Electronic Journal of Statistics
Vol. 11 (2017) 2499–2518
ISSN: 1935-7524
DOI: 10.1214/17-EJS1288

L1 least squares for sparse high-dimensional LDA*

Yanfang Li and Jinzhu Jia

School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing, China,100871
e-mail: liyanfang1110@pku.edu.cn; jzjia@pku.edu.cn

Abstract: This paper studies high-dimensional linear discriminant analysis (LDA). First, we review the ℓ_1 penalized least square LDA proposed in [10], which could circumvent estimation of the annoying high-dimensional covariance matrix. Then detailed theoretical analyses of this sparse LDA are established. To be specific, we prove that the penalized estimator is ℓ_2 consistent in high-dimensional regime and the misclassification error rate of the penalized LDA is asymptotically optimal under a set of reasonably standard regularity conditions. The theoretical results are complementary to the results to [10], together with which we have more understanding of the ℓ_1 penalized least square LDA (or called Lassoed LDA).

Keywords and phrases: High-dimensional LDA, Lasso, sparsity.

Received February 2016.

Contents

*Supported in part by the National Science Foundation of China (11101005, 11571021). This research was also supported by the Key Lab of Mathematical Economics and Quantitative Finance (Ministry of Education), the Key lab of Mathematics and Applied Mathematics (Ministry of Education), and the MOE-Microsoft Key Laboratory of Statistics and Information Technology of Peking University.

1. Introduction

Classification problem is important in a lot of fields such as pattern recognition, bioinformatics etc. There are a few classic classification methods, including LDA (linear discriminant analysis), logistic regression, naive bayes, SVM (support vector machine). LDA is very popular due to its simplicity, robustness and great performances in practice. When the number of features (or predictors) denoted by p is fixed, under some regularity conditions, LDA is proved to be optimal (see standard statistical text books such as [1]).

However, recent technology makes it easy to obtain data with very large number of features, which makes it challenging to apply LDA in practice. One problem is that when p is bigger than n (the number of observations), the covariance matrix of predictors performs very poorly – it is even not invertible. [2] pointed out that the LDA performs poorly and can even perform just as random guessing when p is big. A lot of researchers noticed that in high-dimensional classification problems, it is critical to have a "good" estimation of the covariance matrix. [2] showed that one could use a diagonal matrix instead, which used the idea of Naive Bayes and assumed that features are independent. [7] pointed out that even if a "Naive Bayes" (or independence) rule is used, if there are too many features in the model, the performance of LDA is still poor. So they proposed using only a small number of selected features. But the assumption that features are independent is a little bit annoying. [15] proposed covariance-regularized method to estimate covariance matrix by shrunken method. [12] studied sparse LDA by assuming both the covariance matrix and the difference between the mean vectors of two classes are sparse. Let Σ be the shared covariance matrix and δ be the difference between the mean vectors. [6] proposed a sparse LDA by directly assuming that $\Sigma^{-1}\delta$ is sparse and their method circumvents estimation of the inverse covariance matrix directly.

We'd like to emphasize here that there is a perfect connection between LDA and the least squares. This connection was first established by [8]. Using this connection, in fact, one could solve the LDA problem by directly using the vanilla ℓ_1 penalized least squares, i.e. the Lasso ([13]). Using Lasso to solve sparse LDA has already been proposed in [10]. They call it Lassoed discriminant analysis. [10] showed that under irrepresentable condition ([14]) and some other regularity conditions, the Lassoed discriminant analysis could consistently identify the important features (or predictors) in high-dimensional regime. In this paper, we mainly study the theoretical properties of the Lasso for solving sparse high-dimensional LDA problems, which is a complementary to the results to [10]. The primary contributions of this paper are as follows.

1. Under the restricted eigenvalue condition on the covariance matrix and some other regularity conditions, the Lassoed LDA estimator (defined later) is proved to be ℓ_2 consistent in high-dimensional regime.
2. The misclassification error rate of the Lassoed discriminant analysis tends to be asymptotically optimal.

We want to emphasize that the analysis for sparse LDA via the Lasso is quite

different from that for the Lasso in sparse linear regression model. The main reason is that for the Lasso problem, the response y is a linear combination of predictors $X_1, \ldots, X_p$ plus additive noise. But for LDA, although least squares $\min_\beta \|Y - X\beta\|_2^2$ is used, there is not any functional connection between binary vector Y and the predictor matrix X.

The rest of this paper is organized as follows. In Section 2, we first give a brief and clear review of the Lassoed discriminant analysis proposed in [10], and then we establish the ℓ_2 consistency of the Lasso estimator to solve sparse high-dimensional LDA problems. We also prove that the misclassification error rate of the Lasso tends to be asymptotically optimal. We conclude in Section 3. All proofs are postponed into appendix (Section 4).

2. Sparse LDA and its asymptotic properties

In this section, we first review the procedure of the sparse LDA via the penalized least squares (Lasso). Then we provide the ℓ_2 consistency of the Lasso to solve high-dimensional LDA problems under the restricted eigenvalue condition and some other mild regularity conditions. Restricted eigenvalue condition on the covariance matrix, has proved to be much weaker than the irrepresentable condition. At last, we establish the asymptotic optimality of the corresponding misclassification error rate in the sense that the Lasso tends to achieve the Bayes error.

Throughout this paper, we assume that $\left\{x_i^{(g)}, i = 1, \cdots, n_g\right\}$ are generated independently from normal distributions $\mathrm{N}(\mu^{(g)}, \Sigma)$, which share the same covariance matrix Σ but different means $\mu^{(g)}$, where $g = 1, 2$. Let $\mu = \left(\mu^{(1)} + \mu^{(2)}\right)/2$, $\delta = \mu^{(1)} - \mu^{(2)}$ be the mean of and the difference beween two population means, $n = n_1 + n_2$ be the total sample size, Δ_p^2 be the Mahalanobis squared distance between two populations

$$\Delta_p^2 := \left(\mu^{(1)} - \mu^{(2)}\right)^{\mathrm{T}} \Sigma^{-1} \left(\mu^{(1)} - \mu^{(2)}\right),$$

and $\hat{\mu}^{(g)}$, $\hat{\Sigma}^{(g)}$ be the well-known maximum likelihood estimates. All C's below are positive constant but differ from one to another.

2.1. Review of sparse LDA

To make our paper self-contained, we first review LDA and explain the connection between LDA and least squares in low-dimensional case. Then we breifly introduce one direct method for solve sparse high-dimensional LDA problem proposed in [10], which used the perfect connection between LDA and the least squares.

The classic LDA approaches the classification problem by applying the Bayes rule and classifies a new data point with equal prior weight for each class to $g = 1$, at population level, if and only if

$$\mathrm{U}(x) := (x - \mu)^{\mathrm{T}} \beta^* > 0, \tag{2.1}$$

where $\beta^* = \Sigma^{-1}\delta$, which is called the *fisher classification direction*. However the population parameters in (2.1) are always unknown in practice, their plug-in estimates

$$\hat{\mu} = \frac{1}{2}\left(\hat{\mu}^{(1)} + \hat{\mu}^{(2)}\right), \quad \hat{\Sigma} = \frac{1}{n_1 + n_2}\left(n_1\hat{\Sigma}^{(1)} + n_2\hat{\Sigma}^{(2)}\right)$$

are used instead to get the sample LDA function

$$\mathrm{W}(x) := (x - \hat{\mu})^{\mathrm{T}}\,\hat{\beta}^{lda}, \tag{2.2}$$

where $\hat{\beta}^{lda} = \hat{\Sigma}^{-1}\left(\hat{\mu}^{(1)} - \hat{\mu}^{(2)}\right)$.

The LDA function (2.2) could be explained from another point of view. First we give a label to each data point. Let the label be 1 if the data point is from $\mathrm{N}\left(\mu^{(1)}, \Sigma\right)$ and -1 if it is from $\mathrm{N}\left(\mu^{(2)}, \Sigma\right)$. In fact, any binary code is OK. Then, for simple notation, we pool all of these data points together. Let $z_i = x_i^{(1)}$ for $i = 1, 2, \cdots, n_1$ and $z_{n_1+k} = x_k^{(2)}$ for $k = 1, 2, \cdots, n_2$. Therefore the class label $y_i = 1$ for $i = 1, 2, \cdots, n_1$ and $y_{n_1+k} = -1$ for $k = 1, 2, \cdots, n_2$. We also define the centered version of labels and design matrix as follows. Let $\tilde{y}_i = y_i - \frac{1}{n}\sum_{i=1}^{n} y_i$ and $x_i = z_i - \frac{1}{n}\sum_{i=1}^{n} z_i$. $\tilde{y}_i$'s are the centered class labels and $X = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^{n\times p}$ are called the centered design matrix.

Using notations above, the following lemma gives the connection between LDA and the least quares, see Chaper 4 of [9], from which we can estimate the fisher classification direction(β^*) without the estimates of the mean vectors and the inverse of the covariance matrix separately.

Lemma 2.1. *Consider the following least squares problem*

$$\left(\hat{\beta}^{ols}, \hat{\beta}_0^{ols}\right) = \arg\max_{\beta,\beta_0} \sum_{i=1}^{n} \left[y_i - \left(z_i^{\mathrm{T}}\beta + \beta_0\right)\right]^2.$$

We have that, there exists a positive constant C such that $\hat{\beta}^{ols} = C\hat{\beta}^{lda}$.

Lemma 2.1 gives a perfect connection between LDA and the least squares, though it only holds for low-dimensional settings. The good property of this connecting is that, we could circumvent estimation of the annoying high-dimensional covariance matrix when we extend LDA to high-dimensional regime via the least squares. [10] used this connection and proposed a direct approach to sparse discriminant analysis defined as follows

$$\left(\hat{\beta}(\lambda), \hat{\beta}_0(\lambda)\right) = \arg\max_{\beta,\beta_0} \frac{1}{2n}\sum_{i=1}^{n} \left[y_i - \left(z_i^{\mathrm{T}}\beta + \beta_0\right)\right]^2 + \lambda\|\beta\|_1. \tag{2.3}$$

[10] mainly showed that the Lasso estimator, defined in (2.3), also called Lassoed LDA estimator can consistently identify the important features under the irrepresentable condition and some other regularity conditions. In the next subsection, we will derive the ℓ_2 consistency of the Lasso estimator to solve sparse high-dimensional LDA problems under the restricted eigenvalue condtion on the covariance matrix, together with a set of reasonably standard mild conditions.

To be more specific, the whole procedure using the Lasso to solve sparse LDA could be described in Algorithm 1.

Algorithm 1 Sparse LDA procedure

Require: $(z_i, y_i), i = 1, 2, \ldots, n$

1. Calculate $\hat{\mu}^{(1)}, \hat{\mu}^{(2)}$ ▷ the mean vectors of two classes
2. Solve the Lasso problem defined in (2.3). ▷ discriminant direction
3. Give discriminant function ▷ discriminant function

$$\mathrm{W}(x) = \left(x - \frac{\hat{\mu}^{(1)} + \hat{\mu}^{(2)}}{2}\right)^{\mathrm{T}} \hat{\beta}(\lambda).$$

return $\hat{\beta}(\lambda)$ – fisher discriminant direction; and $W(x)$ – the discriminant function for a new data point with features x.

2.2. ℓ_2 consistency

This subsection will focus on the ℓ_2 consistency of the Lassoed LDA estimator. Though $\hat{\beta}(\lambda)$ is not close to β^*, it can be proved to be ℓ_2 consistent to an intermediate variable $\tilde{\beta}$ defined in (2.4), which is a positive constant multiple of β^*, see Lemma 2.2. From the property of hyperplane, the linear seperator with $\tilde{\beta}$ is the same as that with β^*.

$$\tilde{\beta} = \left[\Sigma + \frac{n_1 n_2}{n^2}\left(\mu^{(1)} - \mu^{(2)}\right)\left(\mu^{(1)} - \mu^{(2)}\right)^{\mathrm{T}}\right]^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)\frac{2n_1 n_2}{n^2}. \quad (2.4)$$

Lemma 2.2. *There exists a positive constant C such that $\tilde{\beta} = C\beta^*$. Specifically, C can be displayed as*

$$C = \left[2\theta_n(1-\theta_n) - \theta_n(1-\theta_n)\left(\mu^{(1)} - \mu^{(2)}\right)^{\mathrm{T}}\tilde{\beta}\right],$$

where $\theta_n = n_1/n$ is the sample size ratio of the first class.

This result could be seen from [1].

In the following, we will derive the conditions under which the ℓ_2 consistency could be hold. Certainly, the Lassoed LDA can not work for an arbitary p. First, we need to control the sample size (n), the number of predictors (p) and the number of relevant features of discriminant direction ($q := \#\{j : \beta_j^* \neq 0\}$) to satisfy the following condition (C1). When Δ_p^2 is too small, which is to say that the two populations $\mathrm{N}(\mu^{(g)}, \Sigma)$ are too close, we can not expect any classifier to peforme well since the Bayes rule in this case is just as random guessing, which could be seen directly from (2.7) in the next subsection. So we need to bound Δ_p^2 away from below. We also need to balance the sample sizes of two classes and control the maximum eigenvalue of the covariance matrix, which are commonly used in high-dimensional settings. These conditions are listed in (C2). The key condition for the ℓ_2 consistency of the Lasso-type estimator in high-dimensional linear model is the restricted eigenvalue condtion, which was first proposed in [3] and whose definition is defined in Definition 2.3.

(C1) $q = o(\sqrt{n/\log p})$ and $\log p = o(n)$.

(C2) $\theta_n(1-\theta_n)$ and Δ_p are bounded away from zero; $\lambda_{\max}(\Sigma)$ is bounded from above.
(C3) X (centered design matrix)satisfies restricted eigenvalue condition RE$(\gamma, 3)$.

Definition 2.3 (Restricted Eigenvalue Condition RE(γ,α))**.** The design matrix X satisfies restricted eigenvalue condition RE(γ,α), if there exists a $\gamma > 0$ such that

$$u^{\mathrm{T}}\left(\frac{1}{n}X^{\mathrm{T}}X\right)u \geq \gamma\|u\|_2^2,$$

for all u satisfying

$$\|u_{S^c}\|_1 \leq \alpha\|u_S\|_1.$$

A few class of matrices have been proved to satisfy the restricted eigenvalue condition with high probability([11, 5]), for example, the matrices of which all entries are i.i.d. from a broad class of multivariate normal distributions. Restricted eigenvalue condition has been proved to be nearly necessary to control the ℓ_2 error in minimax setting in high-dimensional regression. Now under the conditions given above, we give the first result of this paper.

Theorem 2.4. *Under conditions (C1), (C2), (C3), and choosing* $\lambda = C_\lambda\sqrt{\log p/n}$ *for some positive constant* C_λ *depending on* $\{\theta_n, \Delta_p^2, \lambda_{\max}(\Sigma)\}$*, with probability converging to 1, we have*

$$\hat{\beta}(\lambda) - \tilde{\beta} \to 0.$$

More specifically, with probability greater than $1 - O(p^{-1})$*, we have*

$$\left\|\hat{\beta}(\lambda) - \tilde{\beta}\right\|_1 \leq Cq\sqrt{\frac{\log p}{n}},$$
$$\left\|\hat{\beta}(\lambda) - \tilde{\beta}\right\|_2 \leq C\sqrt{\frac{q\log p}{n}},$$

where C *is a sufficiently large constant.*

From the distance between $\hat{\beta}(\lambda)$ and $\tilde{\beta}$, as long as (p, n, q) increase under condition (C1), $\hat{\beta}(\lambda)$ is ℓ_2 consistent to $\tilde{\beta}$ with high probability, which is the first contribution of this paper. To prove Theorem 2.4, we first give a deterministic result in Lemma 2.5. For the sake of briefness, we use Y to shortly denote the centered class labels (defined below) y_i for $i = 1, 2\cdots, n$.

Note that the Lasso estimate $\hat{\beta}(\lambda)$ defined in Equation (2.3) is equivalent to the following Lasso problem without an intercept.

$$\hat{\beta}(\lambda) = \arg\max_{\beta} \frac{1}{2n}\sum_{i=1}^{n}\left[\tilde{y}_i - x_i^{\mathrm{T}}\beta\right]^2 + \lambda\|\beta\|_1, \tag{2.5}$$

where $\tilde{y}_i = y_i - \frac{1}{n}\sum_{i=1}^n y_i$ and $x_i = z_i - \frac{1}{n}\sum_{i=1}^n z_i$ are the centered class labels and the centered observations respectively.

Lemma 2.5. *Suppose that X satisfies restricted eigenvalue condition RE($\gamma, 3$). For any $\beta \in \mathbb{R}^p$ with $\beta_{S^c} = 0$ and λ satisfying the following relationship*

$$\left\| \frac{1}{n} X^{\mathrm{T}}(Y - X\beta) \right\|_\infty \leq \frac{1}{2}\lambda, \tag{2.6}$$

for any Lasso estimator $\hat{\beta}(\lambda)$, we have

1. $\|\hat{\beta}(\lambda) - \beta\|_2 \leq \frac{3\lambda\sqrt{q}}{\gamma}$,
2. $\|\hat{\beta}(\lambda) - \beta\|_1 \leq \frac{12\lambda q}{\gamma}$,
3. $\frac{1}{n}\|X(\hat{\beta}(\lambda) - \beta)\|_2^2 \leq \frac{9\lambda^2 q}{\gamma}$.

This result is a deterministic result – no matter what relationship between Y and $X\beta$ is, as long as inequality (2.6) and restricted eigenvalue condition holds. We could use this lemma to bound the distance between the Lasso estimator $\hat{\beta}(\lambda)$ and $\tilde{\beta}$. For simple linear regression, since $Y - X\tilde{\beta}$ (here $\tilde{\beta}$ is the true parameter) is the noise vector and usually one assume it is (sub)Gaussian distributed, it is very straightforward from Lemma 2.5 that one could choose a suitable λ, such that inequality (2.6) holds and so do results (1), (2) and (3) with high probability. But for classification problems, Y is binary coded and clearly $Y - X\tilde{\beta}$ is not a i.i.d noise vector. We do not have a straightforward choice of λ satisfing inequality (2.6) with high probability. Fortunately, with upper bounds of $\left(\mu^{(1)} - \mu^{(2)}\right)^{\mathrm{T}} \tilde{\beta}$ and $\tilde{\beta}^{\mathrm{T}}\Sigma\tilde{\beta}$ given in Appendix, we could find a suitable λ such that inequality (2.6) holds with high probability.

Theorem 2.6. *Under conditions (C1) and (C2), with probability greater than $1 - O(p^{-1})$, we have*

$$\left\| \frac{1}{n} X^{\mathrm{T}}(y - X\tilde{\beta}) \right\|_\infty \leq C\sqrt{\frac{\log p}{n}},$$

where C is a positive constant depending on $\left\{\theta_n, \Delta_p^2, \lambda_{\max}(\Sigma)\right\}$.

The right hand given in Theorem 2.6 is on the same order of λ used in linear regression. And the proof of Theorem 2.6 is the most difficult part in the paper. Once we have Theorem 2.6, together with Lemma 2.5, we immediately have the consistent results in Theorem 2.4. We postpone all the proofs of Theorem 2.4, Lemma 2.5 and Theorem 2.6 in appendix.

Theorem 2.4 gives ℓ_1 and ℓ_2 consistency of the Lassoed estimator using the penalized least square LDA, which together with the variable selection property established in [10], ensure that $\hat{\beta}(\lambda)$ is a good estimator of the fisher classification direction β^*. The next subsection shows that the misclassification error rate of this sparse LDA method tends to be aymptotically optimal, which once more confirms us to use the Lassoed LDA in practice.

2.3. Asymptotic optimal

Being linear with respect to x, $\mathrm{U}(x)$ and $\mathrm{W}(x)$ are both normal distributed no matther which class x belongs to. Recall that $\mathrm{W}(x) = \left(x - \frac{\hat{\mu}^{(1)}+\hat{\mu}^{(2)}}{2}\right)^{\mathrm{T}} \hat{\beta}(\lambda)$ is the estimated discriminant function and $\mathrm{U}(x) = (x-\mu)^{\mathrm{T}} \beta^*$ is the optimal (true but unknown) discriminant function. Through simple calculations, we have the misclassification error R of $\mathrm{U}(x)$ and the misclassification probability R_n of $\mathrm{W}(x)$ conditional on the training samples,

$$R := P(1|2) = P(2|1) = 1 - \Phi\left(\frac{1}{2}\Delta_p\right), \tag{2.7}$$

$$R_n := \frac{1}{2}\left(\hat{P}(2|1)+\hat{P}(1|2)\right) = 1 - \frac{1}{2}\Phi\left(\frac{\left(\mu^{(1)}-\hat{\mu}\right)^{\mathrm{T}}\hat{\beta}}{\sqrt{\hat{\beta}^{\mathrm{T}}\Sigma\hat{\beta}}}\right) - \frac{1}{2}\Phi\left(-\frac{\left(\mu^{(2)}-\hat{\mu}\right)^{\mathrm{T}}\hat{\beta}}{\sqrt{\hat{\beta}^{\mathrm{T}}\Sigma\hat{\beta}}}\right),$$

where $\hat{\beta}$ could be either $\hat{\beta}^{lda}(\hat{\beta}^{ols})$ or $\hat{\beta}(\lambda)$. It is well known that in the classic settings when p is fixed, the fisher LDA asymptotically attains the optimal misclassification rate defined in (2.7). Now for the high-dimensional extension, one natural question is do we still have the optimality result? We answer this question in this subsection under conditions (C1), (C2) and (C3) together with a new condition (C4).

(C4) $\lambda_{\min}(\Sigma)$ is bounded away from zero.

The results of asymptotic optimal misclassification rate is very similar to the results in [6], where the author proposed a method like Danzig selector for sparse LDA problem. Due to the similarity between Lasso and Danzig selector estimator, our results and proof techniques are very similar – both try to bound functions of a few normal statistics with high probabilities. We borrowed a few techniques from [6].

Theorem 2.7. *Under the same conditions in Theorem 2.4, together with condition (C4) and*

$$\Delta_p^3 \|\beta^*\|_0^{1/2} = o\left(\sqrt{\frac{n}{\log p}}\right), \tag{2.8}$$

we have

$$\frac{R_n(\hat{\beta}(\lambda))}{R} - 1 = O\left(\Delta_p^3\sqrt{\frac{q\log p}{n}}\right),$$

with probability greater than $1 - O(p^{-1})$.

[6] provided a similar result for their Danzig selector type sparse LDA, which could be displayed as

$$\frac{R_n}{R} - 1 = O\left(\Delta_p^2 q\sqrt{\frac{\log p}{n}}\right),$$

by using notations defined in the present paper. These two rates are sligtly different, which could be expected because of the similarity of Danzig selector and the Lasso(see [3]). Note that $\Delta_p^2 = \delta^{\mathrm{T}}\Sigma^{-1}\delta$ and $\beta^* = \Sigma^{-1}\delta$, as long as conditions (C2) and (C4) hold, we have $\Delta_p = O(\|\beta^*\|_2)$. When β_S^* is bounded below and above, $\|\beta^*\|_2 = O(\sqrt{q})$, leading to the fact that $\Delta_p = O(\sqrt{q})$. In this case, the asymptotic rates of Danzig selector type sparse LDA and the Lassoed LDA are exactly the same. The result given in Theorem 2.7 is called asymptotically optimal, whose definition could be found in [12]. If we relax the condition (2.8), we will prove that this sparse LDA method is asymptotically sub-optimal, see [12] too.

Theorem 2.8. *Under the same conditions in Theorem 2.7, except replacing (2.8) with*

$$\Delta_p \|\beta^*\|_0^{1/2} = o\left(\sqrt{\frac{n}{\log p}}\right), \tag{2.9}$$

we have

$$R_n(\hat{\beta}(\lambda)) - R \to 0,$$

with probability greater than $1 - O(p^{-1})$.

The proof of Theorem 2.7 and 2.8, which are both adapted from [6], are postponed in appendix. Same as the analysis after Theorem 2.7, if $\Delta_p = O(\sqrt{q})$, the condition (2.9) can be rewritten as

$$q = o\left(\sqrt{\frac{n}{\log p}}\right),$$

which is contained in condition (C1). So in this case, we can derive the asymptotically sub-optimal property of the penalized least square LDA without any further condition.

3. Conclusions

Efficient high-dimensional discriminant analysis is very demanding in today's real applications. Fisher's LDA is a fundamental method. The extension of LDA to high-dimension is crucial. We studied the asymptotic properties of the Lassoed LDA estimator. The large sample results convince people to use this simple method to solve LDA problem. [10] gives a variable selection consistent result under irrepresentable condition and we now provide an ℓ_2 consistent result under restricted eigenvalue condition. These results look similar to the linear regression results. But the proofs here are much more complicated because the response (class label) and the separator (hyperplane) do not have a straightforward stochastic relationship as the linear regression model.

4. Appendix

This appendix contains technical proofs for our results. We first give a proof of Theorem 2.4 by using the result in Lemma 2.5 and results in Theorem 2.6. Then we prove Lemma 2.5 and Theorem 2.6. Finally, we prove Theorem 2.7 and 2.8.

4.1. Proof of Theorem 2.4

Proof. When λ is chosen to be $c_\lambda\sqrt{\log p/n}$ for the same positive constant c_λ as in Theorem 2.6, from Theorem 2.6 we immediately have

$$\left\|\frac{1}{n}X^{\mathrm{T}}(y-X\tilde{\beta})\right\|_\infty \le \frac{1}{2}\lambda,$$

with probability greater than $1-O(p^{-1})$. Then by Lemma 2.5, we derive the upper bounds for the ℓ_1 and ℓ_2 norm of the distance between $\hat{\beta}(\lambda)$ and $\tilde{\beta}$

$$\left\|\hat{\beta}(\lambda)-\tilde{\beta}\right\|_2 \le \frac{3\lambda\sqrt{q}}{\gamma} = \frac{3c_\lambda}{\gamma}\sqrt{\frac{q\log p}{n}}, \tag{4.1}$$

$$\left\|\hat{\beta}(\lambda)-\tilde{\beta}\right\|_1 \le \frac{12\lambda q}{\gamma} = \frac{12c_\lambda q}{\gamma}\sqrt{\frac{\log p}{n}}. \tag{4.2}$$

From inequalities (4.1), (4.2) together with condition (C1), we immediately have the consistency of $\hat{\beta}(\lambda)$

$$P\left(\|\hat{\beta}(\lambda)-\tilde{\beta}\|_2 \to 0\right) \to 1,$$

$$P\left(\|\hat{\beta}(\lambda)-\tilde{\beta}\|_1 \to 0\right) \to 1. \qquad \square$$

4.2. Proof of Lemma 2.5

Proof. Without any confusion, we use $\hat{\beta}$ to shortly denote the Lassoed estimator $\hat{\beta}(\lambda)$. Note that $\hat{\beta}$ minimizes $\frac{1}{2n}\|Y-X\beta\|^2+\lambda\|\beta\|_1$, for any $\beta\in\mathbb{R}^p$ with $\beta_{S^c}=0$ we have

$$\frac{1}{2n}\|Y-X\hat{\beta}\|^2+\lambda\|\hat{\beta}\|_1 \le \frac{1}{2n}\|Y-X\beta\|^2+\lambda\|\beta\|_1.$$

By arranging the terms and replacing $\hat{\beta}$ with $\beta+\nu$, we have

$$\begin{aligned} 0 &\ge \left[\frac{1}{2n}\|Y-X\beta-X\nu\|^2-\frac{1}{2n}\|Y-X\beta\|^2\right]+[\lambda\|\beta+\nu\|_1-\lambda\|\beta\|_1] \\ &= \frac{1}{2n}\|X\nu\|_2^2-\langle\nu,X^{\mathrm{T}}(Y-X\beta)/n\rangle+\lambda(\|\beta_S+\nu_S\|_1+\|\nu_{S^c}\|_1-\|\beta_S\|_1) \\ &\ge \frac{1}{2n}\|X\nu\|_2^2-(\|\nu_S\|_1+\|\nu_{S^c}\|_1)\|X^{\mathrm{T}}(Y-X\beta)/n\|_\infty \end{aligned}$$

$$
\begin{aligned}
&+\lambda(\|\beta_S+\nu_S\|_1+\|\nu_{S^c}\|_1-\|\beta_S\|_1)\\
\geq\ & \frac{1}{2n}\|X\nu\|_2^2-(\|\nu_S\|_1+\|\nu_{S^c}\|_1)\frac{\lambda}{2}+\lambda(\|\nu_{S^c}\|_1-\|\nu_S\|_1)\\
=\ & \frac{1}{2n}\|X\nu\|_2^2+\frac{\lambda}{2}(\|\nu_{S^c}\|_1-3\|\nu_S\|_1),
\end{aligned}
$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product of two vectors. From the above inequality we have

$$
\frac{1}{2n}\|X\nu\|_2^2\leq\frac{\lambda}{2}(3\|\nu_S\|_1-\|\nu_{S^c}\|_1)\quad\text{and}\quad\|\nu_{S^c}\|_1-3\|\nu_S\|_1\leq 0.
$$

Together with $RE(\gamma,3)$, we immediately have that

$$
\gamma\|\nu\|_2^2\leq\frac{1}{n}\|X\nu\|_2^2\leq\lambda(3\|\nu_S\|_1-\|\nu_{S^c}\|_1)\leq 3\lambda\|\nu_S\|_1\leq 3\sqrt{q}\lambda\|\nu_S\|_2\leq 3\lambda\sqrt{q}\|\nu\|_2.
$$

Consequently,

$$
\|\nu\|_2\leq\frac{3\lambda\sqrt{q}}{\gamma}\quad\text{and}\quad\frac{1}{n}\|X\nu\|_2^2\leq\frac{9\lambda^2 q}{\gamma},
$$

from which, we finally get

$$
\|\nu\|_1=\|\nu_S\|_1+\|\nu_{S^c}\|_1\leq 4\|\nu_S\|_1\leq 4\sqrt{q}\|\nu_S\|_2\leq 4\sqrt{q}\|\nu\|_2=\frac{12\lambda q}{\gamma}.\qquad\square
$$

4.3. Proof of Theorem 2.6

4.3.1. Three useful lemmas

Before giving the proof of Theorem 2.6, we first introduce three technical lemmas which are the main tools we used in this subsection. First the definition of sub-exponential random variable and its corresponding concentration inequality are given.

Definition 4.1 (Sub-exponential)**.** A random variable X is sub-exponential with parameters (σ,b), if for all $|\lambda|<\frac{1}{b}$,

$$
\mathrm{E}\left[\exp[\lambda(X-\mathrm{E}(X))]\right]\leq\exp\left(\frac{1}{2}\sigma^2\lambda^2\right).
$$

Lemma 4.2. *Suppose that (X_1,X_2) follows joint normal distribution with mean $(0,0)$, variances σ_1^2,σ_2^2 and correlation ρ. Let σ^2 be any positive real number such that $\sigma^2\geq\sigma_1\sigma_2$. Then $\frac{X_1X_2}{\sigma^2}-E\left[\frac{X_1X_2}{\sigma^2}\right]$ is a mean 0 sub-exponential random variable with parameter $(\sqrt{8},4)$.*

Proof. By definition, we need to prove for all $|\lambda|<\frac{1}{4}$,

$$
\mathrm{E}\left[\exp\left[\lambda\left(\frac{X_1X_2}{\sigma_1\sigma_2}-E\left[\frac{X_1X_2}{\sigma_1\sigma_2}\right]\right)\right]\right]\leq\exp\left(4\lambda^2\right).\tag{4.3}
$$

Since (X_1, X_2) follows a joint normal distribution, X_2 can be written as

$$X_2 = \rho\frac{\sigma_2}{\sigma_1}X_1 + \epsilon,$$

where ϵ has a normal distribution with mean 0 and variance $(1-\rho^2)\sigma_2^2$ and is independent of X_1. So

$$\begin{aligned}
&E\left[\exp\left[\lambda\left(\frac{X_1X_2}{\sigma_1\sigma_2} - E\left[\frac{X_1X_2}{\sigma_1\sigma_2}\right]\right)\right]\right]\\
=&E\left[\exp\left[\lambda\left(\frac{X_1(\rho\frac{\sigma_2}{\sigma_1}X_1+\epsilon)}{\sigma_1\sigma_2}\right)\right]\right]e^{-\lambda\rho}\\
=&E\left[\exp\left(\lambda\rho\frac{X_1^2}{\sigma_1^2}\right)E\left[\exp\left[\left(\frac{\lambda X_1\epsilon}{\sigma_1\sigma_2}\right)\right] \mid X_1\right]\right]e^{-\lambda\rho}\\
=&E\left[\exp\left(\lambda\rho\frac{X_1^2}{\sigma_1^2}\right)\exp\left(\frac{\lambda^2(1-\rho^2)X_1^2}{2\sigma_1^2}\right)\right]e^{-\lambda\rho}\\
=&E\left[\exp\left(\left[\lambda\rho+\frac{1}{2}\lambda^2(1-\rho^2)\right]\frac{X_1^2}{\sigma_1^2}\right)\right]e^{-\lambda\rho}\\
=&\frac{1}{\sqrt{1-2\lambda\rho-\lambda^2(1-\rho^2)}}e^{-\lambda\rho}. \qquad (4.4)
\end{aligned}$$

Note that when $|\lambda| \le \frac{1}{4}$,

$$\lambda\rho+\frac{1}{2}\lambda^2(1-\rho^2) \le \frac{1}{2}[2\lambda\rho+\lambda(1-\rho^2)] = \frac{1}{2}\lambda[2-(\rho-1)^2] \le \lambda \le \frac{1}{4} < \frac{1}{2}.$$

So (4.4) holds from the moment generating function of Chi-square distribution, and from (4.4), to prove inequality (4.3), it is sufficient for us to verify

$$-\lambda\rho - \frac{1}{2}\log(1-2\lambda\rho-\lambda^2(1-\rho^2)) \le 4\lambda^2.$$

Note that

$$\begin{aligned}
&\log(1+x) \ge x - \frac{x^2}{2}, \text{ for all } x \ge 0,\\
&\log(1+x) \ge x - x^2, \text{ for all } -\frac{1}{2} < x < 0.
\end{aligned}$$

When $-2\lambda\rho-\lambda^2(1-\rho^2) \ge 0$, we have

$$\begin{aligned}
&-\lambda\rho - \frac{1}{2}\log(1-2\lambda\rho-\lambda^2(1-\rho^2)) \le -\lambda\rho - \frac{1}{2}\left[-2\lambda\rho-\lambda^2(1-\rho^2)\right.\\
&\left.-(-2\lambda\rho-\lambda^2(1-\rho^2))^2/2\right] = \frac{1}{2}\lambda^2\left[(1-\rho^2)+\frac{(2\rho+\lambda(1-\rho^2))^2}{2}\right]\\
&\le \frac{1}{2}\lambda^2\left[1+\frac{(2+1/4)^2}{2}\right] \le 2\lambda^2.
\end{aligned}$$

When $-\frac{1}{2} < -2\lambda\rho - \lambda^2(1-\rho^2) \leq 0$, we have

$$\begin{aligned} &-\lambda\rho - \frac{1}{2}\log(1-2\lambda\rho-\lambda^2(1-\rho^2)) \leq -\lambda\rho - \frac{1}{2}\left[-2\lambda\rho-\lambda^2(1-\rho^2)\right. \\ &\left.-(-2\lambda\rho-\lambda^2(1-\rho^2))^2\right]\frac{1}{2}\lambda^2\left[(1-\rho^2)+(2\rho+\lambda(1-\rho^2))^2\right] \\ &\leq \frac{1}{2}\lambda^2[1+(2+1/4)^2] \leq 4\lambda^2. \end{aligned}$$

□

Lemma 4.3 (Gaussian Concentration Inequality)**.** *Let X be a vector of n independent standard normal random variable. Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-Lipschitz function. Then, for all $t > 0$,*

$$P\left(f(X) - Ef(X) \geq t\right) \leq e^{-t^2/(2L^2)}.$$

Lemma 4.4 (Concentration Inequality for Sub-exponential)**.** *For sub-exponential random variable X with parameter (σ, b),*

$$P\left(X - EX \geq t\right) \leq \begin{cases} \exp(-\frac{t^2}{2\sigma^2}) & \text{if } 0 \leq t \leq \frac{\sigma^2}{b} \\ \exp(-\frac{t}{2b}) & \text{if } t > \frac{\sigma^2}{b} \end{cases}$$

The proof of Lemma 4.3 and Lemma 4.4 can be found in [4]. Here we omit them for briefness. Now we begin to prove the main result of this section.

4.3.2. Proof of Theorem 2.6

Proof. Recall the definition of $\tilde{\beta}$,

$$\tilde{\beta} = \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)\left[2\theta_n(1-\theta_n) - \theta_n(1-\theta_n)\left(\mu^{(1)} - \mu^{(2)}\right)^{\mathrm{T}}\tilde{\beta}\right].$$

Multiplying on the left by $\left(\mu^{(1)} - \mu^{(2)}\right)^{\mathrm{T}}$ and $\tilde{\beta}^{\mathrm{T}}\Sigma$ on both sides, and rearranging the terms, we have

$$\left(\mu^{(1)} - \mu^{(2)}\right)^{\mathrm{T}}\tilde{\beta} = \frac{2\omega_n}{1+\omega_n} \quad \text{and} \quad \tilde{\beta}^{\mathrm{T}}\Sigma\tilde{\beta} = \left(\frac{2\omega_n}{1+\omega_n}\right)^2\frac{1}{\Delta_p^2},$$

where $\omega_n = \theta_n(1-\theta_n)\Delta_p^2$, which is bounded below from condition (C2). As a consequence, $\left(\mu^{(1)} - \mu^{(2)}\right)^{\mathrm{T}}\tilde{\beta}$ falls in (c,2) for some $c > 0$ and $\tilde{\beta}^{\mathrm{T}}\Sigma\tilde{\beta} = O\left(1/\Delta_p^2\right)$. Then pluging $\tilde{\beta}$ into $\frac{1}{n}X^{\mathrm{T}}(Y - X\tilde{\beta})$, we have

$$\frac{1}{n}X^{\mathrm{T}}(Y - X\tilde{\beta}) = 2\theta_n(1-\theta_n)\xi_1 - \theta_n(1-\theta_n)\xi_2 - \xi_3,$$

where

$$\xi_1 \quad = \quad \left(\hat{\mu}^{(1)} - \hat{\mu}^{(2)}\right) - \left(\mu^{(1)} - \mu^{(2)}\right),$$

$$\begin{aligned}\xi_2 &= \left[\left(\hat{\mu}^{(1)}-\hat{\mu}^{(2)}\right)\left(\hat{\mu}^{(1)}-\hat{\mu}^{(2)}\right)^{\mathrm{T}}-\left(\mu^{(1)}-\mu^{(2)}\right)\left(\mu^{(1)}-\mu^{(2)}\right)^{\mathrm{T}}\right]\tilde{\beta},\\ \xi_3 &= \left(\hat{\Sigma}-\Sigma\right)\tilde{\beta}.\end{aligned}$$

We will prove the theorem in three steps according to the decomposition, and draw the conclusion in step four.

Step one: From the normal distribution of $\hat{\mu}^{(1)}$ and $\hat{\mu}^{(2)}$, and by the independence of two sample pairs, we have

$$\xi_1=\left(\hat{\mu}^{(1)}-\hat{\mu}^{(2)}\right)-\left(\mu^{(1)}-\mu^{(2)}\right)\sim N\left(0,\frac{n}{n_1n_2}\Sigma\right).$$

Then applying Gaussian concentration inequality(Lemma 4.3), we have

$$P\left(\|\xi_1\|_\infty\geq c_1\sqrt{\frac{\log p}{n}}\right)\leq\frac{1}{p},\tag{4.5}$$

where $c_1=\sqrt{\frac{4\lambda_{\max}(\Sigma)}{\theta_n(1-\theta_n)}}$.

Step two: Decompose ξ_2 first,

$$\begin{aligned}\xi_2 &= \left[\left(\hat{\mu}^{(1)}-\hat{\mu}^{(2)}\right)\left(\hat{\mu}^{(1)}-\hat{\mu}^{(2)}\right)^{\mathrm{T}}-\left(\mu^{(1)}-\mu^{(2)}\right)\left(\mu^{(1)}-\mu^{(2)}\right)^{\mathrm{T}}\right]\tilde{\beta}\\ &= \left[\left(\hat{\mu}^{(1)}-\hat{\mu}^{(2)}\right)-\left(\mu^{(1)}-\mu^{(2)}\right)\right]\left[\left(\hat{\mu}^{(1)}-\hat{\mu}^{(2)}\right)-\left(\mu^{(1)}-\mu^{(2)}\right)\right]^{\mathrm{T}}\tilde{\beta}\\ &\quad+\left[\left(\hat{\mu}^{(1)}-\hat{\mu}^{(2)}\right)-\left(\mu^{(1)}-\mu^{(2)}\right)\right]\left(\mu^{(1)}-\mu^{(2)}\right)^{\mathrm{T}}\tilde{\beta}\\ &\quad+\left(\mu^{(1)}-\mu^{(2)}\right)\left[\left(\hat{\mu}^{(1)}-\hat{\mu}^{(2)}\right)-(\mu^{(1)}-\mu^{(2)})\right]^{\mathrm{T}}\tilde{\beta}\\ &= \xi_1\xi_1^{\mathrm{T}}\tilde{\beta}+\left(\mu^{(1)}-\mu^{(2)}\right)\xi_1^{\mathrm{T}}\tilde{\beta}+\xi_1\left(\mu^{(1)}-\mu^{(2)}\right)^{\mathrm{T}}\tilde{\beta}.\end{aligned}$$

We will prove $\|\xi_2\|_\infty$ tends to zero with high probability by analysing the infinity norm of all the three parts tend to zero separately.

Part 1: Note $\|\xi_1\xi_1^{\mathrm{T}}\tilde{\beta}\|_\infty=\|\xi_1\|_\infty|\xi_1^{\mathrm{T}}\tilde{\beta}|$ and $\xi_1^{\mathrm{T}}\tilde{\beta}$ is a Gaussian random variable with mean zero and variance $\frac{n}{n_1n_2}\tilde{\beta}^{\mathrm{T}}\Sigma\tilde{\beta}$, which is bounded above by $\frac{4}{\theta_n(1-\theta_n)\Delta_p^2n}$. Then (4.5) and Gaussian concentration inequality(Lemma 4.3) yields the following inequality

$$P\left(\|\xi_1\xi_1^{\mathrm{T}}\tilde{\beta}\|_\infty>c_1c_2\sqrt{\frac{\log p}{n}}\right)\leq\frac{2}{p},\tag{4.6}$$

where $c_2=\sqrt{\frac{8}{\theta_n(1-\theta_n)\Delta_p^2}}$.

Part 2: For each $j\in\{1,2,\cdots,p\}$, the jth element of $\left(\mu^{(1)}-\mu^{(2)}\right)\xi_1^{\mathrm{T}}\tilde{\beta}$ is a Gaussian random variable with mean zero and variance $\frac{n}{n_1n_2}\left(\mu_j^{(1)}-\mu_j^{(2)}\right)^2\tilde{\beta}^{\mathrm{T}}\Sigma\tilde{\beta}$,

which is bounded above by

$$\frac{4\|\mu^{(1)}-\mu^{(2)}\|_2^2}{\theta_n(1-\theta_n)n\Delta_p^2} \le \frac{4\lambda_{\max}(\Sigma)}{\theta_n(1-\theta_n)n}.$$

So applying Gaussian concentration inequality, we immediately have

$$P\left(\left\|\left(\mu^{(1)}-\mu^{(2)}\right)\xi_1^{\mathrm{T}}\tilde{\beta}\right\|_\infty \ge c_3\sqrt{\frac{\log p}{n}}\right) \le \frac{1}{p}, \tag{4.7}$$

where $c_3 = \sqrt{\frac{16\lambda_{\max}(\Sigma)}{\theta_n(1-\theta_n)}}$.

Part 3: From the analysis before step 1, we have

$$\left\|\xi_1(\mu^{(1)}-\mu^{(2)})^{\mathrm{T}}\tilde{\beta}\right\|_\infty = \|\xi_1\|_\infty\left|\left(\mu^{(1)}-\mu^{(2)}\right)^{\mathrm{T}}\tilde{\beta}\right| \le 2\|\xi_1\|_\infty.$$

Then from the result given in step one, we immediately have

$$P\left(\left\|\xi_1\left(\mu^{(1)}-\mu^{(2)}\right)^{\mathrm{T}}\tilde{\beta}\right\|_\infty > 2c_1\sqrt{\frac{\log p}{n}}\right) \le P\left(\|\xi_1\|_\infty > c_1\sqrt{\frac{\log p}{n}}\right) \le \frac{1}{p}, \tag{4.8}$$

which together with inequalities (4.6) and (4.7), implies that

$$P\left(\|\xi_2\|_\infty > (c_1c_2+c_3+2c_1)\sqrt{\frac{\log p}{n}}\right) \le \frac{4}{p}.$$

Step three: Let $x_i^{(1)} = u_i + \mu^{(1)}, i = 1, 2, \cdots, n_1$ and $x_k^{(2)} = u_{n_1+k} + \mu^{(2)}, k = 1, 2, \cdots, n_2$, where $\{u_i\}$ for $i = 1, 2, \cdots, n$ are independent and identical Gaussian random variables with mean zero and covariance matrix Σ. And $\hat{\Sigma}$ can be further written as

$$\begin{aligned}\hat{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}u_iu_i^{\mathrm{T}} - \frac{n_1}{n}\left(\hat{\mu}^{(1)}-\mu^{(1)}\right)\left(\hat{\mu}^{(1)}-\mu^{(1)}\right)^{\mathrm{T}} \\ &\quad -\frac{n_2}{n}\left(\hat{\mu}^{(2)}-\mu^{(2)}\right)\left(\hat{\mu}^{(1)}-\mu^{(2)}\right)^{\mathrm{T}}.\end{aligned}$$

Consequently,

$$\begin{aligned}\left(\hat{\Sigma}-\Sigma\right)\tilde{\beta} &= \frac{1}{n}\sum_{i=1}^{n}\left(u_iu_i^{\mathrm{T}}\tilde{\beta}-\mathrm{E}u_iu_i^{\mathrm{T}}\tilde{\beta}\right) - \frac{n_1}{n}\left(\hat{\mu}^{(1)}-\mu^{(1)}\right)\left(\hat{\mu}^{(1)}-\mu^{(1)}\right)^{\mathrm{T}}\tilde{\beta} \\ &\quad -\frac{n_2}{n}\left(\hat{\mu}^{(2)}-\mu^{(2)}\right)\left(\hat{\mu}^{(1)}-\mu^{(2)}\right)^{\mathrm{T}}\tilde{\beta}.\end{aligned}$$

Firstly we analyze the first part in the decomposition above. Since for each $j \in \{1, 2, \cdots, p\}$, u_{ij} and $u_i^{\mathrm{T}}\tilde{\beta}$ are Gaussian distributed with mean zero and variance Σ_{ij}, $\tilde{\beta}^{\mathrm{T}}\Sigma\tilde{\beta}$ respectively. The product of the two variances can be bounded

by $c_4 = \frac{4\lambda_{\max}(\Sigma)}{\Delta_p^2}$ and hence from Lemma 4.2, we have for each $j \in \{1, 2, \cdots, p\}$, $\frac{u_{ij}u_i^{\mathrm{T}}\tilde{\beta}}{\sqrt{c_4}} - \mathrm{E}\frac{u_{ij}u_i^{\mathrm{T}}\tilde{\beta}}{\sqrt{c_4}}$ is a sub-exponential random variable with parameter $(\sqrt{8}, 4)$. So $\eta_j = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{u_{ij}u_i^{\mathrm{T}}\tilde{\beta}}{\sqrt{c_4}} - E\frac{u_{ij}u_i^{\mathrm{T}}\tilde{\beta}}{\sqrt{c_4}})\right]$ follows sub-exponential distribution with parameter $(\sqrt{\frac{8}{n}}, \frac{4}{n})$. By concentration inequality for sub-exponential distribution given in Lemma 4.4, for all $0 \le t \le 2$, we have

$$P\left(|\eta_j| > t\right) \le 2\exp\left(-\frac{nt^2}{16}\right).$$

Consequently,

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n_1}(u_i u_i^{\mathrm{T}}\tilde{\beta} - \mathrm{E}u_i u_i^{\mathrm{T}}\tilde{\beta})\right\|_\infty > c_5\sqrt{\frac{\log p}{n}}\right) \le \frac{2}{p} \tag{4.9}$$

where $c_5 = \sqrt{32c_4}$.

Similar to the analysis of part one in step two, we can obtain

$$P\left(\left\|\frac{n_1}{n}\left(\hat{\mu}^{(1)} - \mu^{(1)}\right)\left(\hat{\mu}^{(1)} - \mu^{(1)}\right)^{\mathrm{T}}\tilde{\beta}\right\|_\infty) > c_6\sqrt{\frac{\log p}{n}}\right) \le \frac{2}{p},$$

$$P\left(\left\|\frac{n_1}{n}\left(\hat{\mu}^{(2)} - \mu^{(2)}\right)\left(\hat{\mu}^{(2)} - \mu^{(2)}\right)^{\mathrm{T}}\tilde{\beta}\right\|_\infty) > c_6\sqrt{\frac{\log p}{n}}\right) \le \frac{2}{p},$$

where $c_6 = \sqrt{\frac{32\lambda_{\max}(\Sigma)}{\Delta_p^2\theta_n^2(1-\theta_n)^2}}$, and from which, together with inequality (4.9), we immediately have

$$P\left(\|\xi_3\|_\infty > (c_5 + 2c_6)\sqrt{\frac{\log p}{n}}\right) \le \frac{6}{p}$$

Step four: Combining all results we have got in step 1-3, there exists a positive constant c depending on $\left\{\theta_n, \Delta_p^2, \lambda_{\max}(\Sigma)\right\}$ such that

$$P\left(\left\|\frac{1}{n}X^{\mathrm{T}}(Y - X\tilde{\beta})\right\|_\infty > c\sqrt{\frac{\log p}{n}}\right) \le O(p^{-1}). \qquad \square$$

4.4. Proof of Theorem 2.7

Proof. Firstly we derive consistency of the numerator in R_n. The difference of the numerator can be decomposed into three parts as in Equation (4.10),

$$\begin{aligned} & \left(\mu^{(1)} - \hat{\mu}\right)^{\mathrm{T}}\hat{\beta} - \left(\mu^{(1)} - \mu\right)^{\mathrm{T}}\tilde{\beta} \\ = \; & (\mu - \hat{\mu})^{\mathrm{T}}\left(\hat{\beta} - \tilde{\beta}\right) + (\mu - \hat{\mu})^{\mathrm{T}}\tilde{\beta} + \frac{1}{2}\left(\mu^{(1)} - \mu^{(2)}\right)^{\mathrm{T}}\left(\hat{\beta} - \tilde{\beta}\right). \end{aligned} \tag{4.10}$$

Using the same technics as the proof of step one in Theorem 2.6, and together with Theorem 2.4, with probability greater than $1 - O(p^{-1})$, we have

$$\left|(\mu - \hat{\mu})^{\mathrm{T}} \left(\hat{\beta} - \tilde{\beta}\right)\right| \leq O\left(\frac{q \log p}{n}\right) \to 0.$$

Note that $(\hat{\mu} - \mu)^{\mathrm{T}} \tilde{\beta}$ is a Gaussian random variable with mean zero and variance $\frac{n}{4n_1 n_2} \tilde{\beta}^{\mathrm{T}} \Sigma \tilde{\beta}$, which is bounded above, so from the Gaussian concentration inequality, we immediately have

$$\left|(\hat{\mu} - \mu)^{\mathrm{T}} \tilde{\beta}\right| \leq O\left(\sqrt{\frac{\log p}{n}}\right) \to 0, \tag{4.11}$$

with probability greater than $1 - O(p^{-1})$. Conditions (C1) and (C4) yields $\Delta_p^2 = O\left(\left\|\mu^{(1)} - \mu^{(2)}\right\|_2^2\right)$, which together with Theorem 2.4 leads to with probability greater than $1 - O(p^{-1})$,

$$\frac{1}{2}\left|(\mu^{(1)} - \mu^{(2)})^{\mathrm{T}} (\hat{\beta} - \tilde{\beta})\right| \leq O\left(\Delta_p \sqrt{\frac{q \log p}{n}}\right) \to 0.$$

By all inequalities obtained above, we immediately have

$$\left|\left(\mu^{(1)} - \hat{\mu}\right)^{\mathrm{T}} \hat{\beta} - \left(\mu^{(1)} - \mu\right)^{\mathrm{T}} \tilde{\beta}\right| \leq O\left(\Delta_p \sqrt{\frac{q \log p}{n}}\right) \to 0, \tag{4.12}$$

with probability greater than $1 - O(p^{-1})$.

Now we begin to prove consistency of the denominator of R_n. Since

$$\begin{aligned}
\left|\hat{\beta}^{\mathrm{T}} \Sigma \hat{\beta} - \tilde{\beta}^{\mathrm{T}} \Sigma \tilde{\beta}\right| &\leq \left|\left(\hat{\beta} - \tilde{\beta}\right)^{\mathrm{T}} \Sigma \left(\hat{\beta} - \tilde{\beta}\right) + 2\left(\hat{\beta} - \tilde{\beta}\right)^{\mathrm{T}} \Sigma \tilde{\beta}\right| \\
&\leq \lambda_{max}(\Sigma) \left\|\hat{\beta} - \tilde{\beta}\right\|_2^2 + 2\left\|\Sigma^{\frac{1}{2}}(\hat{\beta} - \tilde{\beta})\right\|_2 \left\|\Sigma^{\frac{1}{2}} \tilde{\beta}\right\|_2 \\
&\leq \lambda_{max}(\Sigma) \left\|\hat{\beta} - \tilde{\beta}\right\|_2^2 + 2\sqrt{\lambda_{max}(\Sigma)} \left\|\hat{\beta} - \tilde{\beta}\right\|_2 \left\|\Sigma^{\frac{1}{2}} \tilde{\beta}\right\|_2 \\
&\leq \lambda_{max}(\Sigma) \left\|\hat{\beta} - \tilde{\beta}\right\|_2^2 + \frac{4}{\Delta_p} \sqrt{\lambda_{max}(\Sigma)} \left\|\hat{\beta} - \tilde{\beta}\right\|_2 \\
&= O\left(\sqrt{\frac{q \log p}{\Delta_p^2 n}}\right) \to 0,
\end{aligned} \tag{4.13}$$

with probability greater than $1 - O(p^{-1})$, together with the bounds of $\tilde{\beta}^{\mathrm{T}} \Sigma \tilde{\beta}$, we have

$$\left|\frac{\hat{\beta}^{\mathrm{T}} \Sigma \hat{\beta}}{\tilde{\beta}^{\mathrm{T}} \Sigma \tilde{\beta}} - 1\right| \leq O\left(\Delta_p \sqrt{\frac{q \log p}{n}}\right) \to 0,$$

with probability greater than $1 - O(p^{-1})$. Let $\gamma = \Delta_p \sqrt{\frac{q \log p}{n}}$. Then

$$\begin{aligned}
&\left| \frac{(\mu^{(1)} - \hat{\mu})^{\mathrm{T}} \hat{\beta}}{\sqrt{\hat{\beta}^{\mathrm{T}} \Sigma \hat{\beta}}} - \frac{(\mu^{(1)} - \mu)^{\mathrm{T}} \tilde{\beta}}{\sqrt{\tilde{\beta}^{\mathrm{T}} \Sigma \tilde{\beta}}} \right| \\
&\leq \left| \frac{(\mu^{(1)} - \hat{\mu})^{\mathrm{T}} \hat{\beta} - (\mu^{(1)} - \mu)^{\mathrm{T}} \tilde{\beta}}{\sqrt{\hat{\beta}^{\mathrm{T}} \Sigma \hat{\beta}}} \right| + \left| \frac{(\mu^{(1)} - \mu)^{\mathrm{T}} \tilde{\beta}}{\sqrt{\hat{\beta}^{\mathrm{T}} \Sigma \hat{\beta}}} - \frac{(\mu^{(1)} - \mu)^{\mathrm{T}} \tilde{\beta}}{\sqrt{\tilde{\beta}^{\mathrm{T}} \Sigma \tilde{\beta}}} \right| \\
&\leq \left| \frac{(\mu^{(1)} - \hat{\mu})^{\mathrm{T}} \hat{\beta} - (\mu^{(1)} - \mu)^{\mathrm{T}} \tilde{\beta}}{\sqrt{\tilde{\beta}^{\mathrm{T}} \Sigma \tilde{\beta}}} \right| + 2 \left| \frac{\tilde{\beta}^{\mathrm{T}} \Sigma \tilde{\beta} - \hat{\beta}^{\mathrm{T}} \Sigma \hat{\beta}}{\sqrt{\tilde{\beta}^{\mathrm{T}} \Sigma \tilde{\beta}} \sqrt{\hat{\beta}^{\mathrm{T}} \Sigma \hat{\beta}} \left(\sqrt{\tilde{\beta}^{\mathrm{T}} \Sigma \tilde{\beta}} + \sqrt{\hat{\beta}^{\mathrm{T}} \Sigma \hat{\beta}} \right)} \right| \\
&\leq O(\Delta_p \gamma) \\
&\leq O\left(\Delta_p^2 \sqrt{\frac{q \log p}{n}} \right) =: \gamma_n.
\end{aligned}$$

By replacing $\tilde{\beta}$ with

$$\tilde{\beta} = \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \left[2\theta_n(1 - \theta_n) - \theta_n(1 - \theta_n)(\mu^{(1)} - \mu^{(2)})^{\mathrm{T}} \tilde{\beta} \right],$$

the second term in last inequality is $\frac{1}{2} \sqrt{(\mu^{(1)} - \mu^{(2)})^{\mathrm{T}} \Sigma^{-1} (\mu^{(1)} - \mu^{(2)})} = \frac{1}{2} \Delta_p$.

From the property of Φ given in [6] that for any $x < 0$ and $|\delta| \leq 1$, Φ satisfies

$$\left| \frac{\Phi(x + \delta)}{\Phi(x)} - 1 \right| \leq c_1 |\delta| (|x| + 1) e^{c_2 |x\delta|}$$

for some positive constants c_1, c_2 which do not depend on x and δ, we have

$$R_n = R \times (1 + O(1) \gamma_n \Delta_p \exp(O(1) \Delta_p \gamma_n)).$$

So

$$\frac{R_n}{R} - 1 = O\left(\Delta_p^3 \sqrt{\frac{q \log p}{n}} \right). \qquad \square$$

4.5. *Proof of Theorem 2.8*

Proof. When $\Delta_p^2 > M$ for some $M > 0$, then under condition (2.9) and from inequalities (4.12) and (4.13), we have

$$\begin{aligned}
\left| \frac{(\mu^{(1)} - \hat{\mu})^{\mathrm{T}} \hat{\beta}}{\sqrt{\hat{\beta} \Sigma \hat{\beta}}} \right| &\geq C \left| \frac{(\mu^{(1)} - \hat{\mu})^{\mathrm{T}} \tilde{\beta}}{\sqrt{\hat{\beta} \Sigma \hat{\beta}}} \right| \\
&= C \left| (\mu^{(1)} - \hat{\mu})^{\mathrm{T}} \tilde{\beta} \right| \left(\hat{\beta} \Sigma \hat{\beta} \right)^{-\frac{1}{2}}
\end{aligned}$$

$$
\begin{aligned}
&= C\left|(\mu^{(1)}-\hat{\mu})^{\mathrm{T}}\tilde{\beta}\right|\left(\tilde{\beta}\Sigma\tilde{\beta}+o(1)\right)^{-\frac{1}{2}} \\
&= C\left(\left(\frac{(\mu^{(1)}-\hat{\mu})^{\mathrm{T}}\tilde{\beta}}{\sqrt{\tilde{\beta}\Sigma\tilde{\beta}}}\right)^{-2}+o(1)\right)^{-\frac{1}{2}} \\
&= C(4\Delta_p^{-2}+o(1))^{-\frac{1}{2}} \\
&\geq CM^{\frac{1}{2}},
\end{aligned}
$$

which yields $|R_n - R| \leq \exp(-CM)$.

When $\Delta_p^2 \leq M$, from Theorem 2.7, we have

$$R_n = R \times (1 + O(1)\gamma_n\Delta_p \exp(O(1)\Delta_p\gamma_n)), \tag{4.14}$$

which together with $\gamma_n\Delta_p = o(1)$ implies $R_n = R(1+0(1))$. We complete the proof by firstly letting $p, n \to +\infty$ and then $M \to \infty$. □

References

[1] T. W. Anderson. *An introduction to multivariate statistical analysis*. John Wiley & Sons, New York, London, 1958. MR0091588

[2] Peter J. Bickel and Elizaveta Levina. Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004. ISSN 1350-7265. URL http://dx.doi.org/10.3150/bj/1106314847. MR2108040

[3] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. ISSN 0090-5364. URL http://dx.doi.org/10.1214/08-AOS620. MR2533469

[4] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. ISBN 978-0-19-953525-5. URL http://dx.doi.org/10.1093/acprof:oso/9780199535255.001.0001. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. MR3185193

[5] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. ISBN 978-3-642-20191-2. URL http://dx.doi.org/10.1007/978-3-642-20192-9. Methods, theory and applications. MR2807761

[6] Tony Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.*, 106(496):1566–1577, 2011. ISSN 0162-1459. URL http://dx.doi.org/10.1198/jasa.2011.tm11199. MR2896857

[7] Jianqing Fan and Yingying Fan. High-dimensional classification using features annealed independence rules. *Ann. Statist.*, 36(6):2605–2637,

2008. ISSN 0090-5364. URL http://dx.doi.org/10.1214/07-AOS504. MR2485009

[8] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[9] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. ISBN 978-0-387-84857-0. URL http://dx.doi.org/10.1007/978-0-387-84858-7. Data mining, inference, and prediction. MR2722294

[10] Qing Mai, Hui Zou, and Ming Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42, 2012. ISSN 0006-3444. URL http://dx.doi.org/10.1093/biomet/asr066. MR2899661

[11] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.*, 11:2241–2259, 2010. ISSN 1532-4435. MR2719855

[12] Jun Shao, Yazhen Wang, Xinwei Deng, and Sijian Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.*, 39(2):1241–1265, 2011. ISSN 0090-5364. URL http://dx.doi.org/10.1214/10-AOS870. MR2816353

[13] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996. ISSN 0035-9246. URL http://links.jstor.org/sici?sici=0035-9246(1996)58:1<267:RSASVT>2.0.CO;2-G&origin=MSN. MR1379242

[14] Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009. ISSN 1935-7524. URL http://dx.doi.org/10.1214/09-EJS506. MR2576316

[15] Daniela M. Witten and Robert Tibshirani. Covariance-regularized regression and classification for high dimensional problems. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(3):615–636, 2009. ISSN 1369-7412. URL http://dx.doi.org/10.1111/j.1467-9868.2009.00699.x. MR2749910