \leq \|\beta^*\|_1 \|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty \leq C \cdot \sigma_{\max}^{1-q/2} M^{(1-q)/2} \sqrt{R} \left(\frac{\log p}{n}\right)^{\frac{2-q}{4}},$$

and thus (23b) holds. ■

C.7 Proof of Lemma 14

Proof When (β^*, τ^*) is feasible to (5), from the first constraint of (5) we have

$$\|\widehat{\Sigma}\delta\|_\infty = \|\widehat{\Sigma}(\widehat{\beta} - \beta^*)\|_\infty \leq \|\widehat{\Sigma}\widehat{\beta} - \widehat{\mu}_d\|_\infty + \|\widehat{\Sigma}\beta^* - \widehat{\mu}_d\|_\infty \leq \lambda\widehat{\sigma}_{\max}(\widehat{\tau} + \tau^*) + 2\lambda\widehat{\sigma}_{\max}. \quad (34)$$

When $\widehat{\tau} = \sqrt{\widehat{\beta}^\top \widehat{\Sigma}\widehat{\beta}}$, we have

$$\begin{aligned} \widehat{\tau}^2 &= \widehat{\beta}^\top \widehat{\Sigma}\widehat{\beta} = (\beta^* + \delta)^\top \widehat{\Sigma}(\beta^* + \delta) = \tau^{*2} + 2\delta^\top \widehat{\Sigma}\beta^* + \delta^\top \widehat{\Sigma}\delta \\ &= \tau^{*2} + \delta^\top \widehat{\Sigma}\delta + 2\delta^\top(\widehat{\Sigma} - \Sigma)\beta^* + 2\delta^\top \mu_d \\ &\leq \tau^{*2} + \delta^\top \widehat{\Sigma}\delta + 20\sigma_{\max}\Delta\sqrt{\frac{\log p}{n}}\|\delta\|_1 + 2\|\mu_d\|_2\|\delta\|_2. \end{aligned}$$

Plugging the above inequality into (34), we have

$$\|\widehat{\Sigma}\delta\|_\infty \leq \lambda\widehat{\sigma}_{\max}\left[2\tau^* + 2 + \sqrt{\delta^\top \widehat{\Sigma}\delta} + \left(20\sigma_{\max}\Delta\sqrt{\frac{\log p}{n}}\|\delta\|_1\right)^{1/2} + (2\|\mu_d\|_2\|\delta\|_2)^{1/2}\right].$$

Applying Hölder's inequality, we obtain that

$$\delta^\top \widehat{\Sigma}\delta \leq \lambda\widehat{\sigma}_{\max}\|\delta\|_1\left[2\tau^* + 2 + \sqrt{\delta^\top \widehat{\Sigma}\delta} + \left(20\sigma_{\max}\Delta\sqrt{\frac{\log p}{n}}\|\delta\|_1\right)^{1/2} + (2\|\mu_d\|_2\|\delta\|_2)^{1/2}\right].$$

From the above inequality, we may derive that

$$\begin{aligned} \delta^\top \widehat{\Sigma}\delta \leq & C \cdot \lambda\sigma_{\max}\|\delta\|_1\Big\{\lambda\sigma_{\max}\|\delta\|_1 + \tau^* + 1 + \left(20\sigma_{\max}\Delta\sqrt{\frac{\log p}{n}}\|\delta\|_1\right)^{1/2} \\ & + (2\|\mu_d\|_2\|\delta\|_2)^{1/2}\Big\}, \end{aligned}$$

where C is a constant. ■

C.8 Proof of Lemma 15

Proof When $\widehat{\tau} = \sqrt{\widehat{\beta}^\top \widehat{\Sigma}\widehat{\beta}}$, we have

$$\begin{aligned} \widehat{\tau}^2 &= \widehat{\beta}^\top \widehat{\Sigma}\widehat{\beta} = (\beta^* + \delta)^\top \widehat{\Sigma}(\beta^* + \delta) = \tau^{*2} + 2\delta^\top \widehat{\Sigma}\beta^* + \delta^\top \widehat{\Sigma}\delta \\ &= \tau^{*2} + \delta^\top \widehat{\Sigma}\delta + 2\delta^\top(\widehat{\Sigma} - \Sigma)\beta^* + 2\delta^\top \mu_d. \end{aligned}$$

With event $\mathcal{E}_1$, we have that

$$\begin{aligned}|\widehat{\tau}^2 - \tau^{*2}| &= |\delta^\top \widehat{\Sigma} \delta + 2\delta^\top (\widehat{\Sigma} - \Sigma)\beta^* + 2\delta^\top \mu_d| \\ &\le \delta^\top \widehat{\Sigma}\delta + 20\sigma_{\max}\Delta\sqrt{\frac{\log p}{n}}\|\delta\|_1 + 2\|\mu_d\|_2\|\delta\|_2.\end{aligned}$$

Then, using the previous results (20), (24) and (25), we obtain that

$$|\widehat{\tau}^2 - \tau^{*2}| \le C \cdot \Delta(\Delta+1)\sigma_{\max}^{1-q/2} M^{(3-q)/2}\sqrt{R}\left(\frac{\log p}{n}\right)^{1/2-q/4}$$

for some constant C. ■

C.9 Proof of Lemma 16

Proof Note that

$$\begin{aligned}\widehat{\Delta} = \sqrt{\widehat{\beta}^\top \Sigma \widehat{\beta}} &= \sqrt{\beta^\top \Sigma \beta + 2\beta^\top \Sigma\delta + \delta^\top \Sigma \delta} \\ &\le \sqrt{\beta^\top \Sigma\beta}\left(1 + \frac{2\beta^\top \Sigma \delta + \delta^\top \Sigma\delta}{2\beta^\top \Sigma \beta}\right) \\ &= \Delta + \frac{2\mu_d^\top \delta + \delta^\top \Sigma \delta}{2\Delta}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} &= \frac{1}{2\widehat{\Delta}}(\Delta\widehat{\Delta} - \mu_d^\top \beta) \\ &\le \frac{1}{2\widehat{\Delta}}\left(\Delta^2 + \mu_d^\top(\delta - \beta) + \frac{1}{2}\delta^\top \Sigma\delta\right) \\ &= \frac{1}{4\widehat{\Delta}}\delta^\top\Sigma\delta \le \frac{\delta^\top \Sigma \delta}{4(\Delta + \frac{\mu_d^\top \delta}{\Delta})}.\end{aligned} \tag{35}$$

Note that $|\mu_d^\top \delta| \le \|\mu_d\|_2\|\delta\|_2 \le M^{1/2}\Delta\|\delta\|_2$. Using the convergence rate of $\|\delta\|_2$ in (9b) from Theorem 1, when n satisfies that

$$n \ge C \cdot \sigma_{\max}^2 M^{2+2/(2-q)} R^{2/(2-q)} \log p$$

for some constant C, we have that $|\mu_d^\top \delta| \le \Delta^2/2$, and thus it follows from (35) that

$$\frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} \le \frac{\delta^\top \Sigma \delta}{2\Delta} \le \frac{M}{2\Delta}\|\delta\|_2^2.$$

■

C.10 Proof of Lemma 17

Proof We first show that

$$\mathbb{P}\left(\|\widehat{\mu}^{(\ell)}-\mu^{(\ell)}\|_\infty \le \sigma_{\max}\sqrt{\frac{2\log p}{n}},\ \ell=0,1\right) \ge 1-4p^{-1}.$$

Note that $\widehat{\mu}^{(\ell)} \sim N(\mu^{(\ell)},\Sigma/n)$ for $\ell=0,1$, and thus $\widehat{\mu}_j^{(\ell)} \sim N(\mu_j^{(\ell)},\Sigma_{j,j}/n)$, for $j\in[p]$. Hence,

$$\mathbb{P}\left(|\widehat{\mu}_j^{(\ell)}-\mu_j^{(\ell)}|\ge t\right) \le 2\exp\left(-\frac{nt^2}{\Sigma_{j,j}}\right) \le 2\exp\left(-\frac{nt^2}{\sigma_{\max}^2}\right) \quad \text{for all } \ell\in\{0,1\},\ j\in[p].$$

Taking $t=\sigma_{\max}\sqrt{2\log p/n}$ and applying the union bound for all $j\in[p]$, we have

$$\mathbb{P}\left(\|\widehat{\mu}^{(\ell)}-\mu^{(\ell)}\|_\infty \le \sigma_{\max}\sqrt{\frac{2\log p}{n}},\ \ell=0,1\right) \ge 1-4p\exp(-2\log p)=1-4p^{-1}.$$

We next bound the term $\widehat{\beta}^\top(\mu^{(\ell)}-\widehat{\mu}^{(\ell)})$ for $\ell=0,1$. Note that

$$\begin{aligned}\widehat{\beta}^\top(\mu^{(\ell)}-\widehat{\mu}^{(\ell)}) &= (\beta^*+\delta)^\top(\mu^{(\ell)}-\widehat{\mu}^{(\ell)})\\ &\le (\|\beta^*_{S_\eta}\|_1+\|\beta^*_{S_\eta^c}\|_1+\|\delta\|_1)\|\mu^{(\ell)}-\widehat{\mu}^{(\ell)}\|_\infty\\ &\le \left(\sqrt{|S_\eta|}\|\beta^*\|_2+5\|\beta^*_{S_\eta^c}\|_1+4\sqrt{|S_\eta|}\|\delta\|_2\right)\|\mu^{(\ell)}-\widehat{\mu}^{(\ell)}\|_\infty.\end{aligned}$$

Here the last inequality uses (20). Also, note that $\|\beta^*\|\le M^{1/2}\Delta$. With our choice of η in (17) and the upper bound for $\|\delta\|_2$, when n satisfies that

$$n \ge C\cdot\sigma_{\max}^2 M\Delta^{-\frac{4}{2-q}}R^{\frac{2}{2-q}}\log p$$

for some constant C, we have that

$$\widehat{\beta}^\top(\mu^{(\ell)}-\widehat{\mu}^{(\ell)}) \le \sigma_{\max}^{-q/2}M^{\frac{1-q}{2}}\Delta\sqrt{R}\left(\frac{\log p}{n}\right)^{\frac{2-q}{4}}. \tag{36}$$

We then consider the term $\widehat{\Delta}=\sqrt{\widehat{\beta}^\top\Sigma\widehat{\beta}}$. Note that

$$\widehat{\Delta}^2=\widehat{\beta}^\top\Sigma\widehat{\beta}=\Delta^2+2\mu_d^\top\delta+\delta^\top\Sigma\delta,$$

Hence we have

$$|\widehat{\Delta}^2-\Delta^2| \le 2\|\mu_d\|\|\delta\|_2+M\|\delta\|_2^2.$$

When n is sufficiently large, we have that $|\widehat{\Delta}^2-\Delta^2|\le\frac{1}{2}\Delta^2$. Combining this with (36), we have that

$$\left(\frac{\widehat{\beta}^\top(\mu^{(0)}-\widehat{\mu}^{(0)})+\widehat{\beta}^\top(\mu^{(1)}-\widehat{\mu}^{(1)})}{2\widehat{\Delta}}\right)^2 \le C\cdot\sigma_{\max}^{-q}M^{1-q}R\left(\frac{\log p}{n}\right)^{1-q/2}$$

for some constant C. Therefore, Lemma 17 holds true. ■

Appendix D. Review of Gautier's method

In this section, we provide a brief review of Gautier's pivotal method for high-dimensional linear regression in Gautier et al. (2011) that inspires our work. Note that they consider a more complicated high-dimensional instrumental variables model. Here we discuss the particular case where the regressors and instruments are identical for ease of presentation. Specifically, let $X \in \mathbb{R}^{n\times p}$ be a design matrix with n observations and p variables, and let $y \in \mathbb{R}^n$ be the response vector. We consider the following linear model that

$$y = X\beta^* + \varepsilon \quad \text{with} \quad \varepsilon \sim N(0, \sigma^2 I_n),$$

where $\beta^* \in \mathbb{R}^p$ is the unknown regression coefficient with $\|\beta^*\|_0 = s < n \ll p$, and ε is the noise. The Gautier's estimator can be viewed as a variant of the Dantzig selector (Candes and Tao, 2007), and is the optimal solution to the following convex optimization problem that

$$\begin{aligned} (\widehat{\beta}, \widehat{\gamma}) &= \underset{\beta,\gamma}{\operatorname{argmin}} \ \|\beta\|_1 + c\gamma, \\ \text{subject to } & \frac{1}{n}\|X^\top(Y - X\beta)\|_\infty \le \lambda\gamma, \quad \frac{1}{n}\|Y - X\beta\|_2^2 \le \gamma^2, \end{aligned}$$

where c and λ are two tuning parameters, and $\widehat{\gamma}$ is an estimator of σ. The theoretical analysis in Gautier et al. (2011) suggests that the tuning parameter c can be set as a constant between 0 and 1, and the tuning parameter λ can be chosen as

$$\lambda = A \cdot \sqrt{\frac{2\log p}{n}},$$

where A is a constant independent of σ. Therefore, the Gautier's estimator is less sensitive to the parameter tuning than the Dantzig selector, where the tuning parameter depends on σ.

Appendix E. Numerical study on performance of Lasso, Dantzig Seector and Gautier's method

In this section, we provide additional numerical results to compare the performance of Lasso, Dantzig Selector and Gautier's method for linear regression in high dimensions.

We generate the data by a process considered in Candes and Tao (2007). To be more specific, we set $n = 100$, $p = 200$, $s = 5, 10, 20$. We generate the rows of X from the standard Gaussian distribution and then normalize each row of X. For β^*, we set

$$\beta_i^* = u_i(1 + |a_i|) \text{ for } i = 1, \cdots, s,$$

where $u_i = \pm 1$ with probability 1/2, and $a_i \sim N(0,1)$ and is independent of u_i. Meanwhile, we set $\sigma = \sqrt{\frac{s}{n}}$. To fine-tune the parameter, we generate an independent validation set with same sample size $n = 100$ as the training set. We let $\lambda = \widetilde{\lambda}\sqrt{\frac{\log p}{n}}$ for all the three methods, and we tune the factor $\widetilde{\lambda}$ over a range from 0 to 1 for each method. Figure 4 shows the results of the estimation error $\|\widehat{\beta} - \beta^*\|_2$ versus the $\widetilde{\lambda}$ value in the three methods, averaged over 100 replicates under each setting of different p and s. For Gautier's method, the result is not sensitive to the parameter c as long as c is not too small, and we set $c = 20$. Table 13 summarizes the estimation error $\|\widehat{\beta} - \beta^*\|_2$ under different p and s. As can be seen, the three methods have comparable performance in β^* estimation after fine-tuning.

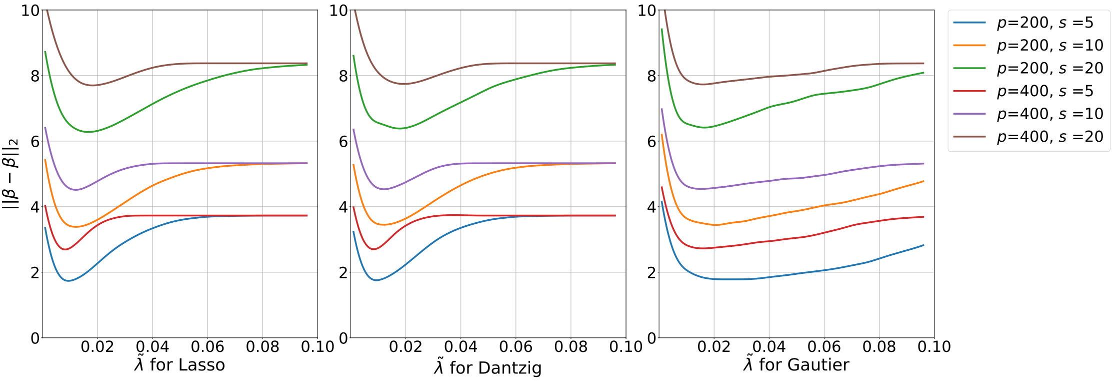


Figure 4: *ℓ_2 estimation error v.s. value of tuning parameter $\widetilde{\lambda}$ in Lasso (left), Dantzig selector (middle) and Gautier's method (right). The results are averaged over 100 replicates.*

Table 13: *The ℓ_2 error of β^* estimation for the regression example. The testing errors are averaged over 100 replicates. The standard deviation of the testing errors are given in brackets.*

Method	(s, p)					
	$(5, 200)$	$(10, 200)$	$(20, 200)$	$(5, 400)$	$(10, 400)$	$(20, 400)$
Lasso	1.801 (0.325)	3.425 (0.517)	6.546 (0.969)	2.738 (0.440)	4.609 (0.492)	7.748 (0.460)
Dantzig Selector	1.802 (0.343)	3.466 (0.495)	6.389 (0.634)	2.757 (0.449)	4.600 (0.500)	7.744 (0.471)
Gautier's Method	1.771 (0.341)	3.412 (0.406)	6.375 (0.620)	2.749 (0.441)	4.653 (0.548)	7.741 (0.483)

Appendix F. Technical derivation on the penalty term in PANDA

In this section, we provide a deep insight on how to non-trivially modify Gautier's pivotal method to our context. To be more specific, we compare the penalty term imposed in Gautier's pivotal method and our proposed PANDA, and explain our choice of a quadratic penalty for τ in (5). For simplicity, we consider the case where $q = 0$ and $|\mathrm{supp}(\beta^*)| \leq s$.

Let $S = \mathrm{supp}(\beta^*)$. For both Gautier's method and PANDA, a key step to derive the upper bound of $\|\delta\|_1 = \|\widehat{\beta} - \beta^*\|_1$ is to show that δ belongs to some restricted subset $\mathcal{C}_{S,\beta^*}$ with high probability, where $\mathcal{C}_{S,\beta^*}$ is defined in (7.1). Note that when $q = 0$, $\|\beta^*_{S^c}\|_1 = 0$, such that $\mathcal{C}_{S,\beta^*}$ reduces to

$$\mathcal{C}_S = \left\{\delta \in \mathbb{R}^p : \|\delta_{S^c}\|_1 \leq 3\|\delta_S\|_1\right\}.$$

In Gautier's method, it is shown that with high probability, (β^*, σ^*) is feasible to the program (D), where β^* is the true regression parameter and

$$\sigma^* := \frac{1}{\sqrt{n}}\|Y - X\beta^*\|_2.$$

Then, by the optimality condition of the solution $\widehat{\beta}$, i.e. $\|\widehat{\beta}\|_1 + c\widehat{\sigma} \leq \|\beta^*\|_1 + c\sigma^*$, $\|\delta_{S^c}\|_1$ can be upper bounded by

$$\begin{aligned}
\|\delta_{S^c}\|_1 &\leq \|\delta_S\|_1 + \frac{c}{\sqrt{n}}\left(\|Y - X\beta^*\|_2 - \|Y - X\widehat{\beta}\|_2\right) \\
&\leq \|\delta_S\|_1 + \frac{c}{\sqrt{n}}\delta^\top \frac{X^\top(Y - X\beta^*)}{\|Y - X\beta^*\|_2} \\
&\leq \|\delta_S\|_1 + c\|\delta\|_1 \frac{\|\frac{1}{n}X^\top(Y - X\beta^*)\|_\infty}{\sigma^*} \\
&\leq \|\delta_S\|_1 + c\lambda\|\delta\|_1,
\end{aligned}$$

where the second inequality uses the convexity of $\|Y - X\beta\|_2$ in β, the third inequality uses Hölder's inequality and the definition of σ^*, and the last inequality is due to the first constraint in (D). With properly chosen c and λ, it can be shown that $\delta \in \mathcal{C}_S$ with high probability.

For PANDA, if we follow the above framework and impose the same penalty $c\tau$, a similar argument leads to

$$\begin{aligned}
\|\delta_{S^c}\|_1 &\leq \|\delta_S\|_1 + c\|\delta\|_1 \frac{\|\widehat{\Sigma}\beta^*\|_\infty}{\sqrt{\beta^*\widehat{\Sigma}\beta^*}} \\
&\leq \|\delta_S\|_1 + c\|\delta\|_1 \frac{\|\widehat{\Sigma}\beta^* - \widehat{\mu}_d\|_\infty + \|\widehat{\mu}_d\|_\infty}{\sqrt{\beta^*\widehat{\Sigma}\beta^*}} \\
&\leq \|\delta_S\|_1 + c\|\delta\|_1 \left(\lambda + \frac{\lambda + \|\widehat{\mu}_d\|_\infty}{\sqrt{\beta^{*\top}\widehat{\Sigma}\beta^*}}\right).
\end{aligned}$$

Note that $\sqrt{\beta^{*\top}\widehat{\Sigma}\beta^*}$ converges to Δ, and thus the term $\frac{\|\widehat{\mu}_d\|_\infty}{\Delta}$ dominates the last term, and the choice of c must rely on the unknown Δ to ensure that $\delta \in \mathcal{C}_S$ with high probability.

In other words, we cannot directly follow Gautier's framework to impose the penalty $c\tau$. Nevertheless, Gautier's method inspires us to impose a quadratic penalty term on τ, by which it turns out that the tuning parameters will no longer rely on the unknown Δ.

Here we remark that in order to guarantee the tuning-insensitive property of our PANDA method, the penalty on τ must be quadratic. Suppose we consider an increasing and convex penalty function $f(\tau)$ instead. Technically, in order to guarantee that $\delta = \widehat{\beta} - \beta^*$ belongs to the restricted set

$$\mathcal{C}_{S,\beta^*} := \{\delta \in \mathbb{R}^p : \ \|\delta_{S^c}\|_1 \leq 3\|\delta_S\|_1 + 4\|\beta^*_{S^c}\|_1\}$$

with high probability, we require f to satisfy that $f(\tau^*) - f(\widehat{\tau}) \leq \frac{1}{2}\|\delta\|_1$, where τ^* is close to Δ. Following the argument in the proof of Lemma 7, we can derive an upper bound for $f(\tau^*) - f(\widehat{\tau})$ as follows:

$$\begin{aligned} f(\tau^*) - f(\widehat{\tau}) &\leq f\left(\sqrt{\beta^{*\top}\widehat{\Sigma}\beta^*}\right) - f\left(\sqrt{\widehat{\beta}^{\top}\widehat{\Sigma}\widehat{\beta}}\right) \\ &\leq \left|\frac{f'(\tau^*)}{\tau^*}\right| \|\widehat{\Sigma}\beta^*\|_\infty \|\delta\|_1 \\ &\leq \left|\frac{f'(\tau^*)}{\tau^*}\right| \left(\|\mu_d\|_\infty + \|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty\right) \|\delta\|_1. \end{aligned}$$

In order to control that $f(\tau^*) - f(\widehat{\tau}) \leq \frac{1}{2}\|\delta\|_1$, we need $\left|\frac{f'(\tau^*)}{\tau^*}\right| \left(\|\mu_d\|_\infty + \|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty\right) \leq \frac{1}{2}$. When n is sufficiently large, the term $\|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty$ here is small, and $\|\mu_d\|_\infty$ can be closely estimated from the sample. Therefore, we require the term $\frac{f'(\tau^*)}{\tau^*}$ to be controlled by some constant that is independent of τ^* or Δ. To satisfy this, the Taylor expansion of f can only have non-zero coefficient for the first-order term, while the coefficients for other orders must be zero, implying that f is a quadratic function.

References

Theodore Wilbur Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley Interscience, New York, 3 edition, 2003.

Alexandre Belloni and Victor Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014.

Peter J Bickel, Elizaveta Levina, et al. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Florentina Bunea, Johannes Lederer, and Yiyuan She. The group square-root lasso: theoretical properties and fast algorithms. *IEEE Transactions on Information Theory*, 60(2): 1313–1325, 2013.

Tony Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577, 2011.

Tony Cai and Linjun Zhang. High dimensional linear discriminant analysis: Optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):675–705, 2019.

Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4):406–413, 2011.

Jianqing Fan and Yingying Fan. High-dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36(6):2605, 2008.

Jianqing Fan, Yang Feng, and Xin Tong. A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):745–771, 2012.

Eric Gautier, Alexandre Tsybakov, and Christiern Rose. High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*, 2011.

Irina Gaynanova and Mladen Kolar. Optimal variable selection in multi-group sparse discriminant analysis. *Electronic Journal of Statistics*, 9(2):2007–2034, 2015. ISSN 1935-7524. doi: 10.1214/15-EJS1064. URL `http://dx.doi.org/10.1214/15-EJS1064`.

Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, and Mark A Caligiuri. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

Mladen Kolar and Han Liu. Optimal feature selection in high-dimensional discriminant analysis. *IEEE Transactions on Information Theory*, 61(2):1063–1083, 2015. doi: 10.1109/TIT.2014.2381241.

Han Liu and Lie Wang. TIGER: A tuning-insensitive approach for optimally estimating gaussian graphical models. *Electronic Journal of Statistics*, 11(1):241–294, 2017.

Han Liu, Lie Wang, and Tuo Zhao. Calibrated multivariate regression with application to neural semantic basis discovery. *Journal of Machine Learning Research: JMLR*, 16:1579, 2015.

Qing Mai, Hui Zou, and Ming Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42, 2012.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*, volume 13. SIAM, 1994.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259, 2010.

Jun Shao, Yazhen Wang, Xinwei Deng, and Sijian Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, 39(2):1241–1265, 2011.

Yu Shi, Daoqing Dai, Chaochun Liu, and Hong Yan. Sparse discriminant analysis for breast cancer biomarker identification and classification. *Progress in Natural Science*, 19(11): 1635–1641, 2009.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

Tingni Sun and Cun-Hui Zhang. Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research*, 14(1):3385–3418, 2013.

Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

Lie Wang. The L_1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151, 2013.

Daniela M Witten and Robert Tibshirani. Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772, 2011.

Okito Yamashita, Masa-aki Sato, Taku Yoshioka, Frank Tong, and Yukiyasu Kamitani. Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage*, 42(4):1414–1429, 2008.

Tuo Zhao and Han Liu. Sparse precision matrix estimation with calibration. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pages 2274–2282, 2013.