



University of Pennsylvania
ScholarlyCommons

Publicly Accessible Penn Dissertations


2019

Minimax Optimality In High-Dimensional Classification, Clustering, And Privacy

Linjun Zhang

University of Pennsylvania, zlj11112222@gmail.com

Follow this and additional works at: <https://repository.upenn.edu/edissertations>

 Part of the [Computer Sciences Commons](#), and the [Statistics and Probability Commons](#)

Recommended Citation

Zhang, Linjun, "Minimax Optimality In High-Dimensional Classification, Clustering, And Privacy" (2019). *Publicly Accessible Penn Dissertations*. 3274.

<https://repository.upenn.edu/edissertations/3274>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/3274>

For more information, please contact repository@pobox.upenn.edu.

Minimax Optimality In High-Dimensional Classification, Clustering, And Privacy

Abstract

The age of “Big Data” features large volume of massive and high-dimensional datasets, leading to fast emergence of different algorithms, as well as new concerns such as privacy and fairness. To compare different algorithms with (without) these new constraints, minimax decision theory provides a principled framework to quantify the optimality of algorithms and investigate the fundamental difficulty of statistical problems. Under the framework of minimax theory, this thesis aims to address the following four problems:

1. The first part of this thesis aims to develop an optimality theory for linear discriminant analysis in the high-dimensional setting. In addition, we consider classification with incomplete data under the missing completely at random (MCR) model.
2. In the second part, we study high-dimensional sparse Quadratic Discriminant Analysis (QDA) and aim to establish the optimal convergence rates.
3. In the third part, we study the optimality of high-dimensional clustering on the unsupervised setting under the Gaussian mixtures model. We propose a EM-based procedure with the optimal rate of convergence for the excess mis-clustering error.
4. In the fourth part, we investigate the minimax optimality under the privacy constraint for mean estimation and linear regression models, under both the classical low-dimensional and modern high-dimensional settings.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Statistics

First Advisor

Tony Cai

Keywords

Classification, Clustering, Differential Privacy, High-dimensional data, Minimax Optimal, Non-convex Optimization

Subject Categories

Computer Sciences | Statistics and Probability

MINIMAX OPTIMALITY IN HIGH-DIMENSIONAL CLASSIFICATION,
CLUSTERING, AND PRIVACY

Linjun Zhang

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

T. Tony Cai
Daniel H. Silberberg Professor, Professor of Statistics

Graduate Group Chairperson

Catherine Schrand, Celia Z. Moh Professor, Professor of Accounting

Dissertation Committee:

T. Tony Cai, Daniel H. Silberberg Professor, Professor of Statistics

Edward I. George, Universal Furniture Professor, Professor of Statistics

Hongzhe Li, Professor of Biostatistics in Biostatistics and Epidemiology

MINIMAX OPTIMALITY IN HIGH-DIMENSIONAL CLASSIFICATION,
CLUSTERING, AND PRIVACY

© COPYRIGHT

2019

Linjun Zhang

This work is licensed under the
Creative Commons Attribution
NonCommercial-ShareAlike 3.0
License

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-sa/3.0/>

Dedicated to my beloved parents and family

ACKNOWLEDGEMENT

First and foremost, I would like to thank my amazing advisor, Tony Cai, who guided me through the transition from a good student to an independent researcher. Tony has great insight into math and statistics, and he is a thinker with so many unique ideas in a broad range of areas in statistics. Whenever my research goes to a dead end, Tony is always able to point out a new direction for me. To me, Tony is not only an advisor but a mentor. Through our so many long discussions, Tony taught me how to find research problems, how to write papers and revisions, and helped me blend into the culture in the US. I am sure the wisdom I learned from Tony will have a long term effect on my future career. I am deeply thankful for him dedicating so much time teaching me and also giving me incredible freedom to pursue my own research ideas. I would also like to thank Tony and his wife for organizing Thanksgiving parties every year, making those holidays especially sweet and precious.

Next I would like to thank Hongzhe Li and Edward George for serving my thesis proposal and defense committees. To Hongzhe, thank you for introducing me to the world of microbime data analysis. Reading more your papers and having more discussions with you gradually changed and shaped my view of statistics. I would certainly like to explore more data-driven problems in my future career. To Edward, thank you for all the interesting discussions we have together. You are always so insightful and gentle in our every discussion. The meetings with you always make my day.

This thesis could never been finished without the support of Wharton Statistics Department. I would like to thank Andreas Buja for writing the recommendation letter of teaching for my academia application. I am very grateful to Weijie Su and Edgar Dobriban for frequently discussing problems with me and providing me with a lot of insights. I would also like to thank Mark Low for creating such a friendly and supportive environment in our department. In addition, special thanks go to Larry Brown, who is an unbelievable pioneer and leader

of modern statistics and provided numerous support for our department. I feel so fortunate to be able to take Larry's *Linear Model* course in my first year and had a few discussions about my research with Larry during my PhD career. His wisdom, humor, and kindness will always be remembered.

I am also specially grateful to Anru Zhang, Zijian Guo, Qingyuan Zhao, and Zhuang Ma, who are my senior fellows. They always patiently answer all the questions I have had and share their own wisdom in both research and life with me. I cannot say enough thanks to them. I am also extremely thankful for Jing Ma, Yichen Wang, and Rong Ma, who are my wonderful collaborators, and thank you for all the inspiring discussions we had.

My peers in the department have been a stable source for new ideas and fun. Special shout-out to Justin Khim, who is super smart and friendly, and has been my officemate for the past five years. The daily conversation with Justin can instantly sweep away all my anxiety and stress. I would also like to thank my peers Bikram Karmakar, Raiden Hasegawa, Gemma Moran and Xuran Wang, with whom I shared a lot of fantastic memories. I am sure our friendship will last well beyond graduation. I would also like to thank all my friends at Penn and other graduates in the department, whose companion has made my graduate life entertaining and enjoyable.

Finally, my family have always been there and I cannot thank them enough. No words can ever express my appreciation. I would like to thank my parents for their constant love and encouragement. My father passed away when I was in my third year of PhD. It is always a pain but that is what makes life, life. I strongly appreciate my mother for her braveness and love for me, which is always my pillar of strength. Last but not least, thank you to my fiancé, Jing Xu, for her love and trust. I cannot imagine getting through everything without her support and love.

ABSTRACT

MINIMAX OPTIMALITY IN HIGH-DIMENSIONAL CLASSIFICATION, CLUSTERING, AND PRIVACY

Linjun Zhang

T. Tony Cai

The age of “Big Data” features large volume of massive and high-dimensional datasets, leading to fast emergence of different algorithms, as well as new concerns such as privacy and fairness. To compare different algorithms with (without) these new constraints, minimax decision theory provides a principled framework to quantify the optimality of algorithms and investigate the fundamental difficulty of statistical problems. Under the framework of minimax theory, this thesis aims to address the following four problems:

1. The first part of this thesis aims to develop an optimality theory for linear discriminant analysis in the high-dimensional setting. In addition, we consider classification with incomplete data under the missing completely at random (MCR) model.
2. In the second part, we study high-dimensional sparse Quadratic Discriminant Analysis (QDA) and aim to establish the optimal convergence rates.
3. In the third part, we study the optimality of high-dimensional clustering on the unsupervised setting under the Gaussian mixtures model. We propose a EM-based procedure with the optimal rate of convergence for the excess mis-clustering error.
4. In the fourth part, we investigate the minimax optimality under the privacy constraint for mean estimation and linear regression models, under both the classical low-dimensional and modern high-dimensional settings.

TABLE OF CONTENTS

| | |
|--|------|
| ACKNOWLEDGEMENT | iv |
| ABSTRACT | vi |
| LIST OF TABLES | xi |
| LIST OF ILLUSTRATIONS | xii |
| PREFACE | xiii |
| CHAPTER 1 : Introduction | 1 |
| 1.1 Liner Discriminant Analysis | 1 |
| 1.2 Quadratic Discriminant Analysis | 2 |
| 1.3 Unsupervised Gaussian Mixture Model | 2 |
| 1.4 Parameter Estimation with Differential Privacy | 3 |
| CHAPTER 2 : High-dimensional Linear Discriminant Analysis: Optimal- ity Algorithm, and Missing Data | 5 |
| 2.1 Introduction | 5 |
| 2.2 Methodology | 10 |
| 2.3 Theoretical properties of AdaLDA and ADAM | 18 |
| 2.4 Numerical results | 26 |
| 2.5 Extension to multiple-class LDA | 37 |
| 2.6 Proofs | 39 |
| CHAPTER 3 : A Convex Optimization Approach to High-dimensional Sparse Quadratic Discriminant Analysis | 54 |
| 3.1 Introduction | 54 |

| | | |
|---|--|-----|
| 3.2 | The Difficulties of High-dimensional QDA | 58 |
| 3.3 | Sparse Quadratic Discriminant Analysis | 61 |
| 3.4 | Theoretical Guarantees | 63 |
| 3.5 | Numerical Studies | 67 |
| 3.6 | Extensions | 73 |
| 3.7 | Proofs | 75 |
| | | |
| CHAPTER 4 : CHIME: Clustering of High-Dimensional Gaussian Mixtures with EM Algorithm and Its Optimality | | 94 |
| 4.1 | Introduction | 94 |
| 4.2 | Methodology | 99 |
| 4.3 | Theoretical Analysis | 103 |
| 4.4 | Low-dimensional Gaussian Mixtures | 113 |
| 4.5 | Simulations | 116 |
| 4.6 | Applications to Glioblastoma Gene Expression Data | 120 |
| 4.7 | Extensions to Multi-class Gaussian Mixtures | 122 |
| 4.8 | Proofs | 124 |
| | | |
| CHAPTER 5 : The Cost of Privacy: Optimal Rates of Convergence for Parameter Estimation with Differential Privacy | | 132 |
| 5.1 | Introduction | 132 |
| 5.2 | A General Lower Bound for Minimax Risk with Differential Privacy | 137 |
| 5.3 | Privacy Cost of High-dimensional Mean Estimation | 141 |
| 5.4 | Privacy Cost of Linear Regression | 148 |
| 5.5 | Simulation Studies | 153 |
| 5.6 | Data Analysis | 157 |
| 5.7 | Discussion | 160 |
| 5.8 | Proofs | 161 |
| | | |
| APPENDIX | | 176 |

BIBLIOGRAPHY 176

LIST OF TABLES

| | | |
|------------|---|----|
| TABLE 1 : | Misclassification errors (%) and model fitting times for Model 1 with complete data | 29 |
| TABLE 2 : | Misclassification errors (%) and model fitting times for Model 2 with complete data | 30 |
| TABLE 3 : | Misclassification errors (%) and model fitting times for Model 3 with complete data | 31 |
| TABLE 4 : | Misclassification errors (%) and model fitting times for Model 4 with complete data | 32 |
| TABLE 5 : | Misclassification errors (%) and model fitting times for Model 5 (the first four rows) and 6 (the last four rows) with complete data . . . | 33 |
| TABLE 6 : | Misclassification errors (%) and model fitting times for Model 1 with missing proportion ϵ | 34 |
| TABLE 7 : | Classification error of Lung cancer data by various methods | 35 |
| TABLE 8 : | Classification error of Leukemia data by various methods | 36 |
| TABLE 9 : | Average classification errors (s.e.) based on $n = 100$ test samples from 100 replications under the setting where covariance matrices are known to be identity. | 68 |
| TABLE 10 : | Average classification errors (s.e.) based on $n = 100$ test samples from 100 replications under the setting where means are known to be $\mathbf{0}_p$ and covariance matrices are known to be diagonal. | 69 |
| TABLE 11 : | Average classification errors (s.d.) based on $n = 200$ test samples from 100 replications under three different models | 71 |
| TABLE 12 : | Classification error(%) with s.d. of prostate cancer data by various methods | 72 |

| | |
|--|-----|
| TABLE 13 : Classification error(%) with s.d. of prostate cancer data by various methods | 73 |
| TABLE 14 : Average mis-clustering errors (s.e.) based on $n = 200$ test samples from 100 replications under three different models | 119 |
| TABLE 15 : Clustering results for the GBM gene expression data with $p = 200$ genes and 82 samples | 122 |
| TABLE 16 : Conventional Mean Estimation | 155 |
| TABLE 17 : High-dimensional Mean Estimation | 155 |
| TABLE 18 : Conventional Linear Regression | 155 |
| TABLE 19 : High-dimensional Linear Regression | 156 |

LIST OF ILLUSTRATIONS

FIGURE 1 : Average mis-clustering errors based on $n = 200$ test samples from 100 replications under Model 1 (left), Model 2 (middle) and Model 3 (right). CHIME performs well in all three models. 118

FIGURE 2 : The discriminant vector $|\hat{\beta}|$ is plotted against the marginal variances. 123

FIGURE 3 : Errors (in \log_{10} -scale) plotted against sample size n , with $(0.5, 10/n^{1.1})$ -differentially privacy guarantee. Top-left: conventional mean estimation with sample size ranging from 1000 to 1000×20 ; top-right: conventional linear regression with sample size ranging from 1000 to 1000×20 ; bottom-left: high-dimensional mean estimation with sample size ranging from 200 to 200×20 ; bottom-right: high-dimensional linear regression with sample size ranging from 200 to 200×20 . The local differentially private algorithms (LDP), differentially private (DP) algorithms and non-private (NP) algorithms are colored in green, red and blue respectively. 157

FIGURE 4 : The bootstrap estimate of $\mathbb{E}[\|\hat{\mu} - \mu\|_2]/d$ for the differentially private sparse mean estimator, compared with its locally differentially private counterpart, as sample size increases from 10 to 120. 159

FIGURE 5 : The bootstrap estimate of $\mathbb{E}[\|\hat{\beta} - \beta\|_2]/d$ for the differentially private OLS estimator, compared with its locally differentially private counterpart, as sample size increases from 100 to 20600. 160

CHAPTER 1 : Introduction

The age of “Big Data” features large volume of massive and high-dimensional datasets, leading to fast emergence of different algorithms, as well as new concerns such as privacy and fairness. To compare different algorithms with (without) these new constraints, minimax decision theory provides a principled framework to quantify the optimality of algorithms and investigate the fundamental difficulty of statistical problems.

Under the framework of minimax theory, a matching upper and lower minimax bound would indicate the optimality of the algorithm and the fundamental difficulty of the problem. However, finding the matching bound is not always easy, especially in the high-dimensional setting. Moreover, unlike the high-dimensional linear regression problem, whose optimality has been deeply studied due to the popularity of LASSO, the optimality of high-dimensional classification and clustering, in contrast, is yet to be well understood. Moreover, in modern data analysis, the large-scale data analysis aggravates privacy concerns. The minimax theory with such privacy constraint is a fundamental but unexplored problem in the statistics literature. To address these problems, this thesis consists of the following four parts.

1.1. Liner Discriminant Analysis

Linear discriminant analysis (LDA), or Fisher’s linear discriminant, is a popular method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events (Hastie et al., 2009). The first chapter aims to develop an optimality theory for LDA in the high-dimensional setting. A data-driven and tuning-free classification rule, which is based on an adaptive constrained ℓ_1 minimization approach, is proposed and analyzed. Minimax lower bounds are obtained and this classification rule is shown to be simultaneously rate optimal over a collection of parameter spaces. In addition, we consider classification with incomplete data under the missing completely at random (MCR) model. An adaptive classifier with theoretical guarantees is introduced and optimal rate of convergence for high-dimensional

linear discriminant analysis under the MCR model is established. The technical analysis for the case of missing data is much more challenging than that for the complete data. We establish a large deviation result for the generalized sample covariance matrix, which serves as a key technical tool and can be of independent interest. An application to lung cancer and leukemia studies is also discussed.

This chapter is based on Cai and Zhang (2018c), joint work with T. Tony Cai.

1.2. Quadratic Discriminant Analysis

In this chapter, we extend the results in the previous chapter by considering the Quadratic Discriminant Analysis (QDA) model where the two covariance matrices of two classes are different. We study high-dimensional sparse QDA and aim to establish the optimal convergence rates for the classification error. Minimax lower bounds are established to demonstrate the necessity of structural assumptions such as sparsity conditions on the discriminating direction and differential graph for the possible construction of consistent high-dimensional QDA rules.

We then propose a classification algorithm called SDAR using constrained convex optimization under the sparsity assumptions. Both minimax upper and lower bounds are obtained and this classification rule is shown to be simultaneously rate optimal over a collection of parameter spaces, up to a logarithmic factor. Simulation studies demonstrate that SDAR performs well numerically. The method is also illustrated through an analysis of prostate cancer data and colon tissue data.

This chapter is based on Cai and Zhang (2018b), joint work with T. Tony Cai.

1.3. Unsupervised Gaussian Mixture Model

Unsupervised learning is an important problem in statistics and machine learning with a wide range of applications. In this paper, we study clustering of high-dimensional Gaussian mixtures and propose a procedure, called CHIME, that is based on the EM algorithm and a

direct estimation method for the sparse discriminant vector. Both theoretical and numerical properties of CHIME are investigated. We establish the optimal rate of convergence for the excess mis-clustering error and show that CHIME is minimax rate optimal. In addition, the optimality of the proposed estimator of the discriminant vector is also established. Simulation studies show that CHIME outperforms the existing methods under a variety of settings. The proposed CHIME procedure is also illustrated in an analysis of a glioblastoma gene expression data set and shown to have superior performance.

Clustering of Gaussian mixtures in the conventional low-dimensional setting is also considered. The technical tools developed for the high-dimensional setting are used to establish the optimality of the clustering procedure that is based on the classical EM algorithm.

This chapter is based on Cai et al. (2019a), joint work with T. Tony Cai and Jing Ma.

1.4. Parameter Estimation with Differential Privacy

Privacy-preserving data analysis is a rising challenge in contemporary statistics, as the privacy guarantees of statistical methods are often achieved at the expense of accuracy. In this paper, we investigate the tradeoff between statistical accuracy and privacy in mean estimation and linear regression, under both the classical low-dimensional and modern high-dimensional settings. A primary focus is to establish minimax optimality for statistical estimation with the (ϵ, δ) -differential privacy constraint. To this end, we find that classical lower bound arguments fail to yield sharp results, and new technical tools are called for.

We first develop a general lower bound argument for estimation problems with differential privacy constraints, and then apply the lower bound argument to mean estimation and linear regression. For these statistical problems, we also design computationally efficient algorithms that match the minimax lower bound up to a logarithmic factor. In particular, for the high-dimensional linear regression, a novel private iterative hard thresholding pursuit algorithm is proposed, based on a privately truncated version of stochastic gradient descent. The numerical performance of these algorithms is demonstrated by simulation studies and

applications to real data containing sensitive information, for which privacy-preserving statistical methods are necessary.

This chapter is based on Cai et al. (2019b), joint work with T. Tony Cai and Yichen Wang.

CHAPTER 2 : High-dimensional Linear Discriminant Analysis: Optimality,
Adaptive Algorithm, and Missing Data

2.1. Introduction

Classification is one of the most important tasks in statistics and machine learning with applications in a broad range of fields. See, for example, Hastie et al. (2009). The problem has been well studied in the low-dimensional setting. In particular, consider the Gaussian case where one wishes to classify a new random vector \mathbf{Z} drawn with equal probability from one of two Gaussian distributions $N_p(\boldsymbol{\mu}_1, \Sigma)$ (class 1) and $N_p(\boldsymbol{\mu}_2, \Sigma)$ (class 2). In the ideal setting where all the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$ are known, Fisher's linear discriminant rule, which is given by

$$C_{\boldsymbol{\theta}}(\mathbf{Z}) = \begin{cases} 1, & (\mathbf{Z} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2})^\top \Omega \boldsymbol{\delta} < 0 \\ 2, & (\mathbf{Z} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2})^\top \Omega \boldsymbol{\delta} \geq 0, \end{cases} \quad (2.1)$$

where $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, and $\Omega = \Sigma^{-1}$ is the precision matrix, is well known to be optimal (Anderson, 2003). Fisher's rule separates the two classes by a linear combination of features and its misclassification error is given by $R_{\text{opt}}(\boldsymbol{\theta}) = \Phi(-\frac{1}{2}\Delta)$, where Φ is the cumulative distribution function of the standard normal distribution and $\Delta = \sqrt{\boldsymbol{\delta}^\top \Omega \boldsymbol{\delta}}$ is the signal-to-noise ratio.

Although Fisher's rule can serve as a useful performance benchmark, it is not practical for real data analysis as the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and Σ are typically unknown and need to be estimated from the data. In applications, it is desirable to construct a data-driven classification rule based on two observed random samples, $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n_1}^{(1)} \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_1, \Sigma)$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n_2}^{(2)} \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_2, \Sigma)$. In the conventional low-dimensional setting, this is easily achieved by plugging in Fisher's linear discriminant rule (2.1) the corresponding sample means and pooled sample covariance matrix for the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and Σ respectively. This classification rule is asymptotically optimal when the dimension p is fixed. See, for

example, Anderson (2003).

Driven by many contemporary applications, much recent attention has been on the high-dimensional setting where the dimension is much larger than the sample size. In this case, the sample covariance matrix is not even invertible and it is difficult to estimate the precision matrix Ω . The standard linear discriminant rule thus fails completely. Several regularized classification methods, including the regularized LDA (Wu et al., 2009), covariance-regularized classification (Witten and Tibshirani, 2009), and hard thresholding (Shao et al., 2011), have been proposed for classification of high-dimensional data. However, all these methods rely on the individual sparsity assumptions on Ω (or Σ) and δ . A fundamental quantity in LDA is the discriminant direction $\beta = \Omega\delta$ and a more flexible assumption is the sparsity of β . In particular, Cai and Liu (2011); Mai et al. (2012) introduced a direct estimation method for the high-dimensional LDA based on the key observation that the ideal Fisher's discriminant rule given in (2.1) depends on the parameters μ_1, μ_2 and Σ primarily through $\beta = \Omega\delta$. They proposed to estimate the discriminant direction β directly instead of estimating Σ and δ separately, under the assumption that β is sparse. The proposed classification rule was shown to be consistent.

Despite much recent progress in methodological development on high-dimensional classification problems, there has been relatively little fundamental study on the optimality theory for the discriminant analysis. Minimax study of high-dimensional discriminant analysis has been considered in Azizyan et al. (2013) and Li et al. (2017) in the special case where the covariance matrix $\Sigma = \sigma^2 I$ for some $\sigma > 0$. However, even in this relatively simple setting there is still a gap between the minimax upper and lower bounds. It is unclear what the optimal rate of convergence for the minimax misclassification risk is and which classification rule is rate optimal under the general Gaussian distribution. The first major goal of the present paper is to provide answers to these questions. Furthermore, although the problem of missing data arises frequently in the analysis of high-dimensional data, compared to the conventional low-dimensional setting, there is a paucity of methods for inference with in-

complete high-dimensional data. The second goal of this paper is to develop an optimality theory for high-dimensional discriminant analysis with incomplete data and to construct in this setting a data-driven adaptive classifier with theoretical guarantees.

Given two random samples, $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n_1}^{(1)} \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_1, \Sigma)$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n_2}^{(2)} \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_2, \Sigma)$, we wish to construct a classifier \hat{C} to classify a future data point \mathbf{Z} drawn from these two distributions with equal prior probabilities, into one of the two classes. Given the observed data, the performance of the classification rule is measured by the misclassification error

$$R_{\boldsymbol{\theta}}(\hat{C}) = \mathbb{P}_{\boldsymbol{\theta}}(\text{label}(\mathbf{Z}) \neq \hat{C}(\mathbf{Z})), \quad (2.2)$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$, $\mathbb{P}_{\boldsymbol{\theta}}$ denotes the probability with respect to $\mathbf{Z} \sim \frac{1}{2}N_p(\boldsymbol{\mu}_1, \Sigma) + \frac{1}{2}N_p(\boldsymbol{\mu}_2, \Sigma)$ and \mathbf{Z} is independent of the observed \mathbf{X} 's. $\text{label}(\mathbf{Z})$ denotes the true class of \mathbf{Z} . For a given classifier \hat{C} , we use the excess misclassification risk relative to the oracle rule (2.1), $R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta})$, to measure the performance of the classifier \hat{C} . Let $n = \min\{n_1, n_2\}$. We consider in this paper a collection of the parameter spaces $\mathcal{G}(s, M_{n,p})$ defined by

$$\begin{aligned} \mathcal{G}(s, M_{n,p}) = \{ \boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p \times p}, \Sigma \succ \mathbf{0}, \\ \|\boldsymbol{\beta}\|_0 \leq s, M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M, M_{n,p} \leq \Delta \leq 3M_{n,p} \}, \end{aligned} \quad (2.3)$$

where $M > 1$ is a constant, $M_{n,p} > 0$ can potentially grow with n and p , and $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ are respectively the largest and smallest eigenvalue of Σ . The notation $\Sigma \succ \mathbf{0}$ means that Σ is symmetric and positive definite. Recall that $\Delta = \sqrt{\boldsymbol{\delta}^\top \Omega \boldsymbol{\delta}}$ and $\boldsymbol{\beta} = \Omega \boldsymbol{\delta}$. The sparsity constraint $\|\boldsymbol{\beta}\|_0 \leq s$, according to the oracle rule (2.1), implies the belief that only a limited number of covariates have discriminating power and contribute to the classification task. In addition, our lower bound results in Theorem 7 show that this sparsity assumption is necessary in general without further constraints of parameter space. Furthermore, we also assume the eigenvalues of the covariance matrix Σ are bounded from below and above. This assumption is commonly used in high-dimensional statistics, ranging from high-dimensional

linear regression (Javanmard and Montanari, 2014), covariance matrix estimation (Cai and Yuan, 2012), classification (Cai and Liu, 2011) and clustering (Cai et al., 2018a).

Combining the upper and lower bounds results given in Section 4.3 leads to the following minimax rates of convergence for the excess misclassification risk.

Theorem 1. *Consider the parameter space $\mathcal{G}(s, M_{n,p})$, s and p approach infinity as n grows to infinity, and $M_{n,p} = o(\sqrt{\frac{n}{s \log p}})$ with $n \rightarrow \infty$,*

1. *If $M_{n,p}$ is a fixed constant not depending on n and p , then for any constant $\alpha \in (0, 1)$, we have*

$$\inf \left\{ r : \inf_{\hat{C}} \sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P} \left(R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \geq r \right) \leq 1 - \alpha \right\} \asymp \frac{s \log p}{n}.$$

2. *If $M_{n,p} \rightarrow \infty$ as $n \rightarrow \infty$, then for sufficiently large n and any constant $\alpha \in (0, 1)$,*

$$\inf \left\{ r : \inf_{\hat{C}} \sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P} \left(R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \geq r \right) \leq 1 - \alpha \right\} \asymp \frac{s \log p}{n} \cdot e^{-(\frac{1}{8} + o(1)) M_{n,p}^2}.$$

It is worth noting that $M_{n,p}$ represents the magnitude of Δ , which is interpreted as the signal-to-noise ratio. As shown in the second case, when the signal-to-noise ratio grows, the classification problem becomes easier and our result precisely characterizes that the convergence rate is exponentially faster with an additional factor $\exp \left(- (1/8 + o(1)) M_{n,p}^2 \right)$.

Furthermore, we propose a three-step data-driven classification rule, called *AdaLDA*, by using an adaptive constrained ℓ_1 minimization approach which takes into account the variability of individual entries. This classification rule is shown to be simultaneously rate optimal over the collection of parameter spaces $\mathcal{G}(s, M_{n,p})$. To the best of our knowledge, this is the first optimality result for classification of high-dimensional Gaussian data. Furthermore, in contrast to many classification rules proposed in the literature, which require to choose tuning parameters, this procedure is data-driven and tuning-free.

In addition, we also consider classification in the presence of missing data. As in the conventional low-dimensional setting, the problem of missing data also arises frequently in the analysis of high-dimensional data from in a range of fields such as genomics, epidemiology, engineering, and social sciences (Libbrecht and Noble, 2015; White et al., 2011; Graham, 2009). Compared to the low-dimensional setting, there are relatively few inferential methods for missing data in the high-dimensional setting. Examples include high-dimensional linear regression (Loh and Wainwright, 2012), sparse principal component analysis (Lounici, 2013), covariance matrix estimation (Cai and Zhang, 2016), and vector autoregressive (VAR) processes (Rao et al., 2017). In this paper, following the missing mechanism considered in the aforementioned papers, we investigate high-dimensional discriminant analysis in the presence of missing observations under the missing completely at random (MCR) model.

We construct a data-driven adaptive classifier with theoretical guarantees based on incomplete data and also develop an optimality theory for high-dimensional linear discriminant analysis under the MCR model. The technical analysis for the case of missing data is much more challenging than that for the complete data, although the classification procedure and the resulting convergence rates look similar. To facilitate the theoretical analysis, we establish a key technical tool, which is a large deviation result for the generalized sample covariance matrix. This is related to the masked covariance matrix estimator considered in Levina and Vershynin (2012) and Chen et al. (2012), see further discussions in Section 2.2.3. This technical tool can be of independent interest as it is potentially useful for other related problems in high-dimensional statistical inference with missing data.

The proposed adaptive classification algorithms can be cast as linear programs and are thus easy to implement. Simulation studies are carried out to investigate the numerical performance of the classification rules. The results show that the proposed classifiers enjoy superior finite sample performance in comparison to existing methods for high-dimensional linear discriminant analysis. The proposed classifiers are also illustrated through an application to the analysis of lung cancer and leukemia datasets. The results show that they

outperform existing methods.

The rest of the paper is organized as follows. In Section 2.2, after basic notation and definitions are reviewed, we introduce an adaptive algorithm for high-dimensional discriminant analysis with the complete data and then propose a more general procedure for the setting of incomplete data. Section 4.3 studies the theoretical properties of these classification rules and related estimators. In addition, minimax lower bounds are given. The upper and lower bounds together establish the optimal rates of convergence for the minimax misclassification risk. Numerical performance of the classification rules are investigated in Section 2.4 and an extension to the multiple-class LDA is discussed in Section 2.5. The proofs of the main results are given in Section 4.8. Technical lemmas are proved in the Supplementary Material (Cai and Zhang, 2018d).

2.2. Methodology

In this section, we firstly introduce an adaptive algorithm for high-dimensional linear discriminant analysis with the complete data. This algorithm is called AdaLDA (**A**daptive **L**inear **D**iscriminant **A**nalysis rule). We then propose a data-driven classifier, called ADAM (**A**daptive linear **D**iscriminant **A**nalysis with randomly **M**issing data), for the incomplete data under the MCR model.

2.2.1. Notation and definitions

We begin with basic notation and definitions. Throughout the paper, vectors are denoted by boldface letters. For a vector $\mathbf{x} \in \mathbb{R}^p$, the usual vector ℓ_0, ℓ_1, ℓ_2 and ℓ_∞ norms are denoted respectively by $\|\mathbf{x}\|_0, \|\mathbf{x}\|_1, \|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_\infty$. Here the ℓ_0 norm counts the number of nonzero entries in a vector. The support of a vector \mathbf{x} is denoted by $\text{supp}(\mathbf{x})$. The symbol \circ denotes the Hadamard product. For $p \in \mathbb{N}$, $[p]$ denotes the set $\{1, 2, \dots, p\}$. For $j \in [p]$, denote by \mathbf{e}_j the j -th canonical basis in \mathbb{R}^p . For a matrix $\Sigma = (\sigma_{ij})_{1 \leq i, j \leq p} \in \mathbb{R}^{p \times p}$, the Frobenius norm is defined as $\|\Sigma\|_F = \sqrt{\sum_{i,j} \sigma_{ij}^2}$ and the spectral norm is defined to be $\|\Sigma\|_2 = \sup_{\|\mathbf{x}\|_2=1} \|\Sigma\mathbf{x}\|_2$. The vector ℓ_∞ norm of the matrix Σ is $|\Sigma|_\infty = \max_{i,j} |\sigma_{ij}|$. For

a symmetric matrix Σ , we use $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ to denote respectively the largest and smallest eigenvalue of Σ . $\Sigma \succ 0$ means that Σ is positive definite. For a positive integer $s < p$, let $\Gamma(s) = \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}_{S^c}\|_1 \leq \|\mathbf{u}_S\|_1, \text{ for some } S \subset [p] \text{ with } |S| = s\}$, where \mathbf{u}_S denotes the subvector of \mathbf{u} confined to S . For two sequences of positive numbers a_n and b_n , $a_n \lesssim b_n$ means that for some constant $c > 0$, $a_n \leq cb_n$ for all n , and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We say an event \mathcal{A}_n holds with high probability if $\liminf_{n \rightarrow \infty} \mathbb{P}(\mathcal{A}_n) = 1$. Finally, $c_0, c_1, c_2, C, C_1, C_2, \dots$ denote generic positive constants that may vary from place to place.

The complete data $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n_1}^{(1)}$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n_2}^{(2)}$ are independent realizations of $\mathbf{X}^{(1)} \sim N_p(\boldsymbol{\mu}_1, \Sigma)$ and $\mathbf{X}^{(2)} \sim N_p(\boldsymbol{\mu}_2, \Sigma)$. We assume $n_1 \asymp n_2$ and define $n = \min\{n_1, n_2\}$. In our asymptotic framework, we let n be the driving asymptotic parameter, s and p approach infinity as n grows to infinity. The missing completely at random (MCR) model assumes that one observes samples $\{\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n_1}^{(1)}\}$ and $\{\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n_2}^{(2)}\}$ with missing values, where the observed coordinates of $\mathbf{X}_t^{(k)}$ are indicated by an independent vector $\mathbf{S}_t^{(k)} \in \{0, 1\}^p$ for $t = 1, \dots, n_k, k = 1, 2$, that is,

$$X_{tj}^{(k)} \text{ is observed if } S_{tj}^{(k)} = 1 \text{ and } X_{tj}^{(k)} \text{ is missing if } S_{tj}^{(k)} = 0; t \in [n_k], j \in [p], k = 1, 2. \quad (2.4)$$

Here $X_{tj}^{(k)}$ and $S_{tj}^{(k)}$ are respectively the j -th coordinate of the vectors $\mathbf{X}_t^{(k)}$ and $\mathbf{S}_t^{(k)}$. Generally, we use the superscript “*” to denote objects related to missing values. The incomplete samples with missing values are denoted by $\mathbf{X}^{(1)*} = \{\mathbf{X}_1^{(1)*}, \dots, \mathbf{X}_{n_1}^{(1)*}\}$ and $\mathbf{X}^{(2)*} = \{\mathbf{X}_1^{(2)*}, \dots, \mathbf{X}_{n_2}^{(2)*}\}$.

Regarding the mechanism for missingness, the MCR model is formally stated as below. This assumption is more general than the one considered previously by Loh and Wainwright (2012) and Lounici (2013).

Assumption 1. (*Missing Completely at Random (MCR)*) $S = \{\mathbf{S}_t^{(k)} \in \{0, 1\}^p : t = 1, \dots, n_k, k = 1, 2\}$ is independent of the values of $\mathbf{X}_t^{(1)}$ and $\mathbf{X}_t^{(2)}$ for $t = 1, \dots, n_k, k = 1, 2$. Here $\mathbf{S}_t^{(k)}$ can be either deterministic or random, but independent of $\mathbf{X}_t^{(1)}$ and $\mathbf{X}_t^{(2)}$.

A major goal of the present paper is to construct a classification rule \hat{C} in the high dimensional setting where $p \gg n$ for both complete and incomplete data.

2.2.2. Data-driven adaptive classifier for complete data

We first consider the case of complete data. In this setting, as mentioned in the introduction, a number of high-dimensional linear discriminant rules have been proposed in the literature. In particular, Cai and Liu (2011) introduced a classification rule called LPD (Linear Programming Discriminant) rule by directly estimating the discriminant direction β through solving the following optimization problem:

$$\hat{\beta}_{\text{LPD}} = \arg \min_{\beta} \left\{ \|\beta\|_1 : \text{subject to } \|\hat{\Sigma}\beta - (\hat{\mu}_2 - \hat{\mu}_1)\|_{\infty} \leq \lambda_n \right\}, \quad (2.5)$$

where $\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}$ are sample means and pooled sample covariance matrix respectively, and $\lambda_n = C\sqrt{\log p/n}$ is the tuning parameter with some constant C . Based on $\hat{\beta}_{\text{LPD}}$, the LPD rule is then given by

$$\hat{C}_{\text{LPD}}(\mathbf{Z}) = \begin{cases} 1, & (\mathbf{Z} - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2})^{\top} \hat{\beta}_{\text{LPD}} < 0 \\ 2, & (\mathbf{Z} - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2})^{\top} \hat{\beta}_{\text{LPD}} \geq 0 \end{cases}. \quad (2.6)$$

The LPD rule is easy to implement and Cai and Liu (2011) proves the consistency of LPD when the tuning parameter λ_n is appropriately chosen. However, it has three drawbacks. One major drawback of the LPD rule is that it uses a common constraint λ_n for all coordinates of $\mathbf{a} = \hat{\Sigma}\beta - (\hat{\mu}_2 - \hat{\mu}_1)$. This essentially treats the random vector \mathbf{a} as homoscedastic, while in fact \mathbf{a} is intrinsically heteroscedastic and the coordinates could have a wide range of variability. The resulting estimator $\hat{\beta}_{\text{LPD}}$ obtained in (2.5) of the discriminant direction β has yet to be shown as rate optimal; secondly, the procedure is not adaptive in the sense that the tuning parameter λ_n is not fully specified and needs to be chosen through an empirical method such as cross-validation. The third drawback is that the LPD rule \hat{C}_{LPD} does not come with theoretical optimality guarantees.

To resolve these drawbacks, we introduce an adaptive algorithm for high-dimensional LDA with complete data, called AdaLDA (**A**daptive **L**inear **D**iscriminant **A**nalysis rule), which takes into account the heteroscedasticity of the random vector $\mathbf{a} = \hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$. AdaLDA is fully data-driven and tuning-free and will be shown to be minimax rate optimal later. Before we describe the classifier in detail, it is helpful to state the following key technical result which provides the motivation for the new procedure.

Lemma 1. *Suppose $\{\mathbf{X}_t^{(1)}\}_{t=1}^{n_1}$ and $\{\mathbf{X}_t^{(2)}\}_{t=1}^{n_2}$ are i.i.d. random samples from $N_p(\boldsymbol{\mu}_1, \Sigma)$ and $N_p(\boldsymbol{\mu}_2, \Sigma)$ respectively with $\Sigma = (\sigma_{ij})_{1 \leq i, j \leq p}$. Let $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, $\boldsymbol{\beta} = \Omega\boldsymbol{\delta}$, $\Delta = \sqrt{\boldsymbol{\beta}^\top \boldsymbol{\delta}}$ and $\mathbf{a} = \hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$, where $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\Sigma}$ are sample means and pooled sample covariance matrix respectively. Then*

$$\text{Var}(a_j) = \frac{n-1}{2n^2}(\sigma_{jj}\Delta^2 + \delta_j^2) + \frac{2}{n}\sigma_{jj}, \quad j = 1, \dots, p.$$

Furthermore, with probability at least $1 - 4p^{-1}$,

$$|a_j| \leq 4\sqrt{\frac{\log p}{n}} \cdot \sqrt{\sigma_{jj}} \cdot \left(\sqrt{\frac{25\Delta^2}{2} + 1} \right), \quad j = 1, \dots, p. \quad (2.7)$$

A major step in the construction of the AdaLDA classifier is using Lemma 1 to construct an element-wise constraint for $\hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$, which relies on an accurate estimation of the right hand side of (2.7). In (2.7), σ_{jj} can be easily estimated by the sample variances $\hat{\sigma}_{jj}$, but Δ^2 is harder to estimate. In the following, we begin by constructing a preliminary estimator $\tilde{\boldsymbol{\beta}}$, estimating Δ^2 by $|\tilde{\boldsymbol{\beta}}^\top(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)|$, and then applying Lemma 1 to refine the estimation of $\boldsymbol{\beta}$. The data-driven adaptive classifier AdaLDA is constructed in three steps.

Step 1 (Estimating Δ^2). Fix $\lambda_0 = 25/2$, we estimate β by a preliminary estimator

$$\begin{aligned} \tilde{\beta} &= \arg \min_{\beta} \|\beta\|_1 \\ \text{subject to } |e_j^\top (\hat{\Sigma}\beta - (\hat{\mu}_2 - \hat{\mu}_1))| &\leq 4\sqrt{\frac{\log p}{n}} \cdot \sqrt{\hat{\sigma}_{jj}} \cdot (\lambda_0 \beta^\top (\hat{\mu}_2 - \hat{\mu}_1) + 1), \quad j \in [p]. \end{aligned} \quad (2.8)$$

Then we estimate Δ^2 by $\hat{\Delta}^2 = |\tilde{\beta}^\top (\hat{\mu}_2 - \hat{\mu}_1)|$.

Step 2 (Adaptive estimation of β). Given $\hat{\Delta}^2$, the final estimator $\hat{\beta}_{\text{AdaLDA}}$ of β is constructed through the following linear optimization

$$\begin{aligned} \hat{\beta}_{\text{AdaLDA}} &= \arg \min_{\beta} \|\beta\|_1 \\ \text{subject to } |e_j^\top (\hat{\Sigma}\beta - (\hat{\mu}_2 - \hat{\mu}_1))| &\leq 4\sqrt{\frac{\log p}{n}} \cdot \sqrt{\hat{\sigma}_{jj}(\lambda_0 \hat{\Delta}^2 + 1)}, \quad j \in [p]. \end{aligned} \quad (2.9)$$

Step 3 (Construction of AdaLDA). The AdaLDA classification rule is obtained by plugging $\hat{\beta}_{\text{AdaLDA}}$ into Fisher's rule (2.1),

$$\hat{C}_{\text{AdaLDA}}(\mathbf{Z}) = \begin{cases} 1, & (\mathbf{Z} - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2})^\top \hat{\beta}_{\text{AdaLDA}} < 0, \\ 2, & (\mathbf{Z} - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2})^\top \hat{\beta}_{\text{AdaLDA}} \geq 0. \end{cases} \quad (2.10)$$

Note that there is a square root on Δ^2 (or $\hat{\Delta}^2$) in both (2.7) and (2.9), but this square root is removed in (2.8). Intuitively, by removing the square root in (2.8), Step 1 becomes a linear program, which provides a computationally efficient but sub-optimal estimator. This estimator is then refined to be rate-optimal in Step 2 by adding back the square root. This two-step idea is in the similar spirit as that in Cai et al. (2016b) for adaptive estimation of precision matrices. Despite this similarity in the ideas for the construction procedures, the problem considered and the technical tools applied in the present paper are very different

from those in Cai et al. (2016b).

This classification rule does not require a tuning parameter and the estimator $\hat{\beta}_{\text{AdaLDA}}$ is solved by optimizing a linear program with an element-wise constraint, adapting to individual variability of $\hat{\Sigma}\beta - (\hat{\mu}_2 - \hat{\mu}_1)$. It will be shown in Section 4.3 that the AdaLDA classification rule is adaptively minimax rate optimal. Our theoretical analysis also shows that the resulting estimator $\hat{\beta}_{\text{AdaLDA}}$ is rate optimally adaptive whenever λ_0 is a sufficiently large constant. In particular, it can be taken as fixed at $\lambda_0 = 25/2$, which is derived from the concentration inequality given in Lemma 1.

Remark 1. Note that the optimization problems (2.8) and (2.9) are both linear programs, so the proposed AdaLDA rule is computationally easy to implement. In contrast, the LPD uses a universal tuning parameter $\lambda_n = C\sqrt{\log p/n}$, whose value is usually chosen by cross-validation. This tuning procedure is computationally costly. In addition, cross-validation tends to overfit (Friedman et al., 2001). Therefore, estimator obtained through cross-validation can be variable and its theoretical properties are unclear, while the AdaLDA procedure does not depend on any unknown parameter and the estimator will be shown to be minimax rate optimal.

2.2.3. ADAM with randomly missing data

We now turn to the case of incomplete data under the MCR model. To generalize AdaLDA to the incomplete data case, we proceed by firstly estimating μ_1 , μ_2 and Σ . The following estimators follow the idea in Cai and Zhang (2016), and for completeness, we present their proposed estimators below. Let

$$n_{ij}^{(k)*} = \sum_{t=1}^{n_k} S_{ti}^{(k)} S_{tj}^{(k)}, \quad 1 \leq i, j \leq p, k = 1, 2.$$

Here $n_{ij}^{(k)*}$ is the number of vectors $\mathbf{X}_t^{(k)}$ in which the i^{th} and j^{th} entries are both observed. In addition, we denote $n_i^{(k)*} = n_{ii}^{(k)*}$ for simplicity and

$$n_{\min}^* = \min_{i,j,k} n_{ij}^{(k)*}. \quad (2.11)$$

In the presence of missing values, the usual sample mean and sample covariance matrix can no longer be calculated. Instead, the ‘‘generalized sample mean’’ is proposed, defined by

$$\begin{aligned} \hat{\boldsymbol{\mu}}_1 &= (\hat{\mu}_{1i}^*)_{1 \leq i \leq p} \quad \text{with} \quad \hat{\mu}_{1i}^* = \frac{1}{n_i^{(1)*}} \sum_{t=1}^{n_1} X_{ti}^{(1)} S_{ti}^{(1)}, \quad 1 \leq i \leq p; \\ \hat{\boldsymbol{\mu}}_2 &= (\hat{\mu}_{2i}^*)_{1 \leq i \leq p} \quad \text{with} \quad \hat{\mu}_{2i}^* = \frac{1}{n_i^{(2)*}} \sum_{t=1}^{n_2} X_{ti}^{(2)} S_{ti}^{(2)}, \quad 1 \leq i \leq p. \end{aligned}$$

The ‘‘generalized sample covariance matrix’’ is then defined by $\hat{\boldsymbol{\Sigma}} = (\hat{\sigma}_{ij}^*)_{1 \leq i,j \leq p}$ with

$$\hat{\sigma}_{ij}^* = \frac{1}{n_{ij}^{(1)*} + n_{ij}^{(2)*}} \left(\sum_{t=1}^{n_1} (X_{ti}^{(1)} - \hat{\mu}_{1i}^*)(X_{tj}^{(1)} - \hat{\mu}_{1j}^*) S_{ti}^{(1)} S_{tj}^{(1)} + \sum_{t=1}^{n_2} (X_{ti}^{(2)} - \hat{\mu}_{2i}^*)(X_{tj}^{(2)} - \hat{\mu}_{2j}^*) S_{ti}^{(2)} S_{tj}^{(2)} \right).$$

For these generalized estimators, we have the following bound under the MCR model.

Lemma 2. *Let $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, $\boldsymbol{\beta} = \Omega \boldsymbol{\delta}$, $\Delta = \sqrt{\boldsymbol{\delta}^\top \Omega \boldsymbol{\delta}}$ and $\mathbf{a}^* = \hat{\boldsymbol{\Sigma}} \boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$. Then conditioning on \mathbf{S} , we have with high probability,*

$$|a_j^*| \leq 4 \sqrt{\frac{\log p}{n_{\min}^*}} \cdot \sqrt{\sigma_{jj}} \cdot \left(\sqrt{64 \Delta^2 + 1} \right), \quad j = 1, \dots, p. \quad (2.12)$$

Remark 2. Although the above result has a form that is similar to Lemma 1, its derivation is quite different and relies on a new technical tool, the large deviation bound for $\hat{\boldsymbol{\Sigma}}$. This is of independent interest and is related to that of the masked sample covariance estimator considered in Levina and Vershynin (2012) and Chen et al. (2012). In particular, the masked sample covariance estimator considered in Chen et al. (2012) applies the mask matrix to the sample covariance matrix, while our proposed estimator $\hat{\boldsymbol{\Sigma}}$ can be interpreted as applying

the mask matrix to each *i.i.d.* sample, and thus is more general. The proof of Lemma 2 uses the idea of Lemma 2.1 in Cai and Zhang (2016), but yields a sharper bound. The detailed proof is given in Section A.3.2 in the supplement (Cai and Zhang, 2018d).

We propose to estimate β adaptively and construct ADAM (**A**daptive linear **D**iscriminant **A**nalysis with randomly **M**issing data) in the following way:

Step 1 (Estimating Δ^2). Fix $\lambda_1 = 64$. We estimate β by a preliminary estimator

$$\begin{aligned} \tilde{\beta} &= \arg \min_{\beta} \|\beta\|_1 \\ &\text{subject to } |e_j^\top (\hat{\Sigma}\beta - (\hat{\mu}_2 - \hat{\mu}_1))| \leq 4\sqrt{\frac{\log p}{n_{min}^*}} \cdot \sqrt{\hat{\sigma}_{jj}^*} \cdot (\lambda_1 \beta^\top (\hat{\mu}_2 - \hat{\mu}_1) + 1), \quad j \in [p]. \end{aligned} \quad (2.13)$$

Then we estimate Δ^2 by $\hat{\Delta}^{*2} = |\tilde{\beta}^\top (\hat{\mu}_2 - \hat{\mu}_1)|$.

Step 2 (Adaptive estimation of β). Given $\hat{\Delta}^{*2}$, the final estimator $\hat{\beta}_{\text{ADAM}}$ of β is constructed by the following linear optimization problem

$$\begin{aligned} \hat{\beta}_{\text{ADAM}} &= \arg \min_{\beta} \|\beta\|_1 \\ &\text{subject to } |e_j^\top (\hat{\Sigma}\beta - (\hat{\mu}_2 - \hat{\mu}_1))| \leq 4\sqrt{\frac{\log p}{n_{min}^*}} \cdot \sqrt{\hat{\sigma}_{jj}^* (\lambda_1 \hat{\Delta}^{*2} + 1)}, \quad j \in [p]. \end{aligned} \quad (2.14)$$

Step 3 (Construction of ADAM). Given the estimator $\hat{\beta}_{\text{ADAM}}$ of the discriminant direction β , we then construct the following ADAM classification rule by plugging $\hat{\beta}_{\text{ADAM}}$ into the oracle rule (2.1):

$$\hat{C}_{\text{ADAM}}(\mathbf{Z}) = \begin{cases} 1, & (\mathbf{Z} - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2})^\top \hat{\beta}_{\text{ADAM}} < 0, \\ 2, & (\mathbf{Z} - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2})^\top \hat{\beta}_{\text{ADAM}} \geq 0. \end{cases} \quad (2.15)$$

As shown in Section 4.3, \hat{C}_{ADAM} has the similar theoretical performance as \hat{C}_{AdaLDA} .

Remark 3. The ADAM algorithm is designed for the MCR model. Extensions to other missing mechanism such as missing not at random (MNAR) is possible but challenging. In such a setting even parametric models are often not identifiable (Miao et al., 2016; Robins and Ritov, 1997). Several authors have studied the problem of identification under MNAR with different conditions (Rotnitzky and Robins, 1997; Sun et al., 2016; Tchetgen Tchetgen and Wirth, 2017). The consistency of our algorithm only relies on consistent estimation of the mean vectors and the covariance matrix. Therefore, if the means and the covariance matrix can be estimated consistently under some MNAR model, for example, by using EM algorithm and imputing the missing values (Schneider, 2001), we can then construct a consistent classification rule based on these estimators. However, such imputation techniques are computationally intensive (Lounici, 2014).

2.3. Theoretical properties of AdaLDA and ADAM

In this section, we develop an optimality theory for high-dimensional linear discriminant analysis for both the complete data and the incomplete data settings. We first investigate the theoretical properties of the AdaLDA and ADAM algorithms proposed in Section 2.2 and obtain the upper bounds for the excess misclassification risk. We then establish the lower bounds for the rate of convergence. The upper and lower bounds together yield the minimax rates of convergence and show that AdaLDA and ADAM are adaptively rate optimal.

2.3.1. Theoretical Analysis of AdaLDA

We begin by considering the properties of the estimator $\hat{\beta}_{\text{AdaLDA}}$ of the discriminant direction β . The following theorem shows that $\hat{\beta}_{\text{AdaLDA}}$ attains the convergence rate of $M_{n,p}\sqrt{s \log p/n}$ over the class of sparse discriminating directions $\mathcal{G}(s, M_{n,p})$ defined in (4.18). The matching lower bound given in Section 3.3 implies that this rate is optimal. Therefore, AdaLDA adapts to both the sparsity pattern of β as well as the signal-to-noise

ratio Δ .

Theorem 2. Consider the parameter space $\mathcal{G}(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$. Suppose $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n_1}^{(1)} \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_1, \Sigma)$, $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n_2}^{(2)} \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_2, \Sigma)$ and $n_1 \asymp n_2$. Assume that $M_{n,p} \sqrt{\frac{s \log p}{n}} = o(1)$. Then

$$\sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{E}[\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} - \boldsymbol{\beta}\|_2] \lesssim M_{n,p} \sqrt{\frac{s \log p}{n}}.$$

We then proceed to characterize the accuracy of the classification rule \hat{C}_{AdaLDA} , measured by the excess misclassification risk $R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta})$. Note that the conditional misclassification rate of \hat{C}_{AdaLDA} given the two samples can be analytically calculated as

$$R_{\boldsymbol{\theta}}(\hat{C}_{\text{AdaLDA}}) = \frac{1}{2} \Phi \left(-\frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\beta}}_{\text{AdaLDA}}}{\sqrt{\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}^\top \Sigma \hat{\boldsymbol{\beta}}_{\text{AdaLDA}}}} \right) + \frac{1}{2} \bar{\Phi} \left(-\frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^\top \hat{\boldsymbol{\beta}}_{\text{AdaLDA}}}{\sqrt{\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}^\top \Sigma \hat{\boldsymbol{\beta}}_{\text{AdaLDA}}}} \right),$$

where $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2$ and $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$.

We are interested in the excess misclassification risk $R_{\boldsymbol{\theta}}(\hat{C}_{\text{AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta})$. That is, we compare \hat{C}_{AdaLDA} with the oracle Fisher's rule, whose risk is given by

$$R_{\text{opt}}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} R_{\boldsymbol{\theta}}(C_{\boldsymbol{\theta}}) = \Phi \left(-\frac{1}{2} \Delta \right).$$

The following theorem provides an upper bound for the excess misclassification risk of the AdaLDA rule.

Theorem 3. Consider the parameter space $\mathcal{G}(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$ and assume the conditions in Theorem 2 hold.

1. If $M_{n,p} \leq C_b$ for some $C_b > 0$, then there exists some constant $C > 0$,

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P} \left(R_{\boldsymbol{\theta}}(\hat{C}_{\text{AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta}) \leq C \cdot \frac{s \log p}{n} \right) \geq 1 - 8p^{-1}.$$

2. If $M_{n,p} \rightarrow \infty$ as $n \rightarrow \infty$, then there exist some constant $C > 0$ and $\delta_n = o(1)$, such that

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P} \left(R_{\boldsymbol{\theta}}(\hat{C}_{\text{AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta}) \leq C \cdot e^{-(\frac{1}{8} + \delta_n)M_{n,p}^2} \cdot \frac{s \log p}{n} \right) \geq 1 - 8p^{-1}.$$

Remark 4. The results in Theorem 3 improve the convergence rate of the misclassification risk of the LPD rule given in Cai and Liu (2011). Consider the first case where $M_{n,p}$ is a constant not depending on n and p , Theorem 3 of Cai and Liu (2011) shows that the convergence rate is $R_{\boldsymbol{\theta}}(\hat{C}_{\text{LPD}}) - R_{\text{opt}}(\boldsymbol{\theta}) = O_P((s \log p/n)^{1/2})$, while Theorem 3 here shows a faster rate $O_P((s \log p/n))$ when $M_{n,p}$ is a constant. Indeed, this improvement is due to a careful analysis of the misclassification error. In the proof of Theorem 3, it can be seen that the first order approximation error is vanishing, and only the second order approximation error, which has a faster convergence rate, remains. The lower bounds given in Section 3.3 show that both convergence rates in Theorem 3 are indeed optimal.

Similarly, upper bounds on the relative misclassification risk can be obtained.

Proposition 1. Consider the parameter space $\mathcal{G}(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$ and assume the conditions in Theorem 2 hold.

1. If $M_{n,p} \leq C_b$ for some $C_b > 0$, then there exists some constant $C > 0$,

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P} \left(\frac{R_{\boldsymbol{\theta}}(\hat{C}_{\text{AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta})}{R_{\text{opt}}(\boldsymbol{\theta})} \leq C \cdot \frac{s \log p}{n} \right) \geq 1 - 8p^{-1}.$$

2. If $M_{n,p} \rightarrow \infty$ as $n \rightarrow \infty$, then there exist some constant $C > 0$, such that

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P} \left(\frac{R_{\boldsymbol{\theta}}(\hat{C}_{\text{AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta})}{R_{\text{opt}}(\boldsymbol{\theta})} \leq C \cdot M_{n,p}^4 \cdot \frac{s \log p}{n} \right) \geq 1 - 8p^{-1}.$$

Remark 5. The results in Proposition 3 show that the relative misclassification risk has a worse convergence rate when the magnitude of signal-to-noise ratio $M_{n,p}$ becomes larger. This is expected as when $M_{n,p}$ becomes larger, the classification problem itself becomes

easier and the oracle misclassification risk is very small, making the oracle classification rule harder to be mimicked.

2.3.2. Theoretical Analysis of ADAM

We now investigate the theoretical properties of the ADAM procedure in the presence of missing data. Similar rates of convergence for estimation and excess misclassification risk can be obtained, but the technical analysis is much more involved under the MCR model.

Under the MCR model, suppose that the missingness pattern $S \in \{0, 1\}^{n_1 \times p} \times \{0, 1\}^{n_2 \times p}$ is a realization of a distribution \mathcal{F} . We consider the distribution space $\Psi(n_0; n, p)$ given by

$$\Psi(n_0; n, p) = \{\mathcal{F} : \mathbb{P}_{S \sim \mathcal{F}}(c_1 n_0 \leq n_{\min}^*(S) \leq c_2 n_0) \geq 1 - p^{-1}\},$$

for some constants $c_1, c_2 > 0$, and $n_{\min}^*(S)$ is defined for S as in (2.11).

Remark 6. This distribution space includes the missing uniformly and completely at random (MUCR) model considered in Loh and Wainwright (2012); Lounici (2013) and Lounici (2014). More specifically, MUCR model assumes each entry $X_{i,j}^{(k)}$ ($k \in [2], i \in [n_k], j \in [p]$) is missing independently with probability ϵ . As shown in Section A.6 in the supplement, when $\frac{1}{(1-\epsilon)^2} \sqrt{\frac{\log p}{n}} = o(1)$ as $n \rightarrow \infty$, the MUCR model is in the distribution space $\Psi(n(1-\epsilon)^2; n, p)$.

In addition, this distribution space allows a more general variant of MUCR model that each entry $X_{i,j}^{(k)}$ is missing independently with different probabilities $\epsilon_{ij}^{(k)}$. If we assume $\tilde{c}_1 \cdot \epsilon \leq \min_{i,j,k} \epsilon_{ij}^{(k)} \leq \max_{i,j,k} \epsilon_{ij}^{(k)} \leq \tilde{c}_2 \cdot \epsilon$ for some constants $\tilde{c}_1, \tilde{c}_2 > 0$, then use the similar technique, this missingness pattern is included in $\Psi(n(1-\epsilon)^2; n, p)$ when $\frac{1}{(1-\epsilon)^2} \sqrt{\frac{\log p}{n}} = o(1)$ as $n \rightarrow \infty$.

The following two theorems provide respectively the convergence rates for the discriminating direction estimator $\hat{\beta}_{\text{ADAM}}$ and the excess misclassification rate of \hat{C}_{ADAM} over the parameter space $\mathcal{G}(s, M_{n,p})$ for θ and the distribution space $\Psi(n_0; n, p)$.

Theorem 4. Consider the parameter space $\mathcal{G}(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$ and the distribution space $\Psi(n_0; n, p)$ with $M_{n,p} \sqrt{\frac{s \log p}{n_0}} = o(1)$. Suppose $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n_1}^{(1)}$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n_2}^{(2)}$ are i.i.d. samples from $N_p(\boldsymbol{\mu}_1, \Sigma)$ and $N_p(\boldsymbol{\mu}_2, \Sigma)$ respectively. Assuming that $\mathbf{X}_1^{*(1)}, \dots, \mathbf{X}_{n_1}^{*(1)}$ and $\mathbf{X}_1^{*(2)}, \dots, \mathbf{X}_{n_2}^{*(2)}$ defined in (2.4) is observed and Assumption 1 with $S = \{\mathbf{S}_t^{(k)}\}_{t \in [n_k], k \in [2]}$ holds. Then the risk of estimating the discriminant direction $\boldsymbol{\beta}$ by ADAM satisfies

$$\sup_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \Psi(n_0; n, p)}} \mathbb{E}[\|\hat{\boldsymbol{\beta}}_{\text{ADAM}} - \boldsymbol{\beta}\|_2] \lesssim M_{n,p} \sqrt{\frac{s \log p}{n_0}}.$$

Theorem 5. Suppose the conditions of Theorem 4 hold.

1. If $M_{n,p} \leq C_b$ for some $C_b > 0$, then there exists some constant $C > 0$, such that

$$\inf_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \Psi(n_0; n, p)}} \mathbb{P} \left(R_{\boldsymbol{\theta}}(\hat{C}_{\text{ADAM}}) - R_{\text{opt}}(\boldsymbol{\theta}) \leq C \cdot \frac{s \log p}{n_0} \right) \geq 1 - 12p^{-1}.$$

2. If $M_{n,p} \rightarrow \infty$ as $n \rightarrow \infty$, then there exist some constant $C > 0$ and $\delta_n = o(1)$, such that

$$\inf_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \Psi(n_0; n, p)}} \mathbb{P} \left(R_{\boldsymbol{\theta}}(\hat{C}_{\text{ADAM}}) - R_{\text{opt}}(\boldsymbol{\theta}) \leq C \cdot e^{-(\frac{1}{8} + \delta_n)M_{n,p}^2} \cdot \frac{s \log p}{n_0} \right) \geq 1 - 12p^{-1}.$$

In the complete data case, we have $n_0 = n$, so the rates of convergence shown in Theorem 4 and 5 match those in Theorems 2 and 3.

Similarly, upper bounds for the relative misclassification risks can be obtained.

Proposition 2. Suppose the conditions of Theorem 4 hold.

1. If $M_{n,p} \leq C_b$ for some $C_b > 0$, then there exists some constant $C > 0$,

$$\inf_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \Psi(n_0; n, p)}} \mathbb{P} \left(\frac{R_{\boldsymbol{\theta}}(\hat{C}_{\text{ADAM}}) - R_{\text{opt}}(\boldsymbol{\theta})}{R_{\text{opt}}(\boldsymbol{\theta})} \leq C \cdot \frac{s \log p}{n} \right) \geq 1 - 12p^{-1}.$$

2. If $M_{n,p} \rightarrow \infty$ as $n \rightarrow \infty$, then there exist some constant $C > 0$, such that

$$\inf_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \Psi(n_0; n, p)}} \mathbb{P} \left(\frac{R_{\boldsymbol{\theta}}(\hat{C}_{\text{ADAM}}) - R_{\text{opt}}(\boldsymbol{\theta})}{R_{\text{opt}}(\boldsymbol{\theta})} \leq C \cdot M_{n,p}^4 \cdot \frac{s \log p}{n} \right) \geq 1 - 12p^{-1}.$$

In addition, in the special case of MUCR model, Theorem 4 and 5 imply the following result.

Corollary 1. *Under the conditions of Theorem 3 and consider the MUCR model with missing probability ϵ . If $(M_{n,p}^2 \frac{s \log p}{n} \vee \sqrt{\frac{\log p}{n}}) \cdot \frac{1}{(1-\epsilon)^2} = o(1)$, then the risk of estimating the discriminant direction $\boldsymbol{\beta}$ by ADAM over the class $\mathcal{G}(s, M_{n,p})$ satisfies*

$$\sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{E}[\|\hat{\boldsymbol{\beta}}_{\text{ADAM}} - \boldsymbol{\beta}\|_2] \lesssim M_{n,p} \sqrt{\frac{s \log p}{n(1-\epsilon)^2}}.$$

Moreover, there exist constant $C > 0$ and $\delta_n = o(1)$, such that the excess misclassification risk over the class $\mathcal{G}(s, M_{n,p})$ satisfies

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P} \left(R_{\boldsymbol{\theta}}(\hat{C}_{\text{ADAM}}) - R_{\text{opt}}(\boldsymbol{\theta}) \leq C \cdot e^{-(\frac{1}{8} + \delta_n)M_{n,p}^2} \cdot \frac{s \log p}{n(1-\epsilon)^2} \right) \geq 1 - 13p^{-1}.$$

This result shows that, although the sample size only loses a proportion of ϵ , the convergence rates for the estimation risk and misclassification rate shrunk at the rate of $n(1-\epsilon)^2$ under the MUCR model.

2.3.3. Minimax lower bounds

To understand the difficulty of the classification problem and the related estimation problem as well as to establish the optimality for the AdaLDA and ADAM classifiers, it is essential to obtain the minimax lower bounds for the estimation risk and the excess misclassification risk. In this section, we only state the results for the missing data setting as the complete data setting can be treated as a special case. The following lower bound results show that the rates of convergence obtained by AdaLDA and ADAM algorithms are indeed optimal,

for both estimation of the discriminant direction β and classification.

Theorem 6. Consider the parameter space $\mathcal{G}(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$ and the distribution space $\Psi(n_0; n, p)$ with $M_{n,p} \sqrt{\frac{s \log p}{n_0}} = o(1)$. For any $n_0 > 1$, suppose $1 \leq s \leq o(\frac{n_0}{\log p})$ and $\frac{\log p}{\log(p/s)} = O(1)$. Then under MCR model, the minimax risk of estimating the discriminant direction β over the class $G(s, M_{n,p})$ and $\Psi(n_0; n, p)$ satisfies

$$\inf_{\hat{\beta}} \sup_{\substack{\theta \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \Psi(n_0; n, p)}} \mathbb{E}[\|\hat{\beta} - \beta\|_2] \gtrsim M_{n,p} \sqrt{\frac{s \log p}{n_0}}.$$

Theorem 7. Consider the parameter space $\mathcal{G}(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$ and the distribution space $\Psi(n_0; n, p)$ with $M_{n,p} \sqrt{\frac{s \log p}{n_0}} = o(1)$. For any $n_0 \geq 1$, suppose $1 \leq s \leq o(\frac{n_0}{\log p})$ and $\frac{\log p}{\log(p/s)} = O(1)$. Then under the MCR model, the minimax risk of the excess misclassification error over the class $G(s, M_{n,p})$ and $\Psi(n_0; n, p)$ satisfies that

1. If $M_{n,p} \leq C_b$ for some $C_b > 0$, then for any $\alpha > 0$, there are some constants $C_\alpha > 0$ such that

$$\inf_{\hat{C}} \sup_{\substack{\theta \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \Psi(n_0; n, p)}} \mathbb{P}(R_\theta(\hat{C}) - R_{\text{opt}}(\theta) \geq C_\alpha \cdot \frac{s \log p}{n_0}) \geq 1 - \alpha.$$

2. If $M_{n,p} \rightarrow \infty$ as $n \rightarrow \infty$, then for any $\alpha > 0$, there are some constants $C_\alpha > 0$ and $\tilde{\delta}_n = o(1)$ such that

$$\inf_{\hat{C}} \sup_{\substack{\theta \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \Psi(n_0; n, p)}} \mathbb{P}(R_\theta(\hat{C}) - R_{\text{opt}}(\theta) \geq C_\alpha \cdot e^{-(\frac{1}{8} + \tilde{\delta}_n) M_{n,p}^2} \cdot \frac{s \log p}{n_0}) \geq 1 - \alpha.$$

Remark 7. In the complete data case, $n_{\min}^* = \min\{n_1, n_2\} = n$, so Theorems 6 and 7 together with Theorems 1-4 imply that both AdaLDA and ADAM algorithms attain the optimal rates of convergence in terms of estimation and classification error.

We should also note that the proof of Theorem 7 is not straightforward. This is partially due to the fact that the excess risk $R_\theta(\hat{C}) - R_{\text{opt}}(\theta)$ does not satisfy the triangle inequality that is required by standard lower bound techniques. A key technique here is to make a

connection to an alternative risk function. For a generic classification rule \hat{C} , we define

$$L_{\boldsymbol{\theta}}(\hat{C}) = \mathbb{P}_{\boldsymbol{\theta}}(\hat{C}(\mathbf{Z}) \neq C_{\boldsymbol{\theta}}(\mathbf{Z})), \quad (2.16)$$

where $C_{\boldsymbol{\theta}}(\mathbf{Z})$ is the Fisher's linear discriminant rule in (2.1). The following lemma enables us to reduce the loss $R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta})$ to the risk function $L_{\boldsymbol{\theta}}(\hat{C})$.

Lemma 3. *Let $\mathbf{Z} \sim \frac{1}{2}N_p(\boldsymbol{\mu}_1, \Sigma) + \frac{1}{2}N_p(\boldsymbol{\mu}_2, \Sigma)$ with parameter $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$. If a classifier \hat{C} satisfying $L_{\boldsymbol{\theta}}(\hat{C}) = o(1)$ as $n \rightarrow \infty$, then for sufficiently large n ,*

$$R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \geq \frac{\sqrt{2\pi}\Delta}{8} e^{\Delta^2/8} \cdot L_{\boldsymbol{\theta}}^2(\hat{C}).$$

Lemma 12 shows the relationship between the risk function $R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta})$ and a more "standard" risk function $L_{\boldsymbol{\theta}}(\hat{C})$, who has the following property which served the same purpose as the triangle inequality.

Lemma 4. *Let $\boldsymbol{\theta} = (\boldsymbol{\mu}, -\boldsymbol{\mu}, I_p)$ and $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\mu}}, -\tilde{\boldsymbol{\mu}}, I_p)$ with $\|\boldsymbol{\mu}\|_2 = \|\tilde{\boldsymbol{\mu}}\|_2 = \Delta/2$. For any classifier C , if $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|_2 = o(1)$ as $n \rightarrow \infty$, then for sufficiently large n ,*

$$L_{\boldsymbol{\theta}}(C) + L_{\tilde{\boldsymbol{\theta}}}(C) \geq \frac{1}{\Delta} e^{-\Delta^2/8} \cdot \|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|_2.$$

Using Lemmas 12 and 4, we can then use Fano's inequality to complete the proof of Theorem 7. The details are shown in Section 4.8.

In addition, similar minimax lower bounds for estimating $\boldsymbol{\beta}$ and the excess misclassification error can be established under the MUCR model. The following result shows that the convergence rates in Corollary 1 are minimax rate optimal.

Theorem 8. *Under the conditions of Theorem 6 and MUCR model with missing probability ϵ , and further assume that $((M_{n,p}^2 \frac{s \log p}{n}) \vee \sqrt{\frac{\log p}{n}}) \cdot \frac{1}{(1-\epsilon)^2} = o(1)$, then the minimax risk of estimating the discriminant direction $\boldsymbol{\beta}$ by ADAM over the class $G(s, M_{n,p})$ under the*

MUCR model satisfies

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\theta} \in G(s, M_{n,p})} \mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2] \gtrsim M_{n,p} \sqrt{\frac{s \log p}{n(1-\epsilon)^2}}.$$

Moreover, if $M_{n,p} \rightarrow \infty$ and $\epsilon < 1 - c_B$ for some $c_B \in (0, 1)$, the minimax risk of the misclassification error over the class $G(s, M_{n,p})$ satisfies that for any $\alpha, \delta > 0$, there are some constants $C_\alpha > 0$, such that

$$\inf_{\hat{C}} \sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P}(R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \geq C_\alpha \cdot e^{-(\frac{1}{8} + \delta)M_{n,p}^2} \cdot \frac{s \log p}{n(1-\epsilon)^2}) \geq 1 - \alpha.$$

2.4. Numerical results

The proposed AdaLDA and ADAM classifiers are easy to implement, and the MATLAB code is available at <https://github.com/linjunz/ADAM>. We investigate in this section the numerical performance of AdaLDA and ADAM using both simulated and real data.

2.4.1. Simulations

In all the simulations, the sample size is $n_1 = n_2 = 100$ while the dimension p varies from 400, 800 to 1200. The probability of being in either of the two classes is equal. We consider the following six models for the covariance matrix Σ and the discriminating direction $\boldsymbol{\beta}$.

Model 1 Erdős-Rényi random graph: Let $\tilde{\Omega} = (\tilde{\omega}_{ij})$ where $\tilde{\omega}_{ij} = u_{ij}\delta_{ij}$, $\delta_{ij} \sim \text{Ber}(\rho)$ being the Bernoulli random variable with success probability $\rho = 0.2$ and $u_{ij} \sim \text{Unif}[0.5, 1] \cup [-1, -0.5]$. After symmetrizing $\tilde{\Omega}$, set $\Omega = \tilde{\Omega} + \{\max(-\phi_{\min}(\tilde{\Omega}), 0) + 0.05\}I_p$ to ensure the positive definiteness. Finally, Ω is standardized to have unit diagonals and $\Sigma = \Omega^{-1}$. The discriminating direction $\boldsymbol{\beta} = (5/\sqrt{s}, \dots, 5/\sqrt{s}, 0, \dots, 0)^\top$ is sparse such that only the first s entries are nonzero.

Model 2 Block sparse model: $\Omega = (\mathbf{B} + \delta I_p)/(1 + \delta)$ where $b_{ij} = b_{ji} = 10 \times \text{Ber}(0.5)$ for $1 \leq i \leq p/2$, $i < j \leq p$; $b_{ij} = b_{ji} = 10$ for $p/2 + 1 \leq i < j \leq p$; $b_{ii} = 1$ for $1 \leq i \leq p$.

Here $\delta = \max(-\phi_{\min}(\mathbf{B}), 0) + 0.05$. The matrix Ω is also standardized to have unit diagonals and $\Sigma = \Omega^{-1}$. The discriminating direction $\boldsymbol{\beta} = (2/\sqrt{s}, \dots, 2/\sqrt{s}, 0, \dots, 0)^\top$ where the first s entries are nonzero.

Model 3 AR(1) model: $(\Sigma_{ij})_{p \times p}$ with $\Sigma_{ij} = 0.9^{|i-j|}$. The discriminating direction $\boldsymbol{\beta} = (2/\sqrt{s}, \dots, 2/\sqrt{s}, 0, \dots, 0)^\top$ where the first s entries are nonzero.

Model 4 Varying diagonals model: We first let $(\Sigma_{ij})_{p \times p}$ with $\Sigma_{ij} = 0.9^{|i-j|}$. Then we add $\mathbf{d} = (10, 10, 10, 10, 10, U_6, \dots, U_p)$ to the diagonal entries of Σ , where U_6, U_7, \dots, U_p *i.i.d.* $\sim U(0, 1)$. The discriminating direction $\boldsymbol{\beta} = (1/\sqrt{s}, \dots, 1/\sqrt{s}, 0, \dots, 0)^\top$ where the first s entries are nonzero.

Model 5 Approximately sparse $\boldsymbol{\beta}$: Let $(\Sigma_{ij})_{p \times p}$ with $\Sigma_{ij} = 0.9^{|i-j|}$. The discriminating direction $\boldsymbol{\beta} = (0.75, (0.75)^2, (0.75)^3, \dots, (0.75)^p)^\top$ being approximately sparse.

Model 6 Sparse δ and Σ : Let $\Omega = (\mathbf{B} + \delta I_p)/(1 + \delta)$ where $b_{ij} = b_{ji} = 10 \times \text{Ber}(0.2)$ for $1 \leq i \leq p/2, i < j \leq p; b_{ij} = b_{ji} = 10$ for $p/2 + 1 \leq i < j \leq p; b_{ii} = 1$ for $1 \leq i \leq p$. In other words, only the first $p/2$ rows and columns of Ω are sparse, whereas the rest of the matrix is not sparse. Here $\delta = \max(-\phi_{\min}(\mathbf{B}), 0) + 0.05$. The matrix Ω is also standardized to have unit diagonals and $\Sigma = \Omega^{-1}$. The mean difference vector $\boldsymbol{\delta} = (2/\sqrt{s}, \dots, 2/\sqrt{s}, 0, \dots, 0)^\top$ where the first $s = 10$ entries are nonzero. Finally, let $\boldsymbol{\beta} = \Omega \boldsymbol{\delta}$.

Given the covariance matrix Σ and the discriminating direction $\boldsymbol{\beta}$ generated by the model above, the means are $\boldsymbol{\mu}_1 = (0, \dots, 0)^\top$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 - \Sigma \boldsymbol{\beta}$. The missing mechanism is chosen such that each entry X_{ki} is observed with probability $p = 1 - \epsilon \in (0, 1)$. We change the missing proportion ϵ from 0 to 0.2. We apply AdaLDA rule when the data is complete, i.e. $\epsilon = 0$, and apply ADAM rule when $\epsilon > 0$. The AdaLDA rule is then compared with the LPD (Cai and Liu, 2011), SLDA (Shao et al., 2011), FAIR (Fan and Fan, 2008), and NSC (Tibshirani et al., 2002) rules whose tuning parameters are chosen by five-fold cross-validation over the grid $\{\sqrt{\log p/n}, \frac{3}{2}\sqrt{\log p/n}, 2\sqrt{\log p/n}, \dots, 5\sqrt{\log p/n}\}$. We also note

that one commonly used method, the Naive Bayes rule is a special case of the NSC rule with tuning parameter $\lambda_\Delta = 0$, so it's not included in the comparison. In the following tables, the fitting times (in seconds) on a regular computer (Intel Core i7-3770, 3.40 GHz) and misclassification errors (in %) of different algorithms are recorded. The misclassification error of a classifier \hat{C} is computed as

$$\frac{1}{N} \sum_{i=1}^N \mathbb{1}\{\text{label}(\mathbf{Z}_i) \neq \hat{C}(\mathbf{Z}_i)\},$$

where Z_i 's are N fresh samples from the same distribution as the training data. Here we let $N = 200$ and \mathbf{Z}_i are drawn from the two classes with the same probability. Due to different signal-to-noise ratios across different models, the misclassification rates for the Fisher's rule vary in different models. For each setting, the number of repetition is set to 100.

According to the simulation results in Tables 1–5, the proposed AdaLDA algorithm, which is purely data-driven and tuning-free, has a much shorter fitting time than that of LPD, which requires choosing tuning parameters via cross-validation. In addition, due to the element-wise constraints in the optimization, the AdaLDA algorithm adapts to the heteroscedasticity of \mathbf{a} in Lemma 1, and has a better, if not comparable, performance than that of the LPD algorithm with optimally chosen tuning parameters and outperforms all the other methods. This advantage is further demonstrated in Model 4, where the diagonals of covariance matrices Σ vary significantly. According to Table 4, the AdaLDA algorithm has a significant improvement over the LPD rule. Furthermore, we considered simulation settings where β is not sparse. In Table 5, the AdaLDA algorithm still performs well and outperforms all the other methods when β is approximately sparse in Model 5. Under Model 6 where δ and Σ are individually sparse and Σ is diagonally dominant, which is a setting favoring SLDA and FAIR. In this setting, the numerical performance of AdaLDA is not as good as SLDA and FAIR, but the differences are small.

In addition, we also investigate the numerical performance of ADAM for incomplete data. According to Table 6, which shows the performance of ADAM across different missing

Table 1: Misclassification errors (%) and model fitting times for Model 1 with complete data

| (s, p) | AdaLDA | LPD | SLDA | FAIR | NSC | Oracle |
|-----------|-------------------------|---------------------------|-------------------------|-------------------------|--------------------------|-------------|
| (10,400) | 17.50(1.51) [0.58s] | 18.50(0.42) [55.02s] | 42.77(1.89) [2.18s] | 30.42(1.34) [1.40s] | 34.05(1.21) [42.56s] | 12.60(0.51) |
| (20,400) | 19.73(0.54) [0.58s] | 20.65(0.72) [49.73s] | 41.60(2.10) [2.50s] | 25.92(0.71) [2.96s] | 26.87(0.75) [43.54s] | 11.05(0.61) |
| (10,800) | 20.15(1.24) [3.39s] | 25.37(1.62) [187.03s] | 41.46(2.14) [8.74s] | 29.55(0.81) [5.15s] | 33.60(1.01) [111.60s] | 15.13(0.42) |
| (20,800) | 28.30(1.07) [3.35s] | 29.10(1.63) [195.59s] | 43.68(2.49) [7.18s] | 31.58(0.97) [5.62s] | 31.62(0.86) [115.15s] | 14.30(0.74) |
| (10,1200) | 26.10(0.73) [9.90s] | 26.32(0.80) [531.74s] | 42.26(2.45) [28.21s] | 31.78(0.75) [21.43s] | 34.73(0.71) [244.57s] | 16.00(0.60) |
| (20,1200) | 32.96(1.72) [9.94s] | 35.70(1.68) [493.31s] | 44.23(2.65) [28.14s] | 37.48(2.31) [26.41s] | 36.67(1.01) [244.28s] | 18.90(0.58) |
| (10,1600) | 24.40(0.52) [21.77s] | 28.44(2.41) [809.22s] | 43.14(3.16) [56.78s] | 32.48(0.89) [34.36s] | 34.55(0.99) [333.84s] | 19.90(0.51) |
| (20,1600) | 26.20(0.71) [21.75s] | 30.87(2.05) [1019.35s] | 44.24(2.49) [54.92s] | 38.52(2.56) [34.64s] | 35.15(0.92) [421.86s] | 17.35(0.39) |

Table 2: Misclassification errors (%) and model fitting times for Model 2 with complete data

| (s, p) | AdaLDA | LPD | SLDA | FAIR | NSC | Oracle |
|-----------|-------------------------|---------------------------|-------------------------|-------------------------|--------------------------|-------------|
| (10,400) | 11.88(0.16) [0.59s] | 12.57(0.15) [66.8s] | 14.05(0.66) [3.12s] | 17.52(0.70) [1.56s] | 17.58(0.78) [38.52s] | 11.35(0.56) |
| (20,400) | 10.53(0.94) [0.62s] | 11.28(0.67) [71.8s] | 12.03(0.43) [1.72s] | 12.28(0.41) [1.25s] | 12.25(0.40) [37.15s] | 7.40(0.45) |
| (10,800) | 13.40(1.01) [3.44s] | 16.60(1.78) [232.64s] | 15.10(0.66) [9.52s] | 18.48(0.72) [5.77s] | 21.98(0.67) [114.5s] | 13.35(0.64) |
| (20,800) | 13.45(0.98) [3.34s] | 16.85(1.75) [245.62s] | 14.48(0.68) [8.58s] | 15.28(0.75) [5.93s] | 16.53(0.68) [111.56s] | 9.85(0.41) |
| (10,1200) | 15.20(0.21) [9.87s] | 17.57(1.04) [577.16s] | 18.20(0.26) [17.16s] | 18.88(0.53) [11.08s] | 21.68(0.73) [243.50s] | 12.93(0.50) |
| (20,1200) | 14.27(0.82) [9.90s] | 15.20(0.87) [600.76s] | 17.40(0.42) [18.72s] | 15.93(0.71) [9.67s] | 17.68(1.01) [245.75s] | 9.72(0.28) |
| (10,1600) | 14.38(0.43) [21.87s] | 15.74(1.01) [1215.31s] | 15.35(0.42) [29.02s] | 16.08(0.73) [18.88s] | 22.07(0.68) [420.08s] | 11.77(0.42) |
| (20,1600) | 15.74(0.65) [21.75s] | 17.60(1.47) [1212.03s] | 16.46(0.53) [30.26s] | 16.97(0.71) [17.16s] | 19.90(0.81) [413.55s] | 12.03(0.30) |

Table 3: Misclassification errors (%) and model fitting times for Model 3 with complete data

| (s, p) | AdaLDA | LPD | SLDA | FAIR | NSC | Oracle |
|-----------|-------------------------|---------------------------|-------------------------|-------------------------|---------------------------|--------------|
| (10,400) | 27.98(0.90) [0.61s] | 29.65(1.12) [79.42s] | 33.77(0.88) [3.43s] | 37.15(1.16) [0.62s] | 30.00(1.00) [38.22s] | 23.12(0.82) |
| (20,400) | 35.17(0.82) [0.62s] | 36.40(0.80) [73.49s] | 40.45(0.71) [2.50s] | 43.73(0.85) [2.65s] | 37.78(0.90) [37.75s] | 29.08(0.93) |
| (10,800) | 28.45(0.80) [3.36s] | 34.75(0.79) [255.77s] | 34.83(0.66) [10.14s] | 42.18(1.10) [5.77s] | 31.87(0.67) [112.40s] | 22.50(0.50) |
| (20,800) | 34.25 (0.68) [3.39s] | 41.52 (0.67) [250.27s] | 41.85 (1.32) [9.52s] | 44.98 (1.46) [5.15s] | 39.13 (0.68) [113.40s] | 29.83 (0.49) |
| (10,1200) | 28.23(0.59) [9.81s] | 34.10(1.05) [903.35s] | 34.53(0.83) [24.18s] | 41.08(0.87) [14.98s] | 30.15(0.56) [250.96s] | 21.15(0.72) |
| (20,1200) | 34.65(0.95) [9.93s] | 41.68(1.33) [929.33s] | 41.53(1.48) [24.03s] | 46.45(1.05) [16.0s] | 38.07(1.42) [256.52s] | 28.05(1.15) |
| (10,1600) | 27.88(0.89) [21.71s] | 34.13(1.02) [1321.28s] | 35.57(1.23) [27.92s] | 41.32(1.12) [17.78s] | 30.73(0.78) [414.58s] | 22.17(0.58) |
| (20,1600) | 33.45(0.72) [21.06s] | 37.82(0.50) [1864.71s] | 41.80(1.52) [41.03s] | 46.05(1.13) [28.70s] | 38.70(0.95) [478.12s] | 27.65(0.42) |

Table 4: Misclassification errors (%) and model fitting times for Model 4 with complete data

| (s, p) | AdaLDA | LPD | SLDA | FAIR | NSC | Oracle |
|-----------|-------------------------|---------------------------|-------------------------|-------------------------|--------------------------|-------------|
| (10,400) | 10.80(0.45) [0.59s] | 16.03(0.57) [77.95s] | 23.87(1.04) [5.30s] | 16.27(0.58) [2.34s] | 12.33(0.77) [38.02s] | 8.18 (0.40) |
| (20,400) | 16.42(0.68) [0.60s] | 24.05(0.62) [78.06s] | 32.27(0.90) [4.68s] | 23.20(1.01) [2.62s] | 20.50(0.80) [37.75s] | 11.22(0.59) |
| (10,800) | 12.03(0.57) [3.41s] | 21.17(0.70) [252.72s] | 29.00(0.81) [7.33s] | 23.95(0.40) [6.08s] | 17.20(1.00) [111.58s] | 10.05(0.33) |
| (20,800) | 18.48(0.85) [3.35s] | 25.40(0.75) [249.07s] | 36.20(0.71) [9.67s] | 26.28(0.69) [5.15s] | 26.22(0.53) [111.34] | 11.20(0.72) |
| (10,1200) | 13.98(0.58) [9.89s] | 21.90(0.82) [630.88s] | 28.63(0.49) [15.91s] | 25.52(0.48) [12.32s] | 17.40(0.70) [244.53s] | 9.27(0.42) |
| (20,1200) | 21.70(0.90) [9.95s] | 29.77(0.46) [631.24s] | 34.33(0.64) [14.82s] | 28.40(0.88) [9.67s] | 25.82(0.72) [245.78s] | 10.70(0.61) |
| (10,1600) | 13.90(0.42) [21.29s] | 22.07(0.65) [1846.61s] | 27.05(0.71) [40.56s] | 28.73(0.90) [27.77s] | 20.20(0.37) [480.78s] | 11.67(0.49) |
| (20,1600) | 23.95(1.03) [21.56s] | 29.57(0.92) [1215.36s] | 34.60(0.78) [27.46s] | 28.98(1.09) [17.47s] | 23.05(0.45) [414.57s] | 15.90(0.82) |

Table 5: Misclassification errors (%) and model fitting times for Model 5 (the first four rows) and 6 (the last four rows) with complete data

| Method | AdaLDA | LPD | SLDA | FAIR | NSC | Oracle |
|------------|-------------------------|---------------------------|-------------------------|-------------------------|--------------------------|-------------|
| $p = 400$ | 27.98(0.66) [0.61s] | 31.68(0.46) [84.19s] | 46.72(0.83) [4.28s] | 34.68(0.61) [2.03s] | 31.90(0.60) [39.23s] | 18.55(0.56) |
| $p = 800$ | 29.50(0.73) [3.37s] | 34.15(1.03) [253.53s] | 46.30(0.64) [10.76s] | 37.20(0.87) [7.49s] | 33.63(0.96) [112.38s] | 19.35(0.73) |
| $p = 1200$ | 27.43(1.25) [10.00s] | 36.40(0.78) [642.98s] | 46.68(0.82) [15.76s] | 39.20(0.74) [12.01s] | 33.45(0.72) [248.86s] | 18.53(0.88) |
| $p = 1600$ | 27.38(0.64) [21.67s] | 37.82(0.65) [1314.69s] | 47.30(0.80) [26.83s] | 41.38(1.12) [17.94s] | 35.98(1.11) [424.20s] | 20.62(0.68) |
| $p = 400$ | 18.27(0.74) [0.60s] | 28.52(0.70) [56.94s] | 20.60(0.64) [4.84s] | 18.98(0.68) [4.62s] | 21.12(0.78) [37.66s] | 4.77(0.25) |
| $p = 800$ | 19.10(0.99) [3.34s] | 26.68(2.06) [202.69s] | 21.30(0.89) [7.18s] | 18.82(0.73) [5.62s] | 21.10(0.61) [111.74s] | 4.30(0.23) |
| $p = 1200$ | 18.50(0.83) [9.93s] | 25.60(2.12) [514.57s] | 18.90(0.60) [15.60s] | 17.25(0.52) [11.39s] | 20.18(0.73) [244.46s] | 4.25(0.33) |
| $p = 1600$ | 18.23(1.08) [21.65s] | 24.65(1.94) [1105.70s] | 18.80(0.68) [27.46s] | 17.32(0.60) [17.32s] | 22.95(0.76) [420.25s] | 4.62(0.24) |

Table 6: Misclassification errors (%) and model fitting times for Model 1 with missing proportion ϵ

| Method $(s, p) \setminus \epsilon$ | ADAM | | | | AdaLDA |
|---------------------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| | 0.2 | 0.15 | 0.1 | 0.05 | 0 |
| (10,400) | 20.55(0.91) [0.68s] | 19.70(1.11) [0.65s] | 19.20(1.34) [0.66s] | 18.40(0.78) [0.65s] | 17.50(1.51) [0.58s] |
| (20,400) | 24.95(1.34) [0.67s] | 23.60(0.26) [0.66s] | 23.67(0.10) [0.64s] | 21.28(0.91) [0.64s] | 19.73(0.54) [0.58s] |
| (10,800) | 26.28(1.10) [3.60s] | 25.30(0.72) [3.60s] | 22.15(0.74) [3.59s] | 21.18(1.08) [3.59s] | 20.15(1.24) [3.39s] |
| (20,800) | 34.00(1.02) [3.63s] | 33.55(1.56) [3.61s] | 32.97(1.14) [3.64s] | 31.90(0.95) [3.60s] | 28.30(1.07) [3.35s] |
| (10,1200) | 30.62(1.12) [10.56s] | 30.39(1.08) [10.73s] | 29.65(1.72) [10.56s] | 27.42(1.16) [10.53s] | 26.10(0.73) [9.90s] |
| (20,1200) | 35.62(0.81) [10.51s] | 34.10(1.62) [10.54s] | 33.95(0.91) [10.52s] | 33.77(1.04) [10.52s] | 32.96(1.72) [9.94s] |
| (10,1600) | 33.47(1.59) [23.01s] | 30.53(0.79) [22.94s] | 27.40(1.61) [23.11s] | 26.40(1.52) [23.12s] | 24.40(0.52) [21.77s] |
| (20,1600) | 37.40(0.91) [23.00s] | 33.79(0.74) [22.95s] | 32.70(1.12) [23.01s] | 31.77(1.01) [22.96s] | 26.20(0.71) [21.75s] |

proportions ϵ under Model 1, ADAM does not lose much accuracy in the presence of missing data when the missing proportion ϵ is small. As expected, the misclassification errors of ADAM grows when ϵ increases. Since the pattern of the performances of ADAM are similar across different models, the simulation results of ADAM under Models 2-6 are given in the supplementary material.

2.4.2. Real data analysis

In addition to the simulation studies, we also illustrate the merits of the AdaLDA and ADAM classifiers in an analysis of two real datasets to further investigate the numerical performance of the proposed methods. One dataset, available at www.chestsurg.org, is the Lung cancer data analyzed by Gordon et al. (2002). Another dataset is the Leukemia data from high-density Affymetrix oligonucleotide arrays that was previously analyzed in Golub

Table 7: Classification error of Lung cancer data by various methods

| | ADAM _($\epsilon=0.1$) | ADAM _($\epsilon=0.05$) | AdaLDA | LPD | SLDA | FAIR | NSC |
|---------------|---|--|--------|-------|-------|-------|-------|
| Testing error | 5.53% | 3.22% | 2.09% | 2.11% | 4.88% | 3.64% | 7.30% |

et al. (1999), and is available at www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. These two datasets were frequently used for illustrating the empirical performance of the classifier for high-dimensional data in recent literature. We will compare AdaLDA and ADAM with the existing methods.

Lung cancer data

We evaluate the proposed methods by classifying between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There are 181 tissue samples (31 MPM and 150 ADCA) and each sample is described by 12533 genes in the lung cancer dataset in Gordon et al. (2002). This dataset has been analyzed in Fan and Fan (2008) using FAIR and NSC. In this section we apply the AdaLDA and ADAM rules to this dataset for disease classification. When ADAM rule is used, we make each entry in the dataset missing uniformly and independently with probability ϵ . In the simulation, given the small sample size, we choose $\epsilon = 0.05$ and $\epsilon = 0.1$.

The sample variances of the genes range over a wide interval. We first compute the sample variances for each gene and drop the lower and upper 6-quantiles to control the condition number of $\hat{\Sigma}$. The average misclassification errors are computed by using 5-fold cross-validation for various methods with 50 repetitions. To reduce the computational costs, in each repetition, only 1500 genes with the largest absolute values of the two sample t statistics are used. We then apply all the aforementioned methods to this reduced dimensional dataset. As seen in the Table 7, the classification result of AdaLDA is better than existing methods, including LPD (Cai and Liu, 2011), SLDA (Shao et al., 2011), FAIR (Fan and Fan, 2008), and NSC (Tibshirani et al., 2002) methods, although only 1500 genes were used. Moreover, in the incomplete data case, ADAM still has satisfactory accuracy.

Table 8: Classification error of Leukemia data by various methods

| | ADAM _($\epsilon=0.1$) | ADAM _($\epsilon=0.05$) | AdaLDA | LPD | SLDA | FAIR | NSC |
|---------------|---|--|--------|-------|-------|-------|-------|
| Testing error | 8.47% | 7.53% | 2.94% | 3.09% | 5.76% | 2.94% | 8.82% |

Leukemia data

Golub et al. (1999) applied gene expression microarray techniques to study human acute leukemia and discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). There are 72 tissue samples (47 ALL and 25 AML) and 7129 genes in the Leukemia dataset. In this section, we apply the AdaLDA rule to this dataset and compare the classification results with those obtained by LPD (Cai and Liu, 2011), SLDA (Shao et al., 2011), FAIR (Fan and Fan, 2008), and NSC (Fan and Fan, 2008) methods. Same as the analysis of lung cancer data, when ADAM rule is used, we make each entry in the dataset missing independently with probability $\epsilon \in \{0.05, 0.1\}$

As in the analysis of the lung cancer data, we first drop genes with extreme sample variances out of lower and upper 6-quantiles. Similar to the analysis of the lung cancer data, the average misclassification errors are computed by using two-fold cross-validation for various methods with 50 repetitions, and to control the computational costs, we use 2000 genes with the largest absolute values of the two sample t statistics in each repetition. After the application of all methods to the same reduced dimensional dataset, classification results are then summarized in Table 8. The AdaLDA has the similar performance as the LPD rule and FAIR, as obtain the misclassification error of about 3%. In contrast, the naive-Bayes rule misclassifies 20.59% testing samples and SLDA misclassifies 5.76% testing samples. Fan and Fan (2008) report a test error rate of 2.94% for FAIR and a test error rate of 8.82% for NSC proposed by Tibshirani et al. (2002). In the presence of missing data, ADAM misclassifies 7.53% and 8.47% testing samples when the missing proportion is 0.05 and 0.1 respectively.

2.5. Extension to multiple-class LDA

We have so far focused on the two-class high-dimensional LDA. The procedure can be extended to the following K -class setting:

$$X_1^{(k)}, \dots, X_{n_k}^{(k)} \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_k, \Sigma), \text{ for } k = 1, \dots, K.$$

For ease of presentation, we focus on the complete data case in this section. For a future observation \mathbf{Z} drawn from these K distributions with prior probabilities π_1, \dots, π_K , the oracle classification rule is given by

$$C_{\boldsymbol{\theta}}(\mathbf{Z}) = \arg \max_{k \in [K]} D_k, \quad (2.17)$$

where $D_1 = 0$ and $D_k = (\mathbf{Z} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_K}{2})^\top \boldsymbol{\beta}_k + \log(\frac{\pi_k}{\pi_1})$ for $k = 2, \dots, K$, with $\boldsymbol{\beta}_k = \Omega(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$.

A similar data-driven adaptive classifier, called K -class AdaLDA, can then be constructed based on the estimation of $\boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_K$ and $\Delta_k = \sqrt{\boldsymbol{\beta}_k^\top \Sigma \boldsymbol{\beta}_k}$, as follows.

Let $\hat{\boldsymbol{\mu}}_k, k \in [K]$ and $\hat{\Sigma}$ be the sample means and pooled sample covariance matrix respectively.

Step 1 (Estimating Δ_k^2). Fix $\lambda_0 = 25/2$. For $k = 2, \dots, K$, we estimate $\boldsymbol{\beta}_k$ by a preliminary estimator

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_k &= \arg \min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \\ &\text{subject to } |e_j^\top (\hat{\Sigma} \boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1))| \leq 4 \sqrt{\frac{\log p}{n}} \cdot \sqrt{\hat{\sigma}_{jj}} \cdot (\lambda_0 \boldsymbol{\beta}^\top (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1) + 1), \quad j \in [p]. \end{aligned} \quad (2.18)$$

Then we estimate Δ_k^2 by $\hat{\Delta}_k^2 = |\tilde{\boldsymbol{\beta}}_k^\top (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1)|$, $k = 2, \dots, K$.

Step 2 (Adaptive estimation of β_k). Given $\hat{\Delta}^2$, the final estimator $\hat{\beta}_k$ of β_k is constructed through the following linear optimization

$$\begin{aligned} \hat{\beta}_k &= \arg \min_{\beta} \|\beta\|_1 \\ \text{subject to } |e_j^\top (\hat{\Sigma}\beta - (\hat{\mu}_k - \hat{\mu}_1))| &\leq 4\sqrt{\frac{\log p}{n}} \cdot \sqrt{\hat{\sigma}_{jj}(\lambda_0 \hat{\Delta}_k^2 + 1)}, \quad j \in [p]. \end{aligned} \quad (2.19)$$

Step 3 (Construction of K -class AdaLDA). The K -class AdaLDA classification rule is obtained by plugging $\hat{\beta}_k$ into Fisher's rule (2.17),

$$\hat{C}_{K\text{-AdaLDA}}(\mathbf{Z}) = \arg \max_{k \in [K]} \hat{D}_k, \quad (2.20)$$

where $\hat{D}_1 = 0$ and $\hat{D}_k = (\mathbf{Z} - \frac{\hat{\mu}_1 + \hat{\mu}_K}{2})^\top \hat{\beta}_k + \log(\hat{\pi}_k / \hat{\pi}_1)$ with $\hat{\pi}_k = n_k / \sum_{j=1}^K n_j$ for $k = 2, \dots, K$.

For theoretical analysis, we consider the following parameter space $\mathcal{G}_K(s, M_{n,p})$ defined by

$$\begin{aligned} \mathcal{G}_K(s, M_{n,p}) &= \{\boldsymbol{\theta} = (\pi_1, \dots, \pi_K, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \Sigma) : \boldsymbol{\mu}_k \in \mathbb{R}^p, \pi_k \in (c, 1-c), \sum_{k=1}^K \pi_k = 1, \Sigma \in \mathbb{R}^{p \times p}, \Sigma \succ 0, \\ &\|\beta_k\|_0 \leq s, M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M, M_{n,p} \leq \Delta \leq 3M_{n,p}\}, \end{aligned}$$

where $M > 1$ and $c \in (0, 1/2)$ are some constants, $M_{n,p} > 0$ can potentially grow with n and p .

Theoretical properties of K -class AdaLDA can be established by applying the same technical argument as before.

Theorem 9. *Consider the parameter space $\mathcal{G}_K(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$. Suppose $\mathbf{X}_1^{(k)}, \dots, \mathbf{X}_{n_k}^{(k)}$ *i.i.d.* $N_p(\boldsymbol{\mu}_k, \Sigma)$ for $k = 1, \dots, K$. Assume that $M_{n,p} \sqrt{\frac{s \log p}{n}} = o(1)$.*

Then for $k \in [K]$,

$$\sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{E}[\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_2] \lesssim M_{n,p} \sqrt{\frac{s \log p}{n}}.$$

The following theorem provides an upper bound for the excess misclassification risk $R_{\boldsymbol{\theta}}(\hat{C}_{K\text{-AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta})$ of the K -class AdaLDA rule.

Theorem 10. *Consider the parameter space $\mathcal{G}_K(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$ and assume the conditions in Theorem 2 hold.*

1. *If $M_{n,p} \leq C_b$ for some $C_b > 0$, then there exists some constant $C > 0$,*

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}_K(s, M_{n,p})} \mathbb{P} \left(R_{\boldsymbol{\theta}}(\hat{C}_{K\text{-AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta}) \leq C \cdot \frac{s \log p}{n} \right) \geq 1 - 8p^{-1}.$$

2. *If $M_{n,p} \rightarrow \infty$ as $n \rightarrow \infty$, then there exist some constant $C > 0$ and $\delta_n = o(1)$, such that*

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}_K(s, M_{n,p})} \mathbb{P} \left(R_{\boldsymbol{\theta}}(\hat{C}_{K\text{-AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta}) \leq C \cdot e^{-(\frac{1}{8} + \delta_n) M_{n,p}^2} \cdot \frac{s \log p}{n} \right) \geq 1 - 8p^{-1}.$$

2.6. Proofs

In this section, we prove the main results, Theorem 2, 3, 4 5, 6 and 7. Theorem 1 follows from Theorems 3 and 7. Since $n_1 \asymp n_2$, without loss of the generality we shall assume $n_1 = n_2 = n$ in the proofs. For reasons of space, the proofs of the technical lemmas are given in the Supplementary Material (Cai and Zhang, 2018d).

2.6.1. Proof of Theorem 2

To prove Theorem 2 we begin by collecting a few important technical lemmas that will be used in the main proofs.

Auxiliary Lemmas

Lemma 5. Suppose $\mathbf{X}_1, \dots, \mathbf{X}_n$ i.i.d. $\sim N_p(\boldsymbol{\mu}, \Sigma)$, and assume that $\hat{\boldsymbol{\mu}}, \hat{\Sigma}$ are the sample mean and sample covariance matrix respectively. Let $\Gamma(s) = \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}_{S^c}\|_1 \leq \|\mathbf{u}_S\|_1, \text{ for some } S \subset [p] \text{ with } |S| = s\}$, then with probability at least $1 - p^{-1}$,

$$\sup_{\mathbf{u} \in \Gamma(s)} \mathbf{u}^\top (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \lesssim \sqrt{\frac{s \log p}{n}};$$

$$\sup_{\mathbf{u}, \mathbf{v} \in \Gamma(s)} \mathbf{u}^\top (\hat{\Sigma} - \Sigma) \mathbf{v} \lesssim \sqrt{\frac{s \log p}{n}}.$$

Lemma 6. Suppose $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. Let $\mathbf{h} = \mathbf{x} - \mathbf{y}$ and $S = \text{supp}(\mathbf{y})$. If $\|\mathbf{x}\|_1 \leq \|\mathbf{y}\|_1$, then $h \in \Gamma(s)$ with $s = |S|$, that is,

$$\|\mathbf{h}_{S^c}\|_1 \leq \|\mathbf{h}_S\|_1.$$

Main proof of Theorem 2

Recall that $\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}$ is constructed by the following two steps.

Step 1. Estimating Δ^2

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \left\{ |e_j^\top (\hat{\Sigma} \boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1))| \leq 4 \sqrt{\frac{\log p}{n}} \cdot \sqrt{\hat{\sigma}_{jj}} \cdot (\lambda_0 \boldsymbol{\beta}^\top (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) + 1), j \in [p] \right\}. \quad (2.21)$$

Then we estimate Δ^2 by $\hat{\Delta}^2 = |\langle \tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1 \rangle|$.

Step 2. Adaptive estimation of $\boldsymbol{\beta}$. Given $\hat{\Delta}^2$, the final estimator $\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}$ of $\boldsymbol{\beta}$ is constructed by the following linear optimization problem

$$\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} = \arg \min_{\boldsymbol{\beta}} \left\{ |e_j^\top (\hat{\Sigma} \boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1))| \leq 4 \sqrt{\frac{\log p}{n}} \cdot \sqrt{\lambda_0 \hat{\sigma}_{jj} \hat{\Delta}^2 + \hat{\sigma}_{jj}}, j \in [p] \right\}. \quad (2.22)$$

Firstly, let's show the consistency of estimating Δ^2 . Recall the definition of $\tilde{\boldsymbol{\beta}}$ and using

Lemma 5, we have with high probability at least $1 - 3p^{-1}$,

$$\begin{aligned}
|(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})| &\leq |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\Sigma}\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\delta}})| + |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\Sigma} - \Sigma)\tilde{\boldsymbol{\beta}}| + |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\boldsymbol{\delta} - \hat{\boldsymbol{\delta}})| \\
&\leq \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \|\hat{\Sigma}\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\delta}}\|_\infty + |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\Sigma} - \Sigma)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})| + |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\Sigma} - \Sigma)\boldsymbol{\beta}| \\
&\quad + |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\boldsymbol{\delta} - \hat{\boldsymbol{\delta}})| \\
&\lesssim \sqrt{s} \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \cdot \|\hat{\Sigma}\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\delta}}\|_\infty + \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \cdot \sqrt{\frac{s \log p}{n}} \cdot \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_2 \\
&\quad + \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_2 \sqrt{\frac{s \log p}{n}} \cdot \|\boldsymbol{\beta}\|_2 + \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_2 \sqrt{\frac{s \log p}{n}},
\end{aligned} \tag{2.23}$$

where the third inequality uses Lemma 5 and the fact that $\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \in \Gamma(s)$. In fact, $\tilde{\boldsymbol{\beta}}$ is a feasible solution to (2.8) due to Lemma 1 and thus $\|\tilde{\boldsymbol{\beta}}\|_1 \leq \|\boldsymbol{\beta}\|_1$. Then by Lemma 7, we have $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \in \Gamma(s)$. In addition, $\|\boldsymbol{\beta}\|_0 \leq s$, so we have $\boldsymbol{\beta} \in \Gamma(s)$.

In addition, by standard derivation of the accuracy of sample variance, since $M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M$, by using the union bound technique, we have with probability at least $1 - p^{-1}$,

$$\max_{i \in [p]} |\hat{\sigma}_{ii} - \sigma_{ii}| \lesssim \sqrt{\frac{\log p}{n}},$$

which implies with probability at least $1 - p^{-1}$,

$$\max_{i \in [p]} |\hat{\sigma}_{ii}| \leq 2M.$$

In addition, since $\Delta \geq M_{n,p} \geq c_L > 0$, then with probability at least $1 - 3p^{-1}$,

$$\begin{aligned}
\|\hat{\Sigma}\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\delta}}\|_\infty &\leq 4\sqrt{\frac{\log p}{n}} \cdot \sqrt{\hat{\sigma}_{jj}} \cdot (\lambda_0\tilde{\boldsymbol{\beta}}^\top(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) + 1) \\
&\lesssim \sqrt{\frac{\log p}{n}} \cdot |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) + 1| + \sqrt{\frac{\log p}{n}} \cdot |\boldsymbol{\beta}^\top(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)| \\
&\leq \sqrt{\frac{\log p}{n}} \cdot (|(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)| + |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)| + 1) \\
&\quad + \sqrt{\frac{\log p}{n}} \cdot (|\boldsymbol{\beta}^\top(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)| + |\boldsymbol{\beta}^\top(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1)|) \\
&\lesssim \sqrt{\frac{\log p}{n}}\Delta\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \sqrt{s} \cdot \frac{\log p}{n}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \sqrt{\frac{\log p}{n}}\Delta^2 + \sqrt{s} \cdot \frac{\log p}{n}\Delta,
\end{aligned}$$

where the last inequality uses the fact that $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_2, \|\boldsymbol{\beta}\|_2 \lesssim \Delta$, since $\Delta = \sqrt{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \Omega (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)} \geq \frac{1}{\sqrt{M}}\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_2$, and $\Delta = \sqrt{\boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta}} \geq \frac{1}{\sqrt{M}}\|\boldsymbol{\beta}\|_2$.

It follows that with probability at least $1 - 6p^{-1}$,

$$\begin{aligned}
|(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})| &\lesssim \sqrt{\frac{s \log p}{n}}\Delta\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 + \frac{s \log p}{n}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 + \sqrt{\frac{s \log p}{n}}\Delta^2\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \\
&\quad + \frac{s \log p}{n}\Delta\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \sqrt{\frac{s \log p}{n}} \cdot \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \\
&\quad + \sqrt{\frac{s \log p}{n}}\Delta\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \cdot \sqrt{\frac{s \log p}{n}} \\
&\lesssim \sqrt{\frac{s \log p}{n}}\Delta\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 + \sqrt{\frac{s \log p}{n}}\Delta^2\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2,
\end{aligned}$$

where the last inequality uses the fact that $\Delta \geq M_{n,p} \geq c_L > 0$.

On the other hand, since

$$|(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})| \geq \lambda_{\min}(\Sigma)\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \geq \frac{1}{M}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2.$$

We then have, with probability at least $1 - 6p^{-1}$,

$$\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \lesssim \sqrt{\frac{s \log p}{n}} \left(\Delta\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \Delta^2 \right),$$

Assuming $M_{n,p}\sqrt{\frac{s \log p}{n}} = o(1)$, which implies $\Delta\sqrt{\frac{s \log p}{n}} = o(1)$, then we have

$$\|\tilde{\beta} - \beta\|_2 \lesssim \frac{\Delta^2 \sqrt{\frac{s \log p}{n}}}{1 - \Delta\sqrt{\frac{s \log p}{n}}}.$$

Since $\|\tilde{\beta}\|_1 \leq \|\beta\|_1$ and combining with Lemma 5, we then have with probability at least $1 - 7p^{-1}$,

$$\begin{aligned} \left| \frac{\hat{\Delta}^2 - \Delta^2}{\Delta^2} \right| &\leq \frac{|\tilde{\beta}^\top(\delta - \hat{\delta})| + |\delta^\top(\tilde{\beta} - \beta)|}{\Delta^2} \leq \frac{\|\beta\|_1 \cdot \|\delta - \hat{\delta}\|_\infty + \|\delta\|_2 \cdot \|\beta - \tilde{\beta}\|_2}{\Delta^2} \\ &\leq \frac{\sqrt{s} \cdot \|\beta\|_2 \cdot \|\delta - \hat{\delta}\|_\infty + \|\delta\|_2 \cdot \|\beta - \tilde{\beta}\|_2}{\Delta^2} \\ &\lesssim \frac{\sqrt{s} \cdot \Delta\sqrt{\frac{\log p}{n}} + \Delta \cdot \frac{\Delta^2 \sqrt{\frac{s \log p}{n}}}{1 - \Delta\sqrt{\frac{s \log p}{n}}}}{\Delta^2} = o(1), \end{aligned}$$

given $\Delta \geq c_L$ and $\Delta\sqrt{\frac{s \log p}{n}} = o(1)$.

Secondly, let's proceed to showing the accuracy of $\hat{\beta}_{\text{AdaLDA}}$. We use $\hat{\beta}$ to denote $\hat{\beta}_{\text{AdaLDA}}$ in this subsection for simplicity. By Lemma 1, β lies in the feasible set of (2.9), so $\|\hat{\beta}\|_1 \leq \|\beta\|_1$. By a similar argument as in (2.23), we have that with probability at least $1 - 3p^{-1}$,

$$\begin{aligned} &|(\hat{\beta} - \beta)^\top \Sigma(\hat{\beta} - \beta)| \\ &\leq |(\hat{\beta} - \beta)^\top (\hat{\Sigma}\hat{\beta} - \hat{\delta})| + |(\hat{\beta} - \beta)^\top (\hat{\Sigma} - \Sigma)\hat{\beta}| + |(\hat{\beta} - \beta)^\top (\delta - \hat{\delta})| \\ &\lesssim \sqrt{s} \|\hat{\beta} - \beta\|_2 \cdot \|\hat{\Sigma}\hat{\beta} - \hat{\delta}\|_\infty + \|\hat{\beta} - \beta\|_2 \cdot \sqrt{\frac{s \log p}{n}} \cdot \|\beta - \hat{\beta}\|_2 \\ &\quad + \|\beta - \hat{\beta}\|_2 \sqrt{\frac{s \log p}{n}} \cdot \|\beta\|_2 + \|\beta - \hat{\beta}\|_2 \sqrt{\frac{s \log p}{n}}. \end{aligned} \tag{2.24}$$

Now since we have $|\frac{\hat{\Delta}^2 - \Delta^2}{\Delta^2}| = o(1)$ with probability at least $1 - 7p^{-1}$, this implies with probability at least $1 - 10p^{-1}$,

$$\|\hat{\Sigma}\hat{\beta} - \hat{\delta}\|_\infty \leq \sqrt{\frac{\log p}{n}} \cdot \sqrt{\hat{\sigma}_{jj}\hat{\Delta}^2 + 2\hat{\sigma}_{jj}} \lesssim \Delta\sqrt{\frac{\log p}{n}}.$$

Then using the fact $|(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})| \geq \lambda_{\min}(\Sigma) \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$ again, we have with probability at least $1 - 10p^{-1}$,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \lesssim \Delta \sqrt{\frac{s \log p}{n}} \cdot \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \sqrt{\frac{s \log p}{n}} \cdot \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2.$$

This implies that there exists some constant $C > 0$, such that with probability at least $1 - 10p^{-1}$,

$$\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} - \boldsymbol{\beta}\|_2 \leq C\Delta \cdot \sqrt{\frac{s \log p}{n}}.$$

In addition, since $\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}\|_1 \leq \|\boldsymbol{\beta}\|_1 \leq \sqrt{p} \|\boldsymbol{\beta}\|_2 \leq \sqrt{pM} \cdot \Delta$, we then have

$$\begin{aligned} & \mathbb{E}[\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} - \boldsymbol{\beta}\|_2] \\ & \leq \mathbb{E}[\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} - \boldsymbol{\beta}\|_2 \cdot 1_{\{\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} - \boldsymbol{\beta}\|_2 > C\Delta \cdot \sqrt{\frac{s \log p}{n}}\}}] + \mathbb{E}[\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} - \boldsymbol{\beta}\|_2 \cdot 1_{\{\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} - \boldsymbol{\beta}\|_2 \leq C\Delta \cdot \sqrt{\frac{s \log p}{n}}\}}] \\ & \leq \sqrt{pM} \cdot \Delta \cdot 10p^{-1} + C\Delta \cdot \sqrt{\frac{s \log p}{n}} \lesssim \Delta \cdot \sqrt{\frac{s \log p}{n}} \lesssim M_{n,p} \cdot \sqrt{\frac{s \log p}{n}}. \end{aligned}$$

2.6.2. Proofs of Theorem 3

For a vector $\boldsymbol{x} \in \mathbb{R}^p$, we define $\|\boldsymbol{x}\|_{2,s} = \sup_{\|\boldsymbol{y}\|_2=1, \boldsymbol{y} \in \Gamma(s)} |\boldsymbol{x}^\top \boldsymbol{y}|$. We start with the following lemma.

Lemma 7. *For two vectors $\boldsymbol{\gamma}$ and $\hat{\boldsymbol{\gamma}}$, if $\|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\|_2 = o(1)$ as $n \rightarrow \infty$, and $\|\boldsymbol{\gamma}\|_2 \geq c$ for some constant $c > 0$, then when $n \rightarrow \infty$,*

$$\|\boldsymbol{\gamma}\|_2 \cdot \|\hat{\boldsymbol{\gamma}}\|_2 - \boldsymbol{\gamma}^\top \hat{\boldsymbol{\gamma}} \asymp \|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\|_2^2.$$

We postpone the proof of Lemma 14 to Section A.6 in the supplement, and continue the proof of Theorem 3.

Let $\delta_n = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \vee \|\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1\|_{2,s} \vee \|\hat{\boldsymbol{\mu}}_2 - \boldsymbol{\mu}_2\|_{2,s}$. We are going to show

$$R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \lesssim e^{-\Delta^2/8} \cdot \Delta \cdot \delta_n^2.$$

Given the estimators $\hat{\omega}$, $\hat{\boldsymbol{\mu}}_k$, and $\hat{\boldsymbol{\beta}}$, the sample \mathbf{Z} is classified as

$$\hat{C}(\mathbf{Z}) = \begin{cases} 1, & (\mathbf{Z} - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2)^\top \hat{\boldsymbol{\beta}} \geq 0 \\ 2, & (\mathbf{Z} - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2)^\top \hat{\boldsymbol{\beta}} < 0. \end{cases}$$

Let $\hat{\Delta} = \sqrt{\hat{\boldsymbol{\beta}}^\top \Sigma \hat{\boldsymbol{\beta}}}$ and $\hat{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}$. The misclassification error is

$$R_{\boldsymbol{\theta}}(\hat{C}) = \frac{1}{2} \Phi\left(-\frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}}\right) + \frac{1}{2} \bar{\Phi}\left(-\frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}}\right),$$

with $R_{\text{opt}}(\boldsymbol{\theta}) = \frac{1}{2} \Phi(-\Delta/2) + \frac{1}{2} \bar{\Phi}(\Delta/2)$. Define an intermediate quantity

$$R^* = \frac{1}{2} \Phi\left(-\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right) + \frac{1}{2} \bar{\Phi}\left(\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right).$$

We first show that $R^* - R_{\text{opt}}(\boldsymbol{\theta}) \lesssim e^{-\Delta^2/8} \cdot \Delta \cdot \delta_n^2$. Applying Taylor's expansion to the two terms in R^* at $\frac{\Delta}{2}$ and $-\frac{\Delta}{2}$ respectively, we obtain

$$R^* - R_{\text{opt}}(\boldsymbol{\theta}) = \frac{1}{2} \left(\frac{\Delta}{2} - \frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}}\right) \Phi'\left(\frac{\Delta}{2}\right) + \frac{1}{2} \left(-\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} + \frac{\Delta}{2}\right) \Phi'\left(-\frac{\Delta}{2}\right) + O\left(e^{-\Delta^2/8} \frac{1}{\Delta} \cdot \delta_n^4\right), \quad (2.25)$$

In fact, the remaining term can be written as

$$\frac{1}{2} \left(\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\Delta}{2}\right)^2 \Phi''(t_{1,n}) + \left(\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\Delta}{2}\right)^2 \Phi''(t_{2,n}),$$

where $t_{1,n}, t_{2,n}$ are some constants satisfying $|t_{1,n}|, |t_{2,n}|$ are between $\frac{\Delta}{2}$ and $\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}$.

Therefore, the remaining term can be bounded by using the facts that

$$\left| \frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\Delta}{2} \right| = O\left(\frac{1}{\Delta} \delta_n^2\right), \text{ and } \Phi''(t_n) = O(e^{-\Delta^2/8} \Delta),$$

for $|t_n|$ is between $\frac{\Delta}{2}$ and $\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\Delta}$.

In fact, for the first term, we can obtain this inequality by letting $\boldsymbol{\gamma} = \Sigma^{1/2} \boldsymbol{\beta}$ and $\hat{\boldsymbol{\gamma}} = \Sigma^{1/2} \hat{\boldsymbol{\beta}}$ in Lemma 14. Then

$$\left| \Delta - \frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} \right| = \left| \|\boldsymbol{\gamma}\|_2 - \frac{\boldsymbol{\gamma}^\top \hat{\boldsymbol{\gamma}}}{\|\hat{\boldsymbol{\gamma}}\|_2} \right| = \left| \frac{\|\boldsymbol{\gamma}\|_2 \|\hat{\boldsymbol{\gamma}}\|_2 - \boldsymbol{\gamma}^\top \hat{\boldsymbol{\gamma}}}{\|\hat{\boldsymbol{\gamma}}\|_2} \right| \lesssim \frac{1}{\Delta} \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2^2 \lesssim \frac{1}{\Delta} \delta_n^2.$$

In addition, since as $\delta_n \rightarrow 0$, $\frac{(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \rightarrow \frac{\Delta}{2}$, we then have $|\Phi''(t_n)| \asymp \Delta \cdot e^{-\frac{(\Delta/2)^2}{2}} = \Delta \cdot e^{-\Delta^2/8}$.

Then (4.30) can be further expanded such that

$$\begin{aligned} R^* - R_{\text{opt}}(\boldsymbol{\theta}) &\asymp \left(-\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} + \frac{\Delta}{2} \right) e^{-\frac{1}{2} \left(\frac{\Delta}{2} \right)^2} + \left(-\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} + \frac{\Delta}{2} \right) e^{-\frac{1}{2} \left(-\frac{\Delta}{2} \right)^2} + O\left(e^{-\Delta^2/8} \frac{1}{\Delta} \cdot \delta_n^4 \right) \\ &= \exp\left(-\frac{\Delta^2}{8} \right) \cdot \left(-\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} + \Delta \right) + O\left(e^{-\Delta^2/8} \frac{1}{\Delta} \cdot \delta_n^4 \right) \\ &\lesssim e^{-\Delta^2/8} \cdot \left| \frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} - \Delta \right| + O\left(e^{-\Delta^2/8} \frac{1}{\Delta} \cdot \delta_n^4 \right) \lesssim e^{-\Delta^2/8} \cdot \delta_n^2. \end{aligned}$$

Eventually we obtain $R^* - R_{\text{opt}}(\boldsymbol{\theta}) \lesssim e^{-\Delta^2/8} \Delta \cdot \delta_n^2$.

To upper bound $R_{\boldsymbol{\theta}}(\hat{C}) - R^*$, applying Taylor's expansion to $R_{\boldsymbol{\theta}}(\hat{C})$,

$$\begin{aligned} R_{\boldsymbol{\theta}}(\hat{C}) &= \frac{1}{2} \left\{ \Phi\left(\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right) + \frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\beta}} - \boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \Phi'\left(\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right) + O\left(e^{-\Delta^2/8} \Delta \cdot \delta_n^2 \right) \right\} \\ &\quad - \frac{1}{2} \left\{ \bar{\Phi}\left(\frac{-\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right) + \frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^\top \hat{\boldsymbol{\beta}} + \boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \Phi'\left(-\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right) + O\left(e^{-\Delta^2/8} \Delta \cdot \delta_n^2 \right) \right\}, \end{aligned}$$

where the remaining term can be obtained similarly as (4.30) by using the fact

$$\left| \frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\beta}} - \boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right| = O(\delta_n) \text{ and } |\Phi''(\cdot)| = O(e^{-\Delta^2/8} \Delta).$$

In fact, when $|\hat{\Delta} - \Delta| \leq |\sqrt{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\Sigma(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}| \lesssim \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \lesssim \delta_n = o(1)$, we have

$$\left| \frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\beta}} - \boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right| \leq \frac{1}{2\Delta} |(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_2 + \boldsymbol{\mu}_2)^\top \hat{\boldsymbol{\beta}}| \lesssim \delta_n.$$

This leads to

$$\begin{aligned} |R_{\boldsymbol{\theta}}(\hat{C}) - R^*| &\lesssim \left| \frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} \Phi' \left(\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right) \right. \\ &\quad \left. + \frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2 + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} \Phi' \left(-\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right) + O \left(e^{-\Delta^2/8} \Delta \cdot \delta_n^2 \right) \right| \\ &= \left| \frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} e^{-\frac{1}{2} \left\{ \frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right\}^2} \right. \\ &\quad \left. + \frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2 + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} e^{-\frac{1}{2} \left\{ \frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right\}^2} + O \left(e^{-\Delta^2/8} \Delta \cdot \delta_n^2 \right) \right|. \end{aligned}$$

Since

$$\boldsymbol{\delta}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1) + \boldsymbol{\delta}/2 + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2) = \boldsymbol{\delta} - (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = 0,$$

then it follows that

$$|R_{\boldsymbol{\theta}}(\hat{C}) - R^*| \lesssim e^{-\Delta^2/8} \Delta \cdot \delta_n^2.$$

Combining the pieces, we obtain

$$R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \lesssim e^{-\Delta^2/8} \cdot \Delta \cdot \delta_n^2.$$

Finally, by Lemma 5 and the derivation in Theorem 2, with probability at least $1 - 12p^{-1}$, $\delta_n \lesssim M_{n,p} \sqrt{\frac{s \log p}{n}}$. In addition, $\Delta \in [M_{n,p}, 3M_{n,p}]$, we then have with probability at least $1 - 12p^{-1}$,

$$R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \lesssim e^{-M_{n,p}^2/8} \cdot M_{n,p}^3 \cdot \frac{s \log p}{n}.$$

Now we consider the two cases. On the one hand, when $M_{n,p}$ is bounded by C_b , we have

$$R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \lesssim e^{-M_{n,p}^2/8} \cdot \frac{s \log p}{n}.$$

On the other hand, when $M_{n,p} \rightarrow \infty$ as n grows,

$$R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \lesssim e^{-\left(\frac{1}{8} - \frac{3 \log M_{n,p}}{M_{n,p}^2}\right) M_{n,p}^2} \cdot M_{n,p}^3 \cdot \frac{s \log p}{n},$$

where $\frac{3 \log M_{n,p}}{M_{n,p}^2}$ is an $o(1)$ term as $n \rightarrow \infty$.

2.6.3. Proofs of Theorems 4 and 5

We proceed to proving Theorems 4 and 5 under the event $\{c_1 n_0 \leq n_{\min}^*(S) \leq c_2 n_0\}$ that happens with probability at least $1 - p^{-1}$. The results then rely on the following lemma.

Lemma 8. *Consider the MCR model and assume that $\hat{\boldsymbol{\mu}}, \hat{\Sigma}$ are the generalized sample mean and sample covariance matrix respectively. If $c_1 n_0 \leq n_{\min}^*(S) \leq c_2 n_0$. then with probability at least $1 - p^{-1}$,*

$$\sup_{\mathbf{u} \in \Gamma(s)} \mathbf{u}^\top (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \lesssim \sqrt{\frac{s \log p}{n_0}};$$

$$\sup_{\mathbf{u}, \mathbf{v} \in \Gamma(s)} \mathbf{u}^\top (\hat{\Sigma} - \Sigma) \mathbf{v} \lesssim \sqrt{\frac{s \log p}{n_0}}.$$

Given Lemma 8, the derivation of Theorems 4 is very similar to the case with AdaLDA in Section 2.6.1, and 5 can be derived from Theorem 4 by using the same logic as in Section 2.6.2, and thus are omitted.

2.6.4. Proofs of the minimax lower bound results (Theorems 6 and 7)

In this section we are going to prove Theorems 6 and 7. We start with providing lemmas that will be used in the proof.

Auxiliary lemmas

The proof of Theorem 6 relies on the following Fano's Lemma.

Lemma 9 (Tsybakov (2009)). *Suppose Θ_p is a parameter space consisting of M parameters $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_M \in \Theta_p$ for some $M > 0$, and $d(\cdot, \cdot) : \Theta_p \times \Theta_p \rightarrow \mathbb{R}^+$ is some distance. Denote $\mathbb{P}_\boldsymbol{\theta}$ to be some probability measure parametrized by $\boldsymbol{\theta}$. If for some constants $\alpha \in (0, 1/8), \gamma > 0$, $KL(\mathbb{P}_{\boldsymbol{\theta}_i}, \mathbb{P}_{\boldsymbol{\theta}_0}) \leq \alpha \log M/n$ for all $1 \leq i \leq M$, and $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \geq \gamma$ for all $0 \leq i \neq j \leq M$, then*

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{i \in [M]} \mathbb{E}_{\boldsymbol{\theta}_i} [d_{\boldsymbol{\theta}_i}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_i)] \gtrsim \gamma.$$

The proof of Theorem 7, however, is not straightforward, since the excess risk $R_{\boldsymbol{\theta}}(\hat{C}) - R_{opt}(\boldsymbol{\theta})$ is not a distance as required in Lemma 9. The key step in our proof of Theorem 7 is to reduce the excess risk $R_{\boldsymbol{\theta}}(\hat{C}) - R_{opt}(\boldsymbol{\theta})$ to $L_{\boldsymbol{\theta}}(\hat{C})$, defined in (2.16). The following lemma suggests that it suffices to provide a lower bound for $L_{\boldsymbol{\theta}}(\hat{C})$, and $L_{\boldsymbol{\theta}}(\hat{C})$ satisfies an approximate triangle inequality (Lemma 4).

Although $L_{\boldsymbol{\theta}}(\hat{C})$ is not a distance function and does not satisfy an exact triangle inequality, the following lemma provides a variant of Fano's Lemma.

Lemma 10 (Tsybakov (2009)). *Let $M \geq 0$ and $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_M \in \Theta_p$. For some constants $\alpha_0 \in (0, 1/8], \gamma > 0$, and any classifier \hat{C} , if $KL(\mathbb{P}_{\boldsymbol{\theta}_i}, \mathbb{P}_{\boldsymbol{\theta}_0}) \leq \alpha_0 \log M/n$ for all $1 \leq i \leq M$, and $L_{\boldsymbol{\theta}_i}(\hat{C}) < \gamma$ implies $L_{\boldsymbol{\theta}_j}(\hat{C}) \geq \gamma$ for all $0 \leq i \neq j \leq M$, then*

$$\inf_{\hat{C}} \sup_{i \in [M]} \mathbb{P}_{\boldsymbol{\theta}_i}(L_{\boldsymbol{\theta}_i}(\hat{C}) \geq \gamma) \geq \frac{\sqrt{M}}{\sqrt{M} + 1} (1 - 2\alpha_0 - \sqrt{\frac{2\alpha_0}{\log M}}).$$

Lemma 11 (Tsybakov (2009)). *Define $\mathcal{A}_{p,s} = \{\mathbf{u} : \mathbf{u} \in \{0, 1\}^p, \|\mathbf{u}\|_0 \leq s\}$. If $p \geq 4s$, then there exists a subset $\{\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_M\} \subset \mathcal{A}_{p,s}$ such that $\mathbf{u}_0 = \{0, \dots, 0\}^\top$, $\rho_H(\mathbf{u}_i, \mathbf{u}_j) \geq s/2$ and $\log(M + 1) \geq \frac{s}{3} \log(\frac{p}{s})$, where ρ_H denotes the Hamming distance.*

Proof of Theorem 6

In this section we prove the lower bound of estimation of β . First we construct a subset of the parameter space Θ that characterizes the hardness of the problem. By Lemma 16, there exist $\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_M \in \mathcal{A}_{p,s} = \{\mathbf{u} \in \{0, 1\}^p : \|\mathbf{u}\|_0 \leq s\}$, such that $\rho_H(\mathbf{u}_i, \mathbf{u}_j) > s/2$ and $\log(M+1) \geq \frac{s}{5} \log(\frac{p}{s})$, denote this collection of \mathbf{u}_i by $\tilde{\mathcal{A}}_{p,s}$. In addition, denote $\mathbf{u}_0 = \mathbf{0}_p$.

Since $\frac{\log p}{\log(p/s)} = O(1)$, so for sufficiently large p , we have $s < p/2$. Define \mathbf{b}_0 be the p -dimensional vector with the last s entries being $\frac{M_{n,p}}{\sqrt{s}}$ and the rest being 0, so we have $\|\mathbf{b}_0\|_2 = M_{n,p}$. Let $r = \lceil p/2 \rceil$. For $\mathbf{u} \in \tilde{\mathcal{A}}_{p,s} = \{\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_M\}$, let $B_{\mathbf{u}}$ be the $p \times p$ symmetric matrix whose i -th row and column are both $\epsilon \cdot u_i \cdot \frac{\mathbf{b}_0}{M_{n,p}}$ for $i \in \{1, \dots, r\}$, where ϵ is to be determined later. The parameter set we considered is

$$\Theta_0 = \{\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1 = \mathbf{b}_0, \boldsymbol{\mu}_2 = -\mathbf{b}_0, \Sigma = (I_p + B_{\mathbf{u}})^{-1}; \mathbf{u} \in \tilde{\mathcal{A}}_{p,s} \cup \{\mathbf{0}_p\}\}.$$

For a given \mathbf{u} , the corresponding discriminating direction is $\beta_{\mathbf{u}} = -2(I_p + B_{\mathbf{u}})\mathbf{b}_0$, which implies

$$\|\beta_{\mathbf{u}} - \beta_{\tilde{\mathbf{u}}}\|_2^2 = 4\|(B_{\mathbf{u}} - B_{\tilde{\mathbf{u}}})\mathbf{b}_0\|_2^2 \geq 4\rho_H(\mathbf{u}, \tilde{\mathbf{u}})\epsilon^2\|\mathbf{b}_0\|_2^2 \geq 2sM_{n,p}^2\epsilon^2.$$

In addition, when $\|B_{\mathbf{u}}\|_2 = o(1)$, for sufficiently large n , we have $\Delta = \sqrt{4\mathbf{b}_0^\top (I_p + B_{\mathbf{u}})\mathbf{b}_0} \in (M_{n,p}, 3M_{n,p})$, which implies that $\Theta_0 \subset \mathcal{G}(s, M_{n,p})$.

We then proceed to bound $KL(\mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}_i}}, \mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}_0}})$ for $i \in [M]$, where $\mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}_i}}, \mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}_0}}$ denote the distributions $N_p(\mathbf{b}_0, (I_p + B_{\mathbf{u}_i})^{-1})$ and $N_p(\mathbf{b}_0, I_p)$ respectively. We then have

$$KL(\mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}_i}}, \mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}_0}}) = \frac{1}{2} [-\log |I_p + B_{\mathbf{u}_i}| - p + \text{tr}(I_p + B_{\mathbf{u}_i})].$$

Note that $\frac{\mathbf{b}_0}{M_{n,p}}$ is a unit vector. If we take ϵ such that $\|B_{\mathbf{u}_i}\|_2 \leq \|B_{\mathbf{u}}\|_F \leq \sqrt{2s \cdot \epsilon^2} = o(1)$, and denote the eigenvalues of $I_p + B_{\mathbf{u}_i}$ by $1 + \Delta_{\lambda_1}, \dots, 1 + \Delta_{\lambda_p}$ with $\Delta_{\lambda_j} = o(1)$. We then

have

$$\begin{aligned} KL(\mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}_i}}, \mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}_0}}) &= \frac{1}{2} \left[- \sum_{j=1}^p \log(1 + \Delta_{\lambda_j}) - p + \sum_{j=1}^p (1 + \Delta_{\lambda_j}) \right] \\ &\asymp \frac{1}{4} \sum_{j=1}^p \Delta_{\lambda_j}^2 = \frac{1}{4} \|B_{\mathbf{u}}\|_F^2 \leq \frac{1}{2} s \epsilon^2 \end{aligned}$$

where we use the fact that $\log(1+x) \asymp x - \frac{x^2}{2}$ when $x = o(1)$. Now let $\epsilon = \frac{1}{5\sqrt{2}} \sqrt{\frac{\log p}{n}}$, then $KL(\mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}_i}}, \mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}_0}}) \leq \alpha \log M/n$ for $\alpha = 1/8$.

In addition, let $\gamma = \frac{1}{10} M_{n,p} \sqrt{\frac{s \log p}{n}}$, then for $0 \leq i \neq j \leq M$ and any $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$, such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{u}_i}\|_2 \leq \gamma$, we have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{u}_j}\|_2 \geq \|\boldsymbol{\beta}_{\mathbf{u}_j} - \boldsymbol{\beta}_{\mathbf{u}_i}\|_2 - \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{u}_i}\|_2 \geq \frac{1}{5} M_{n,p} \sqrt{\frac{s \log p}{n}} - \frac{1}{10} M_{n,p} \sqrt{\frac{s \log p}{n}} = \frac{1}{10} M_{n,p} \sqrt{\frac{s \log p}{n}} = \gamma.$$

Then by Fano's lemma (Lemma 9), we have $\inf_{\hat{\boldsymbol{\beta}}} \sup_{i \in [M]} \mathbb{E} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{u}_i}\|_2 \gtrsim M_{n,p} \sqrt{\frac{s \log p}{n}}$.

For the incomplete data case with $n_0 \geq 1$, we consider a special pattern of missingness S_0 :

$$(S_0)_{ij} = 1_{\{1 \leq i \leq n_0, 1 \leq j \leq p\}} \quad \text{with probability 1.}$$

Under this missingness pattern, $n_{\min}^* = n_0$ with probability 1, and the problem essentially becomes complete data problem with n_0 samples, which implies

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathbf{S} \in \Psi(n_0; n, p)}} \mathbb{E} [\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2] \gtrsim M_{n,p} \sqrt{\frac{s \log p}{n_0}}.$$

Proof of Theorem 7

We proceed by applying Lemma 15 to obtain the minimax lower bound for the excess misclassification error. We first construct a subset of the parameter space Θ that characterizes

the hardness of the problem. Let \mathbf{e}_1 be the basis vector in the standard Euclidean space whose first entry is 1 and zero elsewhere. By Lemma 16, there exist $\mathbf{u}_1, \dots, \mathbf{u}_M \in \check{\mathcal{A}}_{p,s} = \{\mathbf{u} \in \{0, 1\}^p : \mathbf{u}^\top \mathbf{e}_1 = 0, \|\mathbf{u}\|_0 = s\}$, such that $\rho_H(\mathbf{u}_i, \mathbf{u}_j) > s/2$ and $\log(M+1) \geq \frac{s}{5} \log(\frac{p-1}{s})$. Note the first entry in \mathbf{u}_j is 0 for all $j = 1, \dots, M$.

Define the parameter space

$$\Theta_1 = \{\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1 = \epsilon \mathbf{u} + \lambda \mathbf{e}_1, \boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1, \Sigma = \sigma^2 I_p; \mathbf{u} \in \check{\mathcal{A}}_{p,s}\},$$

where $\epsilon = \sigma \sqrt{\log p/n}$, $\sigma^2 = O(1)$ and λ is chosen to ensure $\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})$ such that

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{4\|\epsilon \mathbf{u} + \lambda \mathbf{e}_1\|_2^2}{\sigma^2} = M_{n,p}.$$

To apply Lemma 15, we need to verify two conditions: (i) the upper bound on the KL divergence between $\mathbb{P}_{\boldsymbol{\theta}_u}$ and $\mathbb{P}_{\boldsymbol{\theta}_v}$, and (ii) the lower bound of $L_{\boldsymbol{\theta}_u}(\hat{C}) + L_{\boldsymbol{\theta}_v}(\hat{C})$ for $\mathbf{u} \neq \mathbf{v} \in \check{\mathcal{A}}_{p,s}$.

We calculate the KL divergence first. For $\mathbf{u} \in \check{\mathcal{A}}_{p,s}$, denote $\boldsymbol{\mu}_u = \epsilon \mathbf{u} + \lambda \mathbf{e}_1$. For $\boldsymbol{\theta}_u = (\boldsymbol{\mu}_u, -\boldsymbol{\mu}_u, \sigma^2 I_p) \in \Theta_1$, we consider the distribution $N_p(\boldsymbol{\mu}_u, \sigma^2 I_p)$.

Then, the KL divergence between $\mathbb{P}_{\boldsymbol{\theta}_u}$ and $\mathbb{P}_{\boldsymbol{\theta}_v}$ can be bounded by

$$\text{KL}(\mathbb{P}_{\boldsymbol{\theta}_u}, \mathbb{P}_{\boldsymbol{\theta}_v}) \leq \frac{1}{2} \|\boldsymbol{\mu}_u - \boldsymbol{\mu}_v\|_2^2 \leq \sigma^2 \cdot \frac{s \log p}{n}. \quad (2.26)$$

In addition, by applying Lemma 4, we have that for any $\mathbf{u}, \mathbf{v} \in \check{\mathcal{A}}_{p,s}$,

$$L_{\boldsymbol{\theta}_u}(\hat{C}) + L_{\boldsymbol{\theta}_v}(\hat{C}) \gtrsim \frac{1}{M_{n,p}} e^{-M_{n,p}^2/8} \sqrt{\frac{s \log p}{n}}.$$

So far we have verified the aforementioned conditions (i) and (ii). Lemma 15 immediately

implies that, there is some $C_\alpha \geq 0$, such that

$$\inf_{\hat{C}} \sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P}(L_{\boldsymbol{\theta}}(\hat{C}) \geq C_\alpha \frac{1}{M_{n,p}} e^{-M_{n,p}^2/8} \sqrt{\frac{s \log p}{n}}) \geq 1 - \alpha. \quad (2.27)$$

Finally combining (4.34) with Lemma 12, we obtain the desired lower bound for the excess misclassification error

$$\inf_{\hat{C}} \sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P}(R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \geq C_\alpha \frac{1}{M_{n,p}} e^{-M_{n,p}^2/8} \frac{s \log p}{n}) \geq 1 - \alpha.$$

Under this missingness data case, we consider the same missingness pattern S_0 as described in Section 2.6.4 with $n_{\min} = n_0$. Then we have

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{S} \in \Psi(n_0; n, p)}} \mathbb{P}(R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \geq C_\alpha \frac{1}{M_{n,p}} e^{-M_{n,p}^2/8} \frac{s \log p}{n_0}) \geq 1 - \alpha.$$

This implies that

1. If $M_{n,p} \leq C_b$ for some $C_b > 0$, then

$$\inf_{\hat{C}} \sup_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \Psi(n_0; n, p)}} \mathbb{P}(R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \geq C_\alpha e^{-\frac{1}{8} M_{n,p}^2} \cdot \frac{s \log p}{n_0}) \geq 1 - \alpha.$$

2. If $M_{n,p} \rightarrow \infty$ as $n \rightarrow \infty$, then for any $\delta > 0$,

$$\inf_{\hat{C}} \sup_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \Psi(n_0; n, p)}} \mathbb{P}(R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \geq C_\alpha e^{-(\frac{1}{8} + \delta) M_{n,p}^2} \cdot \frac{s \log p}{n_0}) \geq 1 - \alpha.$$

CHAPTER 3 : A Convex Optimization Approach to High-dimensional Sparse
Quadratic Discriminant Analysis

3.1. Introduction

Discriminant analysis is a commonly used classification technique in statistics and machine learning. It has a wide range of applications, including, for example, face recognition (Wright et al., 2009), text mining (Berry and Castellanos, 2004), business forecasting (Churchill and Iacobucci, 2006) and gene expression analysis (Jombart et al., 2010). In the ideal setting of two known normal distributions $N_p(\boldsymbol{\mu}_1, \Sigma_1)$ (class 1) and $N_p(\boldsymbol{\mu}_2, \Sigma_2)$ (class 2), the goal of the discriminant analysis is to classify a new observation \mathbf{z} , which is drawn from one of the two distributions with prior probabilities π_1 and π_2 respectively, into one of the two classes. In the ideal setting where all the parameters $\boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$ are known, the optimal classifier is the quadratic discriminant rule is given by

$$G_{\boldsymbol{\theta}}^*(\mathbf{z}) = \begin{cases} 1, & (\mathbf{z} - \boldsymbol{\mu}_1)^\top D(\mathbf{z} - \boldsymbol{\mu}_1) - 2\boldsymbol{\delta}^\top \Omega_2(\mathbf{z} - \bar{\boldsymbol{\mu}}) - \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + 2\log\left(\frac{\pi_1}{\pi_2}\right) > 0 \\ 2, & (\mathbf{z} - \boldsymbol{\mu}_1)^\top D(\mathbf{z} - \boldsymbol{\mu}_1) - 2\boldsymbol{\delta}^\top \Omega_2(\mathbf{z} - \bar{\boldsymbol{\mu}}) - \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + 2\log\left(\frac{\pi_1}{\pi_2}\right) \leq 0, \end{cases} \quad (3.1)$$

where $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, $\bar{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}$, and $D = \Omega_2 - \Omega_1$ with $\Omega_i = \Sigma_i^{-1}$ for $i = 1, 2$, see, for example, Anderson (2003). When $\Sigma_1 = \Sigma_2$, the quadratic classification boundary in (3.1) becomes linear, reducing the quadratic discriminant analysis (QDA) to the linear discriminant analysis (LDA).

QDA has been an important technique for classification and is more flexible than the LDA (Hastie et al., 2009). In practice, the parameters $\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1$ and Σ_2 are usually unknown and instead one observes two independent random samples, $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n_1}^{(1)} \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_1, \Sigma_1)$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n_2}^{(2)} \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_2, \Sigma_2)$. It is practically important to construct a data-driven classification rule based on the two samples. In the low-dimensional setting where the dimension p is small relative to the sample sizes, a natural approach is to simply plug the sample means and sample covariance matrices into the oracle QDA rule (3.1).

This approach has been well studied. See, for example, Anderson (2003). Thanks to the explosive growth of big data, high-dimensional data, where the dimension p can be much larger than the sample sizes, are now routinely collected in scientific investigations in a wide range of fields. In such settings, the conventional LDA and QDA rules perform poorly.

For high-dimensional LDA, there already exist a number of proposals and theoretical studies. In particular, assuming sparsity on the discriminating direction, direct estimation methods have been introduced in Cai and Liu (2011) and Mai et al. (2012) and optimality theory is developed in Cai and Zhang (2018a). In contrast, relatively few methods have been introduced for regularized QDA in the high-dimensional setting and developing an optimality theory is technically more challenging. Li and Shao (2015) studied high-dimensional QDA by imposing sparsity assumptions on $\boldsymbol{\delta}$, Σ_1 , Σ_2 and $\Sigma_1 - \Sigma_2$ separately, and then plugging the estimates of these quantities into the oracle QDA rule (3.1). Jiang et al. (2015) introduced a direct estimation approach by assuming that $\Omega_1 - \Omega_2$ and $(\Omega_1 + \Omega_2)\boldsymbol{\delta}$ are sparse, and proposed a consistent classification rule. However, it is unclear whether any of these methods achieves the optimal convergence rate for the classification error.

In the present paper, we propose a sparse QDA rule using convex optimization and aim to establish the optimal convergence rates for the classification error in the high-dimensional settings. It is intuitively clear that QDA is a difficult problem in the high-dimensional setting. For example, it can be seen easily from (3.1) that knowledge of the log-determinant of the covariance matrices $\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)$ is essential for the QDA. However, as shown in Cai et al. (2015), there is no consistent estimator for the log-determinant of the covariance matrices in the high-dimensional setting even when they are known to be diagonal. We begin by establishing rigorously minimax lower bound results, which demonstrate that structural assumptions such as sparsity conditions on the discriminating direction $\boldsymbol{\beta} = \Omega_2\boldsymbol{\delta}$ and differential graph $D = \Omega_2 - \Omega_1$ are necessary for the possible construction of consistent high-dimensional QDA rules. There are two key steps in obtaining the impossibility results: One is the reduction of the classification error to an alternative loss and another is a careful

construction of a collection of least favorable multivariate normal distributions.

We then propose a classification algorithm called SDAR (**S**parse **D**iscriminant **A**nalysis with **R**egularization) to solve the high-dimensional QDA problem under the sparsity assumptions. The SDAR algorithm proceeds by first estimating β and D through constrained convex optimization, and then using the estimators to construct a data-driven classification rule. The first estimation step is in a similar spirit to that in Jiang et al. (2015) by directly estimating the key quantities in the oracle QDA rule. The second classification step is based on a simple but important observation that $\log(|\Sigma_1|/|\Sigma_2|) = \log(|D\Sigma_1 + I_p|)$. As a result, we are able to derive an explicit convergence rate for the classification error of the proposed SDAR algorithm. In addition, we establish a matching minimax lower bound, up to a logarithm factor, that shows the near-optimality of the classifier. Both simulations and real data analysis are carried out to study the numerical performance of the proposed algorithm. The results show that the proposed SDAR algorithm outperforms existing methods in the literature. The methodology and theory developed for high-dimensional QDA for two groups in the Gaussian setting are also extended to multi-group classification and to classification under the Gaussian copula model.

The contributions of the present paper are three-fold. Firstly, we address the necessity of structural assumptions on the parameters for the high-dimensional QDA problem by observing that consistent classification is impossible unless $p = o(n)$ without any such assumptions. Secondly, under the sparsity assumptions, we proposed the SDAR rule, and established an explicit convergence rate of classification error. To the best of our knowledge, this is the first explicit convergence rate for high-dimensional QDA. Lastly, we provide a minimax lower bound, which shows that the convergence rate obtained by the SDAR rule is optimal, up to a logarithmic factor.

The rest of the paper is organized as follows. In Section 3.2, minimax lower bounds are established to show the necessity of imposing structural assumptions for high-dimensional QDA. Section 3.3 presents in detail the data-driven classification procedure SDAR. Theo-

retical properties of SDAR are investigated in Section 3.4 under certain sparsity conditions. The upper and lower bounds together show that the SDAR rule achieves the optimal rate for the classification error up to a logarithmic factor. Simulation studies are given in Section 3.5 where we compare the performance of the proposed algorithm to other existing classification methods in the literature. In addition, the merits of the SDAR classifier are illustrated through an analysis of a prostate cancer dataset and a colon tissue dataset. Section 3.6 discusses extensions to multi-group classification and to classification under the Gaussian copula model. The proofs of main results are given in Section 3.7, and proofs of other results are provided in the supplement.

Notation and definitions

We first introduce basic notation and definitions that will be used throughout the rest of the paper. For an event A , $\mathbb{1}\{A\}$ is the indicator function on A . For an integer $m \geq 1$, $[m]$ denotes the set $\{1, 2, \dots, m\}$. Throughout the paper, vectors are denoted by boldface letters. For a vector \mathbf{u} , $\|\mathbf{u}\|$, $\|\mathbf{u}\|_1$, $\|\mathbf{u}\|_\infty$ denotes the ℓ_2 norm, ℓ_1 norm, and ℓ_∞ norm respectively. We use $\text{supp}(\mathbf{u})$ to denote the support of the vector \mathbf{u} . $\mathbf{0}_p$ is a p -dimensional vector with elements being 0, and $\mathbf{1}_p$ is a p -dimensional vector with elements being 1. For $i \in [p]$, \mathbf{e}_i is the i -th standard basis. For a matrix $M \in \mathbb{R}^{p \times p}$, $\|M\|$, $\|M\|_F$, $\|M\|_1$ denote the spectral norm, Frobenius norm, and matrix l_1 norm respectively. In addition, $|M|_1 = \sum_{i,j} |M_{i,j}|$, $|M|_\infty = \max_{i,j} |M_{i,j}|$, and $|M|$ is the determinant of M . Let $\lambda_i(M)$ denote the i -th eigenvalue of M with $\lambda_1(M) \geq \dots \geq \lambda_p(M)$. Let $M \succ 0$ denote M to be a positive semidefinite matrix and I_p is the $p \times p$ identity matrix. In addition, $M_1 \otimes M_2$ denotes the Kronecker product and $\text{vec}(M)$ is the $p^2 \times 1$ vector obtained by stacking the columns of M . $\text{diag}(M)$ is the linear operator that sets all the off diagonal elements of M to 0. $E_{i,i}$ is a $p \times p$ matrix whose (i, i) -th entry is 1 and 0 else. For a positive integer $s < p$, let $\Gamma(s; p) = \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}_{S^c}\|_1 \leq \|\mathbf{u}_S\|_1, \text{ for some } S \subset [p] \text{ with } |S| = s\}$, where \mathbf{u}_S denotes the subvector of \mathbf{u} confined to S . For two sequences of positive numbers a_n and b_n , $a_n \lesssim b_n$ means that for some constant $c > 0$, $a_n \leq c \cdot b_n$ for all n , and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and

$b_n \lesssim a_n$. $a_n \ll b_n$ means that $\lim_{n \rightarrow \infty} |a_n|/|b_n| = 0$. In our asymptotic framework, we let n be the driving asymptotic parameter, s and p approach infinity as n grows to infinity. We also use $c, c_1, c_2, \dots, C, C_1, C_2$ to denote constants that does not depend on n, p , and their values may vary from place to place.

3.2. The Difficulties of High-dimensional QDA

As mentioned in the introduction, high-dimensional QDA is a difficult problem. In this section, we establish explicit minimax lower bounds that show the necessity of structural assumptions on the discriminating direction $\boldsymbol{\beta} = \Omega_2 \boldsymbol{\delta}$ and differential graph $D = \Omega_2 - \Omega_1$ for constructing consistent high-dimensional QDA rules.

3.2.1. The setup

Suppose we have random samples collected from

$\pi_1 N_p(\boldsymbol{\mu}_1, \Sigma_1) + \pi_2 N_p(\boldsymbol{\mu}_2, \Sigma_2)$, among which n_1 samples belong to class 1: $\mathbf{x}_1, \dots, \mathbf{x}_{n_1} \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_1, \Sigma_1)$, and n_2 samples are in class 2: $\mathbf{y}_1, \dots, \mathbf{y}_{n_2} \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_2, \Sigma_2)$. The goal is to construct a classification rule \hat{G} , which is a function of \mathbf{x}_i 's and \mathbf{y}_i 's, to classify a future data point $\mathbf{z} \sim \pi_1 N_p(\boldsymbol{\mu}_1, \Sigma_1) + \pi_2 N_p(\boldsymbol{\mu}_2, \Sigma_2)$. This model is parametrized by $\boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$. Let $n = \min\{n_1, n_2\}$. For any classification rule $\hat{G} : \mathbb{R}^p \rightarrow \{1, 2\}$, the accuracy is measured by the classification error

$$R_{\boldsymbol{\theta}}(\hat{G}) = \mathbb{E}_{\boldsymbol{\theta}}[\mathbb{1}\{\hat{G}(\mathbf{z}) \neq L(\mathbf{z})\}], \quad (3.2)$$

where $L(\mathbf{z})$ denotes the true class label of \mathbf{z} , that is, $L(\mathbf{z}) = 1$ if $\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)$, and 2 otherwise.

When $\boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$ is known in advance, the oracle classification rule in (3.1) is the Bayes rule and achieves the the minimal classification error, see Anderson (2003). For

ease of presentation, let us define the discriminant function by

$$Q(\mathbf{z}; \boldsymbol{\theta}) = (\mathbf{z} - \boldsymbol{\mu}_1)^\top D(\mathbf{z} - \boldsymbol{\mu}_1) - 2\boldsymbol{\delta}^\top \Omega_2(\mathbf{z} - \bar{\boldsymbol{\mu}}) - \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + 2\log\left(\frac{\pi_1}{\pi_2}\right). \quad (3.3)$$

Then $Q(\mathbf{z}; \boldsymbol{\theta}) = 0$ characterizes the classification boundary of the oracle QDA rule, and (3.1) can be rewritten as

$$G_{\boldsymbol{\theta}}^*(\mathbf{z}) = 1 + \mathbb{1}\{Q(\mathbf{z}; \boldsymbol{\theta}) \leq 0\},$$

and $R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}^*) = \min_{G \in \mathcal{G}} R_{\boldsymbol{\theta}}(G)$, where \mathcal{G} is the set of all classification rules.

In the following the Bayes classification risk $R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}^*)$ is used as the benchmark and the excess risk $R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}^*)$ is used to evaluate the performance of a data-driven classification rule \hat{G} . We say \hat{G} is consistent, or $G_{\boldsymbol{\theta}}^*$ can be mimicked by \hat{G} , if the excess risk $R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}^*) \rightarrow 0$ as the sample size $n \rightarrow \infty$.

3.2.2. Impossibility of QDA in high dimensions

We now characterize the fundamental limits of QDA by showing that, without structural assumptions, $G_{\boldsymbol{\theta}}^*$ cannot be mimicked unless $p \ll n$, which precludes the framework in the high-dimensional settings that motivates our study.

We first consider the simple case where $\Sigma_1 = \Sigma_2 = \Sigma$, and in which case the QDA is reduced to the LDA problem. Under the LDA model in the high-dimensional regime, Bickel and Levina (2004) and Cai et al. (2019a) proposed consistent classification rules under stringent structural conditions on $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$. In this paper, we demonstrate the necessity of these structural assumptions by showing that without structural assumptions, a consistent classification rule is impossible in the high-dimensional LDA problem.

We firstly consider the parameter space

$$\Theta_p^{(1)} = \{\boldsymbol{\theta} = (1/2, 1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, I_p, I_p) : \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p, c_1 \leq \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| \leq c_2\},$$

for some constant $c_1, c_2 > 0$.

Theorem 1. *Suppose that \hat{G} is any classification rule constructed based on the observations $\mathbf{x}_1, \dots, \mathbf{x}_n \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_1, I_p)$, $\mathbf{y}_1, \dots, \mathbf{y}_n \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_2, I_p)$ with $\boldsymbol{\theta} = (1/2, 1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, I_p, I_p) \in \Theta_p^{(1)}$, then when n is sufficiently large,*

$$\inf_{\hat{G}} \sup_{\boldsymbol{\theta} \in \Theta_p^{(1)}} \mathbb{E} \left[R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}^*) \right] \gtrsim \frac{p}{n} \wedge 1.$$

This theorem implies that even when the covariance matrices are equal and known to be identity matrices, as long as the mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ are unknown, no data-driven method is able to mimic $G_{\boldsymbol{\theta}}^*$ in the high dimensional setting where $p \gtrsim n$. Structural assumptions are $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are necessary for a consistent classification rule.

However, for high-dimensional QDA, structural assumptions on $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are not enough and more assumptions are needed. To this end, we consider another scenario where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are known exactly. Let $\boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^* \in \mathbb{R}^p$ be two given vectors and define the parameter space

$$\Theta_p^{(2)}(\boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*) = \{\boldsymbol{\theta} = (1/2, 1/2, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*, \Sigma_1, \Sigma_2) : \Sigma_1, \Sigma_2 \text{ are diagonal matrices}\}.$$

Theorem 2. *Suppose \hat{G} is constructed based on the observations $\mathbf{x}_1, \dots, \mathbf{x}_n \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_1, \Sigma_1)$, $\mathbf{y}_1, \dots, \mathbf{y}_n \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_2, \Sigma_2)$. For any given $\boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^* \in \mathbb{R}^p$ with $\|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_2 \leq C$ where $C > 0$ is some constant, when $\boldsymbol{\theta} = (1/2, 1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2) \in \Theta_p^{(2)}(\boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*)$, we have for sufficiently large n ,*

$$\inf_{\hat{G}} \sup_{\boldsymbol{\theta} \in \Theta_p^{(2)}(\boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*)} \mathbb{E} \left[R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}^*) \right] \gtrsim \frac{p}{n} \wedge 1.$$

This theorem implies that even if we have the prior information that $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ are known and Σ_1, Σ_2 are both diagonal, the quadratic discriminant rule $G_{\boldsymbol{\theta}}^*$ cannot be mimicked consistently if $p \gtrsim n$. The construction of consistent classification rules requires stronger assumptions.

The main strategy of these proofs are discussed in Section 3.4.2, and the detailed proofs of these lower bound results is provided in Section 3.7.1. In addition, the lower bounds are tight, up to a logarithmic factor. Specifically, by using the techniques similar to that in Theorem 4, the plug-in classification rule \hat{G} , which is obtained by plugging in sample means and sample covariance matrices in (3.1), satisfies that $R_{\theta}(\hat{G}) - R_{\theta}(G_{\theta}^*) \lesssim \frac{p \log^2 n}{n} \wedge 1$. This result is further discussed in the supplement.

3.3. Sparse Quadratic Discriminant Analysis

The inconsistency results in Theorems 1 and 2 imply the necessity of imposing structural assumptions on both the mean vectors and covariance matrices. In this section, we consider the QDA problem under the assumptions that the discriminating direction $\beta = \Omega_2 \delta$ and the differential graph D are both sparse. This sparsity assumption, according to (3.3), implies that the classification boundary of the oracle rule depends only on a small number of features in \mathbf{z} . It is also worth noting that the differential graph D corresponds to the change of interactions in two different graphs Ω_1 and Ω_2 . The problem of interaction selection is important in its own right and has been studied extensively recently in dynamic network analysis under various environmental and experimental conditions, see Bandyopadhyay et al. (2010); Zhao et al. (2014); Xia et al. (2015); Hill et al. (2016).

To see that these two sparsity assumptions are sufficient to obtain a consistent estimator for the optimal classification rule G_{θ}^* , we begin by rewriting $Q(\mathbf{z}; \theta)$, defined in (3.3). Recall that $\delta = \mu_2 - \mu_1$, $\bar{\mu} = \frac{\mu_1 + \mu_2}{2}$, $D = \Omega_2 - \Omega_1$ and $\beta = \Omega_2 \delta$, then

$$\begin{aligned} Q(\mathbf{z}; \theta) &= (\mathbf{z} - \mu_1)^{\top} D (\mathbf{z} - \mu_1) - 2\beta^{\top} (\mathbf{z} - \bar{\mu}) - \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + 2 \log\left(\frac{\pi_1}{\pi_2}\right) \\ &= (\mathbf{z} - \mu_1)^{\top} D (\mathbf{z} - \mu_1) - 2\beta^{\top} (\mathbf{z} - \bar{\mu}) - \log(|D\Sigma_1 + I_p|) + 2 \log\left(\frac{\pi_1}{\pi_2}\right). \end{aligned} \quad (3.4)$$

A simple but essential observation of (3.4) is that the first three quantities in the above oracle QDA rule G_{θ}^* depends on either D or β , and the fourth term $\log(\pi_1/\pi_2)$ is easy to

estimate. In the present paper, we shall show that under the sparsity assumptions on these two quantities, D and $\boldsymbol{\beta}$ can be estimated directly and efficiently, and the classification rule based on these two estimates enjoys desirable theoretical guarantees.

Remark 1. By symmetry, $Q(\mathbf{z}; \boldsymbol{\theta})$ can also be rewritten in a form that depends on $(\Omega_1 + \Omega_2)\boldsymbol{\delta}$ and D . The reason that we consider $(\Omega_2\boldsymbol{\delta}, D)$ as the key quantity is that this could be easily extended to the case with K multiple groups. In this generalized setting, we consider using the first group as a benchmark, and computing the likelihood ratio of other groups versus the first one. As a result, the key quantity in the multiple classification case is $\{(\Omega_k(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1), \Omega_k - \Omega_1)\}_{k=2}^K$. See more discussion in Section 3.6.

In the following, we proceed to estimate D and $\boldsymbol{\beta}$ through constrained convex optimization. Let the first sample covariance matrix be $\hat{\Sigma}_1 = n_1^{-1} \sum_{i=1}^{n_1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)^\top$, where $\hat{\boldsymbol{\mu}}_1 = n_1^{-1} \sum_{i=1}^{n_1} \mathbf{x}_i$ and define $\hat{\Sigma}_2$ and $\hat{\boldsymbol{\mu}}_2$ similarly. Since D satisfies the equation $\Sigma_1 D \Sigma_2 = \Sigma_1 - \Sigma_2$ and $\Sigma_2 D \Sigma_1 = \Sigma_1 - \Sigma_2$, a sensible estimation procedure is to solve $\hat{\Sigma}_1 D \hat{\Sigma}_2 / 2 + \hat{\Sigma}_2 D \hat{\Sigma}_1 / 2 - \hat{\Sigma}_1 + \hat{\Sigma}_2 = 0$ for D . We estimate D through the following constrained ℓ_1 minimization approach

$$\hat{D} = \arg \min_{D \in \mathbb{R}^{p \times p}} \left\{ |D|_1 : \left| \frac{1}{2} \hat{\Sigma}_1 D \hat{\Sigma}_2 + \frac{1}{2} \hat{\Sigma}_2 D \hat{\Sigma}_1 - \hat{\Sigma}_1 + \hat{\Sigma}_2 \right|_\infty \leq \lambda_{1,n} \right\}, \quad (3.5)$$

where $\lambda_{1,n} = c_1 \sqrt{\frac{\log p}{n}}$ is a tuning parameter with some constant $c_1 > 0$ that will be specified later.

Remark 2. The estimator \hat{D} defined in (3.5) is similar to that in Zhao et al. (2014), but has better numerical performance due to symmetrization. In addition, we are able to solve (3.5) in a more computationally efficient way. Zhao et al. (2014) vectorized D and transformed the optimization problem (3.5) to a linear programming with a $p^2 \times p^2$ constraint matrix $\hat{\Sigma}_1 \otimes \hat{\Sigma}_2$, which is computationally demanding for large p . In contrast, we solve (3.5) by using the primal-dual interior point method (Candes and Romberg, 2005), and keep the matrix form of D in each step of conjugate gradient descent, by using the matrix multiplications $\frac{1}{2} \hat{\Sigma}_1 D \hat{\Sigma}_2 + \frac{1}{2} \hat{\Sigma}_2 D \hat{\Sigma}_1$ instead of computing $(\frac{1}{2} \hat{\Sigma}_1 \otimes \hat{\Sigma}_2 + \frac{1}{2} \hat{\Sigma}_2 \otimes \hat{\Sigma}_1) \text{vec}(D)$

repeatedly. As a result, the computational complexity is reduced to $O(p^3)$ from $O(p^4)$, and our method is able to handle the problem with larger dimension p . The code is available at <https://github.com/linjunz/SDAR>.

We then proceed to estimating β . Similarly, since the true β satisfies that $\Sigma_2\beta = \mu_2 - \mu_1$, following Cai and Liu (2011), β can be estimated by the following procedure

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|\beta\|_1 : \|\hat{\Sigma}_2\beta - \hat{\mu}_2 + \hat{\mu}_1\|_\infty \leq \lambda_{2,n} \right\}, \quad (3.6)$$

where $\lambda_{2,n} = c_2 \sqrt{\frac{\log p}{n}}$ is a tuning parameter with some constant $c_2 > 0$.

We estimate π_1 and π_2 by $\hat{\pi}_1 = \frac{n_1}{n_1+n_2}$ and $\hat{\pi}_2 = \frac{n_2}{n_1+n_2}$ respectively. Given the solutions \hat{D} and $\hat{\beta}$ to (3.5) and (3.6) and the estimates $\hat{\pi}_1$ and $\hat{\pi}_2$, we then propose the following classification rule: classify z to class 1 if and only if

$$(z - \hat{\mu}_1)^\top \hat{D}(z - \hat{\mu}_1) - 2\hat{\beta}^\top \left(z - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2} \right) - \log(|\hat{D}\hat{\Sigma}_1 + I_p|) + \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) > 0.$$

We shall call this rule the Sparse quadratic Discriminant Analysis rule with Regularization (SDAR), and denote it by \hat{G}_{SDAR} . Analytically, it's written as

$$\begin{aligned} \hat{G}_{\text{SDAR}}(z) = 1 + & \hspace{20em} (3.7) \\ \mathbb{1}\{(z - \hat{\mu}_1)^\top \hat{D}(z - \hat{\mu}_1) - 2\hat{\beta}^\top \left(z - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2} \right) - \log(|\hat{D}\hat{\Sigma}_1 + I_p|) + \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) \leq 0\}. & \end{aligned}$$

The SDAR rule is easy to implement as both (3.5) and (3.6) can be solved by linear programming. We shall show in the next sections that the SDAR rule has desirable properties both theoretically and numerically.

3.4. Theoretical Guarantees

We now study the accuracy of the estimators \hat{D} and $\hat{\beta}$ in (3.5) and (3.6), and the performance of the resulting classifier \hat{G}_{SDAR} in (3.7). We first establish the rates of convergence

for the estimation and classification error and then provide matching minimax lower bounds, up to logarithm factors. These results together show the near-optimality of the SDAR rule.

3.4.1. Upper bounds

To overcome the limitations illustrated in Section 3.2, we consider the following parameter space of $\boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$. Especially, we assume here that both the discriminating direction $\boldsymbol{\beta}$ and the differential graph D are sparse. Let $f_{Q,\boldsymbol{\theta}}$ be the probability density of $Q(\mathbf{z}; \boldsymbol{\theta})$ defined in (3.3), we consider the following parameter space.

$$\begin{aligned} \Theta_p(s_1, s_2) = \{ & \boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2) : \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p, \Sigma_1, \Sigma_2 \succ 0, |D|_0 \leq s_1, \|\boldsymbol{\beta}\|_0 \leq s_2 \\ & \|D\|_F, \|\boldsymbol{\beta}\|_2 \leq M_0, M_1^{-1} \leq \lambda_{\min}(\Sigma_k) \leq \lambda_{\max}(\Sigma_k) \leq M_1, k = 1, 2, \\ & \sup_{|x| < \delta} f_{Q,\boldsymbol{\theta}}(x) < M_2, c \leq \pi_1, \pi_2 \leq 1 - c\}, \end{aligned} \quad (3.8)$$

for some constants $M_0 > 0, M_1 > 1, \delta, M_2 > 0$ and $c \in (0, 1/2)$.

Remark 3. Note that we assume sparsity on both the discriminant direction $\boldsymbol{\beta}$ and the differential graph D , whose necessities are shown by Theorem 1 and 2. The upper bound on $\|\boldsymbol{\beta}\|_2$ is a general assumption in LDA, see Cai and Liu (2011); Neykov et al. (2015); and Cai et al. (2019a), and we assume the same on $\|vec(D)\|_2 = \|D\|_F$ in the QDA setting. Moreover, the condition on the bounded density is commonly assumed in discriminant analysis, see condition (C1) in Cai and Liu (2011), and discussions in Li and Shao (2015) and Jiang et al. (2015). In the following we present a condition on $\boldsymbol{\theta}$ such that this bounded density assumption holds. Note that the term $\mathbf{z}^\top D \mathbf{z} + \boldsymbol{\beta}^\top \mathbf{z}$ is equal in distribution to a weighted non-central chi-square distribution, by using the similar proof as that of Lemma 7.2 in Xu et al. (2014), the condition $\sup_{|x| < \delta} f_{Q,\boldsymbol{\theta}}(x) < M_2$ holds when either the two largest positive eigenvalues of D $\lambda_1(D), \lambda_2(D)$ or the two largest negative eigenvalues of D $\tilde{\lambda}_1(D), \tilde{\lambda}_2(D)$ are of the same order, that is $0 < \liminf_{n \rightarrow \infty} \frac{\lambda_1(D)}{\lambda_1(D) + \lambda_2(D)} < \limsup_{n \rightarrow \infty} \frac{\lambda_1(D)}{\lambda_1(D) + \lambda_2(D)} < 1$ or $0 < \liminf_{n \rightarrow \infty} \frac{\tilde{\lambda}_1(D)}{\tilde{\lambda}_1(D) + \tilde{\lambda}_2(D)} < \limsup_{n \rightarrow \infty} \frac{\tilde{\lambda}_1(D)}{\tilde{\lambda}_1(D) + \tilde{\lambda}_2(D)} < 1$.

At first, we show that over the parameter space $\Theta_p(s_1, s_2)$, the estimators $\hat{D}, \hat{\boldsymbol{\beta}}$ obtained in

(3.5) and (3.6) converge to the true parameters D and β . This theorem will then be used to establish the consistency of the proposed classification rule.

Theorem 3. *Consider the parameter space $\Theta_p(s_1, s_2)$, and assume that $n_1 \asymp n_2$, $s_1 + s_2 \lesssim \frac{n}{\log p}$, where $n = \min\{n_1, n_2\}$. In optimization problems (3.5) and (3.6), let $\lambda_{i,n} = c_i \sqrt{\log p/n}$ with $c_i > 0$, $i = 1, 2$ being sufficiently large constants. Then as n goes to infinity, the estimators obtained in (3.5) and (3.6) satisfies that, with probability at least $1 - p^{-1}$,*

$$\|\hat{D} - D\|_F \lesssim \sqrt{\frac{s_1 \log p}{n}}; \quad \|\hat{\beta} - \beta\|_2 \lesssim \sqrt{\frac{s_2 \log p}{n}}.$$

The above theorem shows that although our estimating procedure (3.6) is different from Zhao et al. (2014), the same convergence rate can be obtained and requires milder theoretical conditions. In fact, Zhao et al. (2014) assumes that $\|\Omega_1\|_1$ and $\|\Omega_2\|_1$ are both bounded, and additionally requires that the off-diagonal elements of Σ_1 and Σ_2 are vanishing as $n \rightarrow \infty$, which is much stronger than conditions in (4.18). In addition, the above bound implies that when $\Sigma_1 = \Sigma_2$, that is, $s_1 = 0$, we have $\hat{D} = D = 0$ when $\lambda_{1,n}$ is suitably chosen. This implies that when the two covariance matrices are equal, SDAR rule (3.7) would adaptively be reduced to the LPD rule in Cai and Liu (2011) designed for high-dimensional LDA.

We now turn to the performance of the classification rule \hat{G}_{SDAR} . The behavior of \hat{G}_{SDAR} is measured by the excess risk $R_{\theta}(\hat{G}_{\text{SDAR}}) - R_{\theta}(G_{\theta}^*)$, defined in (4.32). The following theorem provides the upper bound for the excess classification error.

Theorem 4. *Consider the parameter space $\Theta_p(s_1, s_2)$, and assume that $n_1 \asymp n_2$, $s_1 + s_2 \lesssim \frac{n}{\log p \cdot \log^2 n}$. Then when n goes to infinity, the proposed SDAR classification rule in (3.7) satisfies that, for sufficiently large n ,*

$$\sup_{\theta \in \Theta_p(s_1, s_2)} \mathbb{E} \left[R_{\theta}(\hat{G}_{\text{SDAR}}) - R_{\theta}(G_{\theta}^*) \right] \lesssim (s_1 + s_2) \cdot \frac{\log p}{n} \cdot \log^2 n.$$

The result in Theorem 4 shows that \hat{G}_{SDAR} is able to mimic $G_{\boldsymbol{\theta}}^*$ consistently over the parameter space $\Theta_p(s_1, s_2)$, and to the best of our knowledge, gives the first explicit convergence rate of classification error for the high-dimensional QDA problem.

Remark 4. Related work studying the convergence of classification error includes Li and Shao (2015) and Jiang et al. (2015), but both Theorem 3 in Li and Shao (2015) and Theorem 4 in Jiang et al. (2015) only show the consistency of their proposed classification rules instead of explicit convergence rates. Although in Corollary 3 of Jiang et al. (2015), the authors showed a convergence rate for the classification error of order $s_1 s_2^2 \sqrt{\log p/n}$ under some regularity conditions, this result is based on the assumption that an intercept term η , defined in their paper, is known. Jiang et al. (2015) proposed to estimate η based on the idea of cross validation and in their theorem 3 they showed the consistency of this estimation without explicit convergence rate. In contrast, our paper shows that the convergence rate $O((s_1 + s_2) \log p \cdot \log^2 n/n)$ is achievable, which is much faster than their results. In addition, the assumptions here are weaker.

The major technical challenge of this improvement is the characterization of the distribution of $Q(\mathbf{z}; \boldsymbol{\theta})$, which involves the sum of weighted non-central chi-square random variables. In the next section we will show that this convergence rate is indeed optimal up to logarithm factors.

3.4.2. Minimax lower bound for sparse QDA

In this section we establish the minimax lower bound for the convergence rate of $R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}^*)$, and thus show the optimality of \hat{G}_{SDAR} up to logarithm factors.

Theorem 5. *Consider the parameter space $\Theta_p(s_1, s_2)$ defined in (4.18). Suppose $n_1 \asymp n_2$, $1 \leq s_1, s_2 \leq o(\frac{n}{\log p})$, and \hat{G} is constructed based on the observations $\mathbf{x}_1, \dots, \mathbf{x}_n \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_1, \Sigma_1)$, $\mathbf{y}_1, \dots, \mathbf{y}_n \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_2, \Sigma_2)$. Then the minimax risk of the classification error over $\Theta_p(s_1, s_2)$ satisfies*

$$\inf_{\hat{G}} \sup_{\boldsymbol{\theta} \in \Theta_p(s_1, s_2)} \mathbb{E} \left[R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}^*) \right] \gtrsim (s_1 + s_2) \cdot \frac{\log p}{n}.$$

The challenge of proving Theorem 5 is that the excess risk $R_{\theta}(\hat{G}) - R_{\theta}(G_{\theta}^*)$ does not satisfy the triangle inequality (or subadditivity), which is essential to the standard minimax lower bound techniques. To overcome this challenge, we define an alternative risk function $L_{\theta}(\hat{G})$ as follows,

$$L_{\theta}(\hat{G}) := \mathbb{P}_{\theta} \left(\hat{G}(\mathbf{z}) \neq G_{\theta}^*(\mathbf{z}) \right). \quad (3.9)$$

This loss function $L_{\theta}(\hat{G})$ is essentially the probability that \hat{G} produces a different label than G_{θ}^* , and satisfies the triangle inequality, as shown in Lemma 13. The connection between $R_{\theta}(\hat{G}) - R_{\theta}(G_{\theta}^*)$ and $L_{\theta}(\hat{G})$ is presented by the following lemma, which shows that it's sufficient to provide a lower bound for $L_{\theta}(\hat{G})$ to prove Theorem 5.

Lemma 1. *Suppose $\theta \in \Theta_p(s_1, s_2)$. There exists a constant $c > 0$, doesn't depend on n, p , such that for some classification rule G , if $L_{\theta}(G) < c$, then,*

$$L_{\theta}^2(G) \lesssim \mathbb{P}_{\theta}(G(\mathbf{z}) \neq L(\mathbf{z})) - \mathbb{P}_{\theta}(G_{\theta}(\mathbf{z}) \neq L(\mathbf{z})).$$

Based on Lemma 1, we use Fano's inequality on a carefully designed least favorable multivariate normal distributions to complete the proof of Theorems 2 and 5. The details are shown in Section 3.7.

3.5. Numerical Studies

In this section we firstly conduct simulation studies to investigate the impossibility results shown in Section 3.2.2, and then study numerical properties of the proposed SDAR method under various settings.

3.5.1. Impossibility results

We would like to illustrate the impossibility results Theorem 1 and Theorem 2 in a numerical fashion in this subsection.

Let us start with Theorem 1, which shows the sparsity condition on β is necessary. In the

simulation, we consider the simple case where both covariance matrices are known to be identity but the means are unknown: $\mathbf{x}_1, \dots, \mathbf{x}_n \sim N_p(\boldsymbol{\mu}_1, I_p)$ and $\mathbf{y}_1, \dots, \mathbf{y}_n \sim N_p(\boldsymbol{\mu}_2, I_p)$ and let $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = \boldsymbol{\mu} = \frac{1}{\sqrt{p}} \cdot \mathbf{1}_p$, satisfying $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2 = 2$.

We consider nine cases where $(n, p) = (100, 200), (150, 200), (200, 200), (100, 300), (200, 300), (300, 300), (200, 600), (400, 600), (600, 600)$. In each setting, we compare the oracle classification rule $G_{\boldsymbol{\theta}}^*$ in (3.1) with the plug-in classification rule \hat{G} where we estimate $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ by the sample means. The testing sample size is set to 100 and the simulation is repeated 100 times in each setting. The simulations results is summarized in the following table.

Table 9: Average classification errors (s.e.) based on $n = 100$ test samples from 100 replications under the setting where covariance matrices are known to be identity.

| | n | $R_{\boldsymbol{\theta}}(\hat{G})$ | $R_{\boldsymbol{\theta}}(G_{opt})$ |
|-------|-----|------------------------------------|------------------------------------|
| p=200 | 100 | 0.242 (0.054) | 0.155 (0.035) |
| | 150 | 0.232 (0.051) | 0.155 (0.035) |
| | 200 | 0.219 (0.039) | 0.155 (0.035) |
| p=300 | 100 | 0.265 (0.048) | 0.149(0.032) |
| | 200 | 0.223 (0.047) | 0.149(0.032) |
| | 300 | 0.208 (0.038) | 0.149(0.032) |
| p=600 | 200 | 0.269 (0.045) | 0.158 (0.035) |
| | 400 | 0.230 (0.035) | 0.158 (0.035) |
| | 600 | 0.201 (0.035) | 0.158 (0.035) |

To illustrate Theorem 2, we consider a simple case where $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = (1, 0, 0, \dots, 0)^\top$ and the covariance matrices are known to be diagonal. Two classes are $N_p(\boldsymbol{\mu}_1, I_p)$ and $N_p(\boldsymbol{\mu}_2, \Sigma_2)$, where $\Sigma_2 = (I_p + \sum_{i=1}^{p/2} \frac{2}{\sqrt{p}} E_{i,i})^{-1}$ and $E_{i,i}$ is a $p \times p$ matrix whose (i, i) -th entry is 1 and 0 else.

We consider nine cases where $(n, p) = (100, 200), (150, 200), (200, 200), (100, 300), (200, 300), (300, 300), (200, 600), (400, 600), (600, 600)$. In each setting, we compare the oracle classifi-

cation rule G_{opt} , that is (3.1), with the plug-in classification rule \hat{G} where we estimate Σ_1, Σ_2 by the diagonals of sample covariance matrices. The following table summarizes the simulation results where the testing sample size is set to 100 and the simulation is repeated 100 times.

Table 10: Average classification errors (s.e.) based on $n = 100$ test samples from 100 replications under the setting where means are known to be $\mathbf{0}_p$ and covariance matrices are known to be diagonal.

| | n | $R_{\theta}(\hat{G})$ | $R_{\theta}(G_{opt})$ |
|-------|-----|-----------------------|-----------------------|
| p=200 | 100 | 0.274 (0.049) | 0.193 (0.038) |
| | 150 | 0.260 (0.036) | 0.193 (0.038) |
| | 200 | 0.252 (0.033) | 0.193 (0.038) |
| p=300 | 100 | 0.271 (0.043) | 0.151(0.034) |
| | 200 | 0.238 (0.048) | 0.151(0.034) |
| | 300 | 0.224 (0.039) | 0.151(0.034) |
| p=600 | 200 | 0.296 (0.032) | 0.183 (0.046) |
| | 400 | 0.255 (0.055) | 0.183 (0.046) |
| | 600 | 0.245 (0.037) | 0.183 (0.046) |

3.5.2. SDAR on synthetic data

In this section, we provide extensive numerical evidence to show the empirical performance of SDAR by comparing it to its competitors, including the sparse QDA (SQDA, Li and Shao (2015)), the direct approach for sparse LDA (LPD, Cai and Liu (2012)), the conventional LDA (LDA), the conventional QDA (QDA) and the oracle procedure (Oracle). The oracle procedure uses the true underlying model and serves as the optimal risk bound for comparison. We evaluate all methods via three synthetic datasets.

In all simulations, the sample size is $n_1 = n_2 = 200$ while the number of variables p varies from 100, 200, 400 to 600. The sparsity levels are set to be $s_1 = 10, s_2 = 20$. The discriminating direction $\beta = (1, \dots, 1, 0, \dots, 0)^\top$ is sparse such that only the first $s_1 = 10$

entries are nonzero. Given the inverse covariance matrix of the second sample Ω_2 , the mean for class 1 is $\boldsymbol{\mu}_1 = (0, \dots, 0)^\top$ and the mean for class 2 is set to be $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 - \Sigma_2 \boldsymbol{\beta}$. In addition, the differential graph D is a random sparse symmetric matrix with its nonzero positions generated by uniform sample. Each nonzero entry on D is *i.i.d.* and from a standard normal distribution $N(0, 1)$. Lastly, we let $\Omega_1 = D + \Omega_2$, and $\Omega_1 = \Sigma_1^{-1}, \Omega_2 = \Sigma_2^{-1}$. We use the following three models to generate Ω_2 .

Model 1: Block sparse model: We generate $\Omega_2 = U^T \Lambda U$, where $\Lambda \in \mathbb{R}^{p \times p}$ is a diagonal matrix and its entries are *i.i.d.* and uniform on $[1, 2]$, and $U \in \mathbb{R}^{p \times p}$ is a random matrix with *i.i.d.* entries from $N(0, 1)$. In the simulation, the tuning parameters for SDAR method are chosen over a grid $\{\frac{k}{2} \sqrt{\frac{\log p}{n}}\}_{k=1:15}$.

Model 2: AR(1) model: $\Omega_2 = (\Omega_{ij})_{p \times p}$ with $\Omega_{ij} = \rho^{|i-j|}$. In the simulation, the tuning parameters for the SDAR method are chosen by cross validation over a grid $\{\frac{k}{4} \sqrt{\frac{\log p}{n}}\}_{k=1:15}$. The simulation results from 100 replications are summarized as follows, with $\rho = 0.5$.

Model 3: Erdős-Rényi random graph: Let $\tilde{\Omega}_2 = (\tilde{\omega}_{ij})$ where $\tilde{\omega}_{ij} = u_{ij} \delta_{ij}$, $\delta_{ij} \sim \text{Ber}(1, \rho)$ being the Bernoulli random variable with success probability 0.05 and $u_{ij} \sim \text{Unif}[0.5, 1] \cup [-1, -0.5]$. After symmetrizing $\tilde{\Omega}_2$, set $\Omega_2 = \tilde{\Omega}_2 + \{\max(-\phi_{\min}(\tilde{\Omega}_2), 0) + 0.05\} \mathbf{I}_p$ to ensure the positive definiteness. In the simulation, the tuning parameters for SDAR method are chosen over a grid $\{\frac{k}{2} \sqrt{\frac{\log p}{n}}\}_{k=1:15}$.

In each model, the number of repetition is set to be 100, and the classification errors are evaluated based on the test data with size 100 that is generated from a Gaussian mixture model $\frac{1}{2} N_p(\boldsymbol{\mu}_1, \Sigma_1) + \frac{1}{2} N_p(\boldsymbol{\mu}_2, \Sigma_2)$. We compare the proposed SDAR method with the oracle QDA rule (3.1). The simulation results are summarized in Table 11.

This simulation result shows that the proposed SDAR algorithm outperforms the LPD algorithm when there are strong interactions among features ($D \neq 0$). As expected, the conventional LDA and QDA works poorly in the high-dimensional setting, and the perfor-

Table 11: Average classification errors (s.d.) based on $n = 200$ test samples from 100 replications under three different models

| p | | 100 | 200 | 400 | 600 |
|---------|--------------------|--------------|--------------|--------------|--------------|
| Model 1 | LDA | 0.200(0.019) | 0.224(0.028) | 0.269(0.022) | 0.302(0.024) |
| | QDA | 0.236(0.026) | 0.274(0.023) | 0.418(0.025) | 0.432(0.027) |
| | SQDA (Shao et al.) | 0.202(0.022) | 0.231(0.027) | 0.301(0.023) | 0.347(0.025) |
| | LPD | 0.151(0.020) | 0.163(0.021) | 0.208(0.028) | 0.256(0.025) |
| | SDAR | 0.075(0.019) | 0.089(0.022) | 0.091(0.029) | 0.102(0.027) |
| | Oracle | 0.044(0.010) | 0.023(0.007) | 0.039(0.010) | 0.047(0.009) |
| Model 2 | LDA | 0.231(0.022) | 0.214(0.021) | 0.335(0.025) | 0.378(0.027) |
| | QDA | 0.249(0.025) | 0.296(0.029) | 0.405(0.026) | 0.446(0.028) |
| | SQDA (Shao et al.) | 0.214(0.023) | 0.243(0.024) | 0.327(0.023) | 0.376(0.025) |
| | LPD | 0.163(0.018) | 0.156(0.019) | 0.220(0.027) | 0.253(0.024) |
| | SDAR | 0.065(0.015) | 0.042(0.014) | 0.081(0.020) | 0.092(0.019) |
| | Oracle | 0.045(0.010) | 0.025(0.007) | 0.031(0.008) | 0.045(0.008) |
| Model 3 | LDA | 0.279(0.028) | 0.305(0.032) | 0.340(0.031) | 0.387(0.029) |
| | QDA | 0.298(0.024) | 0.356(0.025) | 0.406(0.026) | 0.457(0.025) |
| | SQDA (Shao et al.) | 0.242(0.024) | 0.294(0.029) | 0.335(0.026) | 0.374(0.026) |
| | LPD | 0.236(0.023) | 0.205(0.020) | 0.234(0.031) | 0.252(0.027) |
| | SDAR | 0.078(0.022) | 0.077(0.026) | 0.096(0.028) | 0.112(0.026) |
| | Oracle | 0.065(0.013) | 0.039(0.009) | 0.031(0.008) | 0.048(0.010) |

mance of conventional QDA is even worse due to overfitting. In the setting where $D = 0$, the estimated \hat{D} would equal to $D = 0$ for properly chosen λ_1 , according to Theorem 3. As we estimate β and D separately, the proposed SDAR rule in this case would adaptively reduced to LPD. For reasons of space we do not present the detailed numerical results for this case.

3.5.3. Real data

In addition to the simulation studies, we also illustrate the merits of the SDAR classifier in the analysis of two real datasets to further investigate the numerical performance of the proposed method. One is the prostate cancer data in Singh, et al. (2002), which is available at <ftp://stat.ethz.ch/Manuscripts/dettling/prostate.rda>, and another dataset is the colon tissues data analyzed in Alon et al. (1999) by using the Oligonucleotide microarray technique, available at <http://microarray.princeton.edu/oncology/affydata/index.html>. These two datasets were frequently used for illustrating the empirical performance of

the classifier for high-dimensional data in recent literature, see Dettling (2004) and Efron (2010). We will compare SDAR with the existing methods, including the sparse QDA (SQDA, Li and Shao (2015)), the direct approach for sparse LDA (LPD, Cai and Liu (2012)), the conventional LDA (LDA), the conventional QDA (QDA).

Prostate cancer data

The prostate cancer data consists of genetic expression levels for $p = 6033$ genes from 102 individuals (50 normal control subjects and 52 prostate cancer patients). The SDAR classifier allows us to model the interactions among genes and thus improve the classification accuracy. For this data, we follow the same data cleaning routine in Cai and Liu (2011), retaining only the top 200 genes with the largest absolute values of the two sample t -statistics. The average classification errors using 5-fold cross-validation for various methods with 50 repetitions are reported in Table 12. The proposed SDAR method outperforms all the other methods

Table 12: Classification error(%) with s.d. of prostate cancer data by various methods

| | SDAR | SQDA (Shao et al.) | LPD | LDA | QDA |
|---------------|-------------|--------------------|--------------|--------------|--------------|
| Testing error | 2.20 (1.11) | 3.10 (1.26) | 11.20 (1.87) | 32.20 (3.67) | 35.30 (4.18) |

Colon tissues data

The colon tissues data analyzed gene expression difference between tumor and normal colon tissues using the Oligonucleotide microarray technique, consisting 20 observations from normal tissues and 42 observations from tumor tissues, measured in $p = 2000$ genes.

Similarly to the analysis of the prostate cancer data, to control the computational costs, we use 200 genes with the largest absolute values of the two sample t -statistics. Classification results by using 5-fold cross-validation with 50 repetitions are summarized in Table 13. In this example, the SDAR is still the best among all classifiers.

Table 13: Classification error(%) with s.d. of prostate cancer data by various methods

| | SDAR | SQDA (Shao et al.) | LPD | LDA | QDA |
|---------------|--------------|--------------------|--------------|--------------|--------------|
| Testing error | 19.05 (2.40) | 23.20 (2.36) | 26.67 (2.75) | 38.20 (3.14) | 39.30 (4.71) |

3.6. Extensions

We have so far focused on high-dimensional QDA for two groups in the Gaussian setting. The methodology and theory developed in the earlier sections can be extended to multi-group classification and to classification under the Gaussian copula model.

3.6.1. Multi-group classification

We first turn to multi-group classification. Suppose there are K classes $N_p(\boldsymbol{\mu}_k, \Sigma_k)$ with prior probability π_k for $1 \leq k \leq K$ respectively, and an observation \mathbf{z} is drawn from the same distribution. In the ideal setting where all the parameters are known, the oracle rule classifies \mathbf{z} to class k if and only if

$$k = \arg \min_{k \in [K]} \{Q_k(\mathbf{z})\},$$

where the discriminating function $Q_k(\mathbf{z})$ is

$$Q_k(\mathbf{z}) = \begin{cases} 1, & k = 1 \\ \frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_k)^\top D_k(\mathbf{z} - \boldsymbol{\mu}_k) - \boldsymbol{\beta}_k^\top (\mathbf{z} - \bar{\boldsymbol{\mu}}_k) - \frac{1}{2} \log |D_k \Sigma_1 + I_p| + \log \pi_k, & k \geq 2, \end{cases}$$

with $\bar{\boldsymbol{\mu}}_k = \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_k}{2}$, $D_k = \Omega_1 - \Omega_k$, $\boldsymbol{\beta}_k = \Omega_1(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$, and $\Omega_k = \Sigma_k^{-1}$. When the parameters are unknown and random samples from K classes (with prior probabilities $\{\pi_k\}_{k=1}^K$) are available: $\mathbf{x}_1^{(k)}, \dots, \mathbf{x}_{n_k}^{(k)} \stackrel{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_k, \Sigma_k)$, $k = 1, \dots, K$, by assuming the sparsity on D_k 's and $\boldsymbol{\beta}_k$'s, they can then be estimated by solving a similar linear programming as in (3.5) and

(3.6). For $k = 2, 3, \dots, K$, D_k and β_k are estimated by

$$\hat{D}_k = \arg \min_{D \in \mathbb{R}^{p \times p}} \left\{ |D|_1 : \left| \frac{1}{2} \hat{\Sigma}_1 D \hat{\Sigma}_k + \frac{1}{2} \hat{\Sigma}_2 D \hat{\Sigma}_1 - \hat{\Sigma}_1 + \hat{\Sigma}_k \right|_\infty \leq \lambda_{1,n} \right\}, \quad (3.10)$$

where $\lambda_{1,n}$ is a tuning parameter with constant $c_1 > 0$.

$$\hat{\beta}_k = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|\beta\|_1 : \|\hat{\Sigma}_1 \beta - \hat{\mu}_k + \hat{\mu}_1\|_\infty \leq \lambda_{2,n} \right\}, \quad (3.11)$$

where $\lambda_{2,n}$ is a tuning parameter with constant $c_2 > 0$.

Given these estimators and $\hat{\pi}_k = n_k / (\sum_{k=1}^K n_k)$, the discriminating function is then estimated by

$$\hat{Q}_k(\mathbf{z}) = \begin{cases} 1, & k = 1 \\ \frac{1}{2}(\mathbf{z} - \hat{\mu}_k)^\top \hat{D}_k(\mathbf{z} - \hat{\mu}_k) - \hat{\beta}_k^\top(\mathbf{z} - \hat{\mu}_k) - \frac{1}{2} \log |\hat{D}_k \hat{\Sigma}_1 + I_p| + \log \hat{\pi}_k, & k \geq 2, \end{cases}$$

Then the SDAR classification rule for multi-group classification is constructed as

$$\hat{G}(\mathbf{z}) = \arg \min_{k \in [K]} \{\hat{Q}_k(\mathbf{z})\}.$$

By applying the same techniques we developed for Theorems 3 and 4, similar convergence rates can be obtained for both estimation and classification errors.

3.6.2. Classification under Gaussian copula model

The Gaussianity assumption can be related by incorporating semiparametric Gaussian copula model into the QDA framework. This larger semiparametric Gaussian copula model enables robust estimation and classification, and has been studied widely in statistics and machine learning, including linear discriminant analysis (Han et al., 2013; Mai and Zou, 2015), correlation matrix estimation (Han and Liu, 2017), graphical models (Liu et al., 2012; Xue and Zou, 2012), and linear regression (Cai and Zhang, 2018c).

The Semiparametric Discriminant Analysis (SeDA) model, introduced by Lin and Jeon (2003), assumes that there are K groups of p -dimensional observations $\mathbf{x}_1^{(1)}, \dots, \mathbf{x}_{n_1}^{(1)} \sim \mathbf{X}^{(1)}$, $\mathbf{x}_1^{(2)}, \dots, \mathbf{x}_{n_2}^{(2)} \sim \mathbf{X}^{(2)}$, ..., $\mathbf{x}_1^{(K)}, \dots, \mathbf{x}_{n_K}^{(K)} \sim \mathbf{X}^{(K)}$, and there are some unknown strictly increasing functions $f_{11}, \dots, f_{1p}, \dots, f_{K1}, \dots, f_{Kp}$ such that

$$f_k(\mathbf{X}^{(k)}) = (f_{k1}(X_1^{(k)}), \dots, f_{kp}(X_p^{(k)})) \sim N_p(\boldsymbol{\mu}_k, \Sigma_k) \text{ for } k = 1, \dots, K.$$

The linear SeDA model in the high-dimensional setting was recently studied by Han et al. (2013) and Mai and Zou (2015) under the assumption that Σ_k 's are all equal. By applying the LPD idea in Cai and Liu (2011), consistent classification rules were proposed under this semiparametric linear discriminant analysis model.

The current paper presents a framework to extend the high-dimensional semiparametric LDA to high-dimensional semiparametric QDA. Estimating the mean vectors and covariance matrices similarly as in Han et al. (2013); Mai and Zou (2015) and then plugging these estimators in (3.5) and (3.6) would lead to a generalized classification rule under the semiparametric quadratic discriminant analysis model. We omit further detailed discussion for reasons of space.

3.7. Proofs

We present the proofs of Theorems 1, 2, 3, 4 in this section. The proof of Theorem 5 is similar to Theorems 1, 2, so we present its proof in the supplement.

3.7.1. Proof of Theorem 1 and 2

We prove Theorem 1 and 2 for the case where $p \lesssim n$. In the case where $\limsup_{n \rightarrow \infty} p/n = \infty$, the right hand side of Theorem 1 (and 2) is of constant order and we can consider only the first n -dimension of p -dimensional vector, and assume the rest is known.

We begin by collecting a few important technical lemmas that will be used in the proofs of the minimax lower bounds.

Technical lemmas

Lemma 2 (Azizyan et al. (2013)). *For any $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta_p(s_1, s_2)$ and any classification rule \hat{G} , recall that $G_{\tilde{\boldsymbol{\theta}}}^*$ is the optimal rule w.r.t. $\tilde{\boldsymbol{\theta}}$. If*

$$L_{\boldsymbol{\theta}}(G_{\tilde{\boldsymbol{\theta}}}^*) + L_{\boldsymbol{\theta}}(\hat{G}) + \sqrt{\frac{KL(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})}{2}} \leq 1/2,$$

then

$$L_{\boldsymbol{\theta}}(G_{\tilde{\boldsymbol{\theta}}}^*) - L_{\boldsymbol{\theta}}(\hat{G}) - \sqrt{\frac{KL(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})}{2}} \leq L_{\tilde{\boldsymbol{\theta}}}(\hat{G}) \leq L_{\boldsymbol{\theta}}(G_{\tilde{\boldsymbol{\theta}}}^*) + L_{\boldsymbol{\theta}}(\hat{G}) + \sqrt{\frac{KL(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})}{2}},$$

where the KL divergence of two probability density functions $\mathbb{P}_{\boldsymbol{\theta}_1}$ and $\mathbb{P}_{\boldsymbol{\theta}_2}$ is defined by

$$KL(\mathbb{P}_{\boldsymbol{\theta}_1}, \mathbb{P}_{\boldsymbol{\theta}_2}) = \int \mathbb{P}_{\boldsymbol{\theta}_1}(x) \log \frac{\mathbb{P}_{\boldsymbol{\theta}_1}(x)}{\mathbb{P}_{\boldsymbol{\theta}_2}(x)} dz.$$

Lemma 3 (Tsybakov (2009)). *Let $M \geq 0$ and $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_M \in \Theta_p(s_1, s_2)$. For some constants $\alpha \in (0, 1/8), \gamma > 0$, and any classification rule \hat{G} , if $KL(\mathbb{P}_{\boldsymbol{\theta}_i}, \mathbb{P}_{\boldsymbol{\theta}_0}) \leq \alpha \log M/n$ for all $1 \leq i \leq M$, and $L_{\boldsymbol{\theta}_i}(\hat{G}) < \gamma$ implies $L_{\boldsymbol{\theta}_j}(\hat{G}) \geq \gamma$ for all $0 \leq i \neq j \leq M$, then*

$$\inf_{\hat{G}} \sup_{i \in [M]} \mathbb{E}_{\boldsymbol{\theta}_i}[L_{\boldsymbol{\theta}_i}(\hat{G})] \gtrsim \gamma.$$

To use Fano's type minimax lower bound, we need a covering number argument, provided by the following Lemma 16.

Lemma 4 (Tsybakov (2009)). *Define $\mathcal{A}_{p,s} = \{\mathbf{u} : \mathbf{u} \in \{0, 1\}^p, \|\mathbf{u}\|_0 = s\}$. If $p \geq 4s$, then there exists a subset $\{\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_M\} \subset \mathcal{A}_{p,s}$ such that $\mathbf{u}_0 = \{0, \dots, 0\}^\top$, $\rho_H(\mathbf{u}_i, \mathbf{u}_j) \geq s/2$ and $\log(M+1) \geq \frac{s}{5} \log(\frac{p}{s})$, where ρ_H denotes the Hamming distance.*

Main proof of Theorem 1

At first we construct the following least favorable subset, which characterizes the difficulty of the general QDA problem. Let's consider the parameter space

$$\Theta_1 = \{\boldsymbol{\theta}_u = (1/2, 1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, I_p, I_p) : \\ \boldsymbol{\mu}_1 = \lambda_1 \mathbf{e}_1 + \sum_{i=2}^p \frac{\lambda_2}{\sqrt{n}} \cdot u_i \cdot \mathbf{e}_i, \mathbf{u} \in \mathcal{A}_{p,p/4}, \boldsymbol{\mu}_2 = \mathbf{0}_p\},$$

where $\mathcal{A}_{p,p/4}$ is defined in Lemma 16, and λ_1, λ_2 are of constant order and chosen later.

According to Lemma 16, there is a subset of Θ_1 with logarithm cardinality being of order p , such that for any $\boldsymbol{\theta}_u, \boldsymbol{\theta}_{u'}$ in this subset, we have $\rho_H(\mathbf{u}, \mathbf{u}') \geq p/8$. We are going to apply Lemma 15 to this subset to complete the proof of Theorem 1.

For $\mathbf{u} \in \mathcal{A}_{p,p/4}$, let $\boldsymbol{\mu}_u = \lambda_1 \mathbf{e}_1 + \sum_{i=2}^p \frac{\lambda_2}{\sqrt{n}} \cdot u_i \cdot \mathbf{e}_i$. Note that for two multivariate normal distributions $\mathbb{P}_{\boldsymbol{\theta}_u} = N_p(\boldsymbol{\mu}_u, I_p)$ and $\mathbb{P}_{\boldsymbol{\theta}_{u'}} = N_p(\boldsymbol{\mu}_{u'}, I_p)$, the KL divergence between them are upper bounded by

$$KL(\mathbb{P}_{\boldsymbol{\theta}_u}, \mathbb{P}_{\boldsymbol{\theta}_{u'}}) = \frac{1}{2} \|\boldsymbol{\mu}_u - \boldsymbol{\mu}_{u'}\|_2^2 \leq \frac{\lambda_2^2 \cdot p}{4n}.$$

To use Lemma 15 to prove Theorem 1, we further need to show that for any $\boldsymbol{\theta}_u, \boldsymbol{\theta}_{u'}$,

$$[R_{\boldsymbol{\theta}}(G) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}_u}^*)] + [R_{\boldsymbol{\theta}}(G) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}_{u'}}^*)] \gtrsim \frac{p}{n}.$$

By Lemma 1 and 13,

$$\begin{aligned} & [R_{\boldsymbol{\theta}}(G) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}_u}^*)] + [R_{\boldsymbol{\theta}}(G) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}_{u'}}^*)] \\ & \gtrsim L_{\boldsymbol{\theta}_u}^2(G) + L_{\boldsymbol{\theta}_{u'}}^2(G) \geq \frac{1}{2} (L_{\boldsymbol{\theta}_u}(G) + L_{\boldsymbol{\theta}_{u'}}(G))^2 \geq \frac{1}{2} (L_{\boldsymbol{\theta}_u}(G_{\boldsymbol{\theta}_{u'}}^*) - \sqrt{\frac{KL(\mathbb{P}_{\boldsymbol{\theta}_u}, \mathbb{P}_{\boldsymbol{\theta}_{u'}})}{2}})^2. \end{aligned}$$

Since now that $KL(\mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}}}, \mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}'}}) \leq \frac{\lambda_2^2 p}{4n}$, it's then sufficient to show $L_{\boldsymbol{\theta}_{\mathbf{u}}}(G_{\boldsymbol{\theta}_{\mathbf{u}'}}^*) \geq c\sqrt{\frac{p}{n}}$ for some $c > \frac{\lambda_2}{2\sqrt{2}}$.

Without loss of generality, we assume that the coordinates of \mathbf{u} and \mathbf{u}' are ordered such that $u_i = u'_i = 1$ for $i = 2, \dots, m_1$, $u_i = 1 - u'_i = 1$ for $i = m_1 + 1, \dots, m_2$, $u_i = 1 - u'_i = 0$ for $i = m_2 + 1, \dots, m_3$ and $u_i = u'_i = 0$ for $i = m_3 + 1, \dots, p$. We then have $\rho_H(\mathbf{u}, \mathbf{u}') = m_3 - m_1 \geq \frac{p}{8}$.

Recall that when $\Sigma_1 = \Sigma_2 = I_p$ and $\boldsymbol{\mu}_2 = \mathbf{0}_p$, the oracle rule is given by

$$G_{\boldsymbol{\theta}}^*(\mathbf{z}) = 1 + \mathbb{1}\{-\boldsymbol{\mu}_1^\top(\mathbf{z} - \frac{\boldsymbol{\mu}_1}{2}) > 0\}.$$

Then

$$G_{\boldsymbol{\theta}_{\mathbf{u}}}^*(\mathbf{z}) = 1 + \mathbb{1}\left\{-\frac{\lambda_2}{\sqrt{n}}\left(\sum_{i=2}^{m_1} z_i + \sum_{i=m_1+1}^{m_2} z_i\right) - \lambda_1 z_1 + \frac{1}{2}\lambda_1^2 + \frac{\lambda_2^2(p-1)}{8n} > 0\right\},$$

and

$$G_{\boldsymbol{\theta}_{\mathbf{u}'}}^*(\mathbf{z}) = 1 + \mathbb{1}\left\{-\frac{\lambda_2}{\sqrt{n}}\left(\sum_{i=2}^{m_1} z_i + \sum_{i=m_2+1}^{m_3} z_i\right) - \lambda_1 z_1 + \frac{1}{2}\lambda_1^2 + \frac{\lambda_2^2(p-1)}{8n} > 0\right\}.$$

Let $Z_1 = -\lambda_1 z_1 - \frac{\lambda_2}{\sqrt{n}} \sum_{i=2}^{m_1} z_i + \frac{1}{2}\lambda_1^2 + \frac{\lambda_2^2(p-1)}{8n}$, $Z_2 = \frac{\lambda_2}{\sqrt{n}} \sum_{i=m_1+1}^{m_2} z_i$ and $Z_3 = \frac{\lambda_2}{\sqrt{n}} \sum_{i=m_2+1}^{m_3} z_i$, then

$$G_{\boldsymbol{\theta}_{\mathbf{u}}}^*(\mathbf{z}) = 1 + \mathbb{1}\{Z_1 - Z_2 > 0\} \text{ and } G_{\boldsymbol{\theta}_{\mathbf{u}'}}^*(\mathbf{z}) = 1 + \mathbb{1}\{Z_1 - Z_3 > 0\},$$

and therefore

$$\begin{aligned}
L_{\boldsymbol{\theta}_u}(G_{\boldsymbol{\theta}_{u'}}^*) &= \mathbb{P}_{\boldsymbol{\theta}_u}(G_{\boldsymbol{\theta}_{u'}}^*(z) \neq G_{\boldsymbol{\theta}_u}^*(z)) \\
&= \mathbb{P}_{\boldsymbol{\theta}_u}(Z_2 \leq Z_1 \leq Z_3) + \mathbb{P}_{\boldsymbol{\theta}_u}(Z_3 \leq Z_1 \leq Z_2) \\
&\geq \mathbb{P}_{\boldsymbol{\theta}_u}(Z_2 \leq Z_1 \leq Z_3) \\
&= \frac{1}{2} \mathbb{P}_{z \sim N_p(\boldsymbol{\mu}_u, I_p)}(Z_2 \leq Z_1 \leq Z_3) + \frac{1}{2} \mathbb{P}_{z \sim N_p(\mathbf{0}_p, I_p)}(Z_2 \leq Z_1 \leq Z_3) \\
&\geq \frac{1}{2} \mathbb{P}_{z \sim N_p(\mathbf{0}_p, I_p)}(Z_2 \leq Z_1 \leq Z_3),
\end{aligned}$$

Then, since $Z_1 \sim N\left(\frac{1}{2}\lambda_1^2 + \frac{\lambda_2^2(p-1)}{8n}, \lambda_1^2 + \lambda_2^2 p/(4n)\right)$, the density of Z_1 , $f(z)$ satisfies,

$$f(z) \geq \frac{1}{\sqrt{2\pi(\lambda_1^2 + \lambda_2^2 p/(4n))}} \exp\left(-\frac{(z - \lambda_1^2/2 - \lambda_2^2(p-1)/(8n))^2}{2(\lambda_1^2 + \lambda_2^2 p/(4n))^2}\right),$$

leading to

$$f(z) \geq c_1(\lambda_1, \lambda_2), \text{ for } z \in [-\lambda_2\sqrt{p/n}, \lambda_2\sqrt{p/n}],$$

for some constant $c_1(\lambda_1, \lambda_2) = \frac{1}{\sqrt{2\pi(\lambda_1^2 + \lambda_2^2 p/(4n))}} \exp\left(-\frac{(\lambda_2\sqrt{p/n} + \lambda_1^2/2 + \lambda_2^2(p-1)/(8n))^2}{2(\lambda_1^2 + \lambda_2^2 p/(4n))^2}\right)$.

In addition, since $m_3 - m_1 \in (\frac{p}{8}, \frac{p}{2})$, $Z_3 - Z_2$ is normally distributed with mean 0 and variance of order $\frac{p}{n}$, and therefore we claim that for some constant c_2 ,

$$\mathbb{E}[(Z_3 - Z_2) \cdot \mathbb{1}\{-\lambda_2\sqrt{\frac{p}{n}} < Z_2 < Z_3 < \lambda_2\sqrt{\frac{p}{n}}\}] \geq c_2\lambda_2\sqrt{\frac{p}{n}}.$$

In fact,

$$\begin{aligned}
& \mathbb{E}[(Z_3 - Z_2) \cdot \mathbb{1}\{-\lambda_2\sqrt{\frac{p}{n}} < Z_2 < Z_3 < \lambda_2\sqrt{\frac{p}{n}}\}] \\
& \geq \mathbb{E}[(Z_3 - Z_2) \cdot \mathbb{1}\{-\lambda_2\sqrt{\frac{p}{n}} < Z_2 < -\frac{\lambda_2}{2}\sqrt{\frac{m_2 - m_1}{n}}, \frac{\lambda_2}{2}\sqrt{\frac{m_3 - m_2}{n}} < Z_3 < \lambda_2\sqrt{\frac{p}{n}}\}] \\
& \geq \lambda_2\sqrt{\frac{p}{n}} \cdot \mathbb{P}(-\lambda_2\sqrt{\frac{p}{n}} < Z_2 < -\frac{\lambda_2}{2}\sqrt{\frac{m_2 - m_1}{n}}) \cdot \mathbb{P}(\frac{\lambda_2}{2}\sqrt{\frac{m_3 - m_2}{n}} < Z_3 < \lambda_2\sqrt{\frac{p}{n}}) \\
& \geq \lambda_2\sqrt{\frac{p}{8n}} \cdot \mathbb{P}_{Z \sim N(0,1)}(-\sqrt{\frac{p}{m_2 - m_1}} < Z < -\frac{1}{2}) \cdot \mathbb{P}_{Z \sim N(0,1)}(\frac{1}{2} < Z < \sqrt{\frac{p}{m_3 - m_2}}) \\
& \geq \lambda_2\sqrt{\frac{p}{8n}} \cdot \mathbb{P}_{Z \sim N(0,1)}(-\sqrt{2} < Z < -\frac{1}{2}) \cdot \mathbb{P}_{Z \sim N(0,1)}(\frac{1}{2} < Z < \sqrt{2}) := c_2\lambda_2\sqrt{\frac{p}{n}},
\end{aligned}$$

where $c_2 = \sqrt{\frac{1}{8}}\mathbb{P}_{Z \sim N(0,1)}(-\sqrt{2} < Z < -\frac{1}{2}) \cdot \mathbb{P}_{Z \sim N(0,1)}(\frac{1}{2} < Z < \sqrt{2})$ is of constant order and the inequality above uses $\sqrt{m_2 - m_1} + \sqrt{m_3 - m_2} \geq \sqrt{m_3 - m_1} \geq \sqrt{p/8}$, $m_2 - m_1, m_3 - m_2 \leq m_3 - m_1 \leq p/2$.

Then we have

$$\begin{aligned}
& \mathbb{P}_{\mathbf{z} \sim N_p(\mathbf{0}_p, I_p)}(Z_2 \leq Z_1 \leq Z_3) \geq \mathbb{P}_{\mathbf{z} \sim N_p(\mathbf{0}_p, I_p)}\left(Z_2 \leq Z_1 \leq Z_3, -\lambda_2\sqrt{\frac{p}{n}} < Z_2 < Z_3 < \lambda_2\sqrt{\frac{p}{n}}\right) \\
& = \mathbb{E}_{Z_2} \left[\int_{Z_2}^{Z_3} f(z_1) dz_1 \cdot \mathbb{1}\{-\lambda_2\sqrt{\frac{p}{n}} < Z_2 < Z_3 < \lambda_2\sqrt{\frac{p}{n}}\} \right] \\
& \geq c_1(\lambda_1, \lambda_2) \cdot \mathbb{E}_{Z_2}[(Z_3 - Z_2) \cdot \mathbb{1}\{-\lambda_2\sqrt{\frac{p}{n}} < Z_2 < Z_3 < \lambda_2\sqrt{\frac{p}{n}}\}] \\
& \geq c_1(\lambda_1, \lambda_2)c_2\lambda_2 \cdot \sqrt{\frac{p}{n}}.
\end{aligned}$$

Since $p \lesssim n$, we have $c_1(\lambda_1, \lambda_2) \rightarrow \infty$ when $\lambda_1, \lambda_2 \rightarrow 0$. Therefore, we can choose λ_1, λ_2 to be sufficiently small such that $c_1(\lambda_1, \lambda_2)c_2\lambda_2\sqrt{\frac{p}{n}} \geq \frac{\lambda_2}{2\sqrt{2}}\sqrt{\frac{p}{n}}$. This completes the proof.

Proof of Theorem 2

At first we construct the following least favorable subset, which characterizes the difficulty of the general QDA problem. For simplicity of notation, we use the letters λ_1, λ_2 in this section, whose values are different from those in Section 3.7.1.

Since the KL -divergence and ℓ_2 norm are invariant to translations and orthogonal transformations, without loss of generality, we assume that $\boldsymbol{\mu}_1^* = -\boldsymbol{\mu}_2^* = \lambda_1 \mathbf{e}_1 + \tilde{\lambda}_1 \mathbf{e}_2$ for some constants $\lambda_1, \tilde{\lambda}_1 > 0$ whose values are determined later, with $2\sqrt{\lambda_1^2 + \tilde{\lambda}_1^2} = \|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_2$. In addition, we assume that $p/4$ is an integer.

Now let's consider

$$\Theta_2 = \{\boldsymbol{\theta}_{\mathbf{u}} = (1/2, 1/2, \lambda_1 \mathbf{e}_1 + \tilde{\lambda}_1 \mathbf{e}_2, -\lambda_1 \mathbf{e}_1 - \tilde{\lambda}_1 \mathbf{e}_2, \Sigma_1^{\mathbf{u}}, \Sigma_2) : \\ \Sigma_1^{\mathbf{u}} = (I_p + \tilde{\lambda}_2 E_{2,2} + \frac{\lambda_2}{\sqrt{n}} \sum_{i=3}^{p/2} u_i E_{i,i})^{-1}, \mathbf{u} \in \mathcal{A}_{p,p/4}, \Sigma_2 = I_p + \tilde{\lambda}_2 E_{2,2}\},$$

where $\mathcal{A}_{p,p/4}$ is defined in Lemma 16 .

According to Lemma 16, there is a subset of Θ_1 with logarithm cardinality being of order p , such that for any $\boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{u}'}$ in this subset, we have $\rho_H(\mathbf{u}, \mathbf{u}') \geq p/8$. We are going to apply Lemma 15 to this subset to complete the proof of Theorem 2.

At first we note that for two multivariate normal distribution $N_p(\boldsymbol{\mu}_1^*, \Sigma_1^{\mathbf{u}})$ and $N_p(\boldsymbol{\mu}_1^*, \Sigma_1^{\mathbf{u}'})$, using the fact that $\log(1+x) \asymp x - x^2/2 + o(x^2)$ for $x = o(1)$, the KL divergence between them are upper bounded by

$$\begin{aligned} KL &= \frac{1}{2} \left[\log \frac{|\Sigma_1^{\mathbf{u}'}|}{|\Sigma_1^{\mathbf{u}}|} - p + \text{tr}((\Sigma_1^{\mathbf{u}'})^{-1} \Sigma_1^{\mathbf{u}}) \right] \\ &= \frac{1}{2} \left[\sum_{i=3}^p \log \frac{1 + \frac{\lambda_2}{\sqrt{n}} u'_i}{1 + \frac{\lambda_2}{\sqrt{n}} u_i} - \rho_H(\mathbf{u}, \mathbf{u}') + \sum_{i=3}^p \frac{1 + \frac{\lambda_2}{\sqrt{n}} u_i}{1 + \frac{\lambda_2}{\sqrt{n}} u'_i} \right] \\ &= \frac{1}{2} \left[- \sum_{i=3}^p \log \left(1 + \frac{\frac{\lambda_2}{\sqrt{n}} (u_i - u'_i)}{1 + \frac{\lambda_2}{\sqrt{n}} u'_i} \right) + \sum_{i=3}^p \frac{\frac{\lambda_2}{\sqrt{n}} (u_i - u'_i)}{1 + \frac{\lambda_2}{\sqrt{n}} u'_i} \right] \\ &= \frac{1}{4} \sum_{i=3}^p \frac{1}{n} (u_i - u'_i)^2 + o\left(\frac{p}{n}\right) \leq \frac{\lambda_2^2 p}{16n} + o\left(\frac{p}{n}\right) \leq \frac{\lambda_2^2 p}{8n}. \end{aligned}$$

Therefore we have $KL(\mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}}}, \mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}'}}) \leq \lambda_2^2 p / (8n)$. To use Lemma 15 to prove Theorem 2, we

further need to show that for any $\boldsymbol{\theta}_u, \boldsymbol{\theta}_{u'}$,

$$[R_{\boldsymbol{\theta}}(G) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}_u}^*)] + [R_{\boldsymbol{\theta}}(G) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}_{u'}}^*)] \gtrsim \frac{p}{n}.$$

By Lemma 1 and 13,

$$\begin{aligned} & [R_{\boldsymbol{\theta}}(G) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}_u}^*)] + [R_{\boldsymbol{\theta}}(G) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}_{u'}}^*)] \\ & \geq L_{\boldsymbol{\theta}_u}^2(\hat{G}) + L_{\boldsymbol{\theta}_{u'}}^2(\hat{G}) \geq \frac{1}{2}(L_{\boldsymbol{\theta}_u}(\hat{G}) + L_{\boldsymbol{\theta}_{u'}}(\hat{G}))^2 \geq \frac{1}{2}(L_{\boldsymbol{\theta}_u}(G_{\boldsymbol{\theta}_{u'}}^*) - \sqrt{\frac{KL(\mathbb{P}_{\boldsymbol{\theta}_u}, \mathbb{P}_{\boldsymbol{\theta}_{u'}})}{2}})^2. \end{aligned}$$

Since now that $KL(\mathbb{P}_{\boldsymbol{\theta}_u}, \mathbb{P}_{\boldsymbol{\theta}_{u'}}) \leq \lambda_2^2 \frac{p}{8n}$, it's then sufficient to show $L_{\boldsymbol{\theta}_u}(G_{\boldsymbol{\theta}_{u'}}^*) \geq c\sqrt{\frac{p}{n}}$ for some $c > \lambda_2/4$.

Recall that

$$G_{\boldsymbol{\theta}}^*(z) = \mathbb{1}\{(z - \boldsymbol{\mu}_1)^\top D(z - \boldsymbol{\mu}_1) - 2\boldsymbol{\delta}^\top \Omega_2(z - \boldsymbol{\mu}_1) + \boldsymbol{\delta}^\top \Omega_2 \boldsymbol{\delta} - \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) > 0\},$$

where $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, $D = \Omega_2 - \Omega_1$.

Without loss of generality, we assume that $u_i = u'_i = 1$ when $i = 3, \dots, m_1$, $u_i = 1 - u'_i = 1$ when $i = m_1 + 1, \dots, m_2$, $u_i = 1 - u'_i = 0$ when $i = m_2 + 1, \dots, m_3$ and $u_i = u'_i = 0$ when $i = m_3 + 1, \dots, p$.

Then with a little abuse of notation, we have $\mathbf{z} \sim \frac{1}{2}N_p(\boldsymbol{\mu}_1, \Sigma_1^u) + \frac{1}{2}N_p(\boldsymbol{\mu}_2, \Sigma_2)$ with $\boldsymbol{\mu}_1 \mathbf{1} - \boldsymbol{\mu}_2 = \lambda_1 \mathbf{e}_1 + \tilde{\lambda}_1 \mathbf{e}_2$. Using the fact that $\log(1 + \frac{\lambda_2}{\sqrt{n}}) = \frac{\lambda_2}{\sqrt{n}} - \frac{\lambda_2^2}{2n} + o(\frac{1}{n})$, we have

$$G_{\boldsymbol{\theta}_u}^*(z) = 1 + \mathbb{1}\left\{\frac{\lambda_2}{\sqrt{n}} \left(\sum_{i=3}^{m_1} (z_i^2 - 1) + \sum_{i=m_1+1}^{m_2} (z_i^2 - 1) \right) + 4\lambda_1 z_1 + 4\frac{\tilde{\lambda}_1}{1 + \tilde{\lambda}_2} z_2 + \frac{p}{8n} + o\left(\frac{p}{n}\right) > 0\right\},$$

and

$$G_{\boldsymbol{\theta}_{\mathbf{u}'}}^*(\mathbf{z}) = 1 + \mathbb{1}\left\{\frac{\lambda_2}{\sqrt{n}} \left(\sum_{i=3}^{m_1} (z_i^2 - 1) + \sum_{i=m_2+1}^{m_3} (z_i^2 - 1) \right) + 4\lambda_1 z_1 + 4\frac{\tilde{\lambda}_1}{1 + \tilde{\lambda}_2} z_2 + \frac{p}{8n} + o\left(\frac{p}{n}\right) > 0\right\}.$$

Let $Z_1 = -(4\lambda_1 z_1 + 4\frac{\tilde{\lambda}_1}{1 + \tilde{\lambda}_2} z_2 + \frac{\lambda_2}{\sqrt{n}} \sum_{i=3}^{m_1} (z_i^2 - 1) + \frac{p}{8n})$, $Z_2 = \frac{\lambda_2}{\sqrt{n}} \sum_{i=m_1+1}^{m_2} (z_i^2 - 1)$, $Z_3 = \frac{\lambda_2}{\sqrt{n}} \sum_{i=m_2+1}^{m_3} (z_i^2 - 1)$, then

$$G_{\boldsymbol{\theta}_{\mathbf{u}}}^*(\mathbf{z}) = \mathbb{1}\{-Z_1 + Z_2 + o\left(\frac{p}{n}\right) > 0\} \text{ and } G_{\boldsymbol{\theta}_{\mathbf{u}'}}^*(\mathbf{z}) = \mathbb{1}\{-Z_1 + Z_3 + o\left(\frac{p}{n}\right) > 0\},$$

and

$$\begin{aligned} L_{\boldsymbol{\theta}_{\mathbf{u}}}(G_{\boldsymbol{\theta}_{\mathbf{u}'}}^*) &= \mathbb{P}_{\boldsymbol{\theta}_{\mathbf{u}}}(G_{\boldsymbol{\theta}_{\mathbf{u}'}}^*(\mathbf{z}) \neq G_{\boldsymbol{\theta}_{\mathbf{u}}}^*(\mathbf{z})) \\ &\geq \frac{1}{2} \mathbb{P}_{\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1^{\mathbf{u}})} \left(Z_2 + o\left(\frac{p}{n}\right) \leq Z_1 \leq Z_3 + o\left(\frac{p}{n}\right) \right) \\ &\quad + \frac{1}{2} \mathbb{P}_{\mathbf{z} \sim N_p(\boldsymbol{\mu}_2, \Sigma_2)} \left(Z_3 + o\left(\frac{p}{n}\right) \leq Z_1 \leq Z_2 + o\left(\frac{p}{n}\right) \right) \\ &\geq \frac{1}{2} \mathbb{P}_{\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_2)} (Z_2 \leq Z_1 \leq Z_3) + o\left(\frac{p}{n}\right). \end{aligned}$$

By central limit theorem, $\frac{\sqrt{n}}{\lambda_2 \sqrt{m_2 - m_1}} Z_2$, $\frac{\sqrt{n}}{\lambda_2 \sqrt{m_3 - m_2}} Z_3$ converges to the standard normal distribution $N(0, 1)$. Since $m_3 - m_2 = \rho_H(\mathbf{u}, \mathbf{u}') \geq p/8$, and $\limsup_{n, p \rightarrow \infty} \frac{p}{n} \leq C_1$, similar as the derivation in Section 3.7.1, there exists a constant c_2 , such that n, p are sufficiently large,

$$\begin{aligned} &\mathbb{E}[(Z_3 - Z_2) \cdot \mathbb{1}\{-\lambda_2 \sqrt{\frac{p}{n}} < Z_2 < Z_3 < \lambda_2 \sqrt{\frac{p}{n}}\}] \\ &\geq \mathbb{E}[(Z_3 - Z_2) \cdot \mathbb{1}\{-\lambda_2 \sqrt{\frac{p}{n}} < Z_2 < -\frac{\lambda_2}{2} \sqrt{\frac{p}{n}}, \frac{\lambda_2}{2} \sqrt{\frac{p}{n}} < Z_3 < \lambda_2 \sqrt{\frac{p}{n}}\}] \\ &\geq \lambda_2 \sqrt{\frac{p}{n}} \cdot \mathbb{P}(-\lambda_2 \sqrt{\frac{p}{n}} < Z_2 < -\frac{\lambda_2}{2} \sqrt{\frac{m_2 - m_1}{n}}) \cdot \mathbb{P}(\frac{\lambda_2}{2} \sqrt{\frac{m_3 - m_1}{n}} < Z_3 < \lambda_2 \sqrt{\frac{p}{n}}) \\ &\geq \lambda_2 \sqrt{\frac{p}{8n}} \cdot \mathbb{P}_{Z \sim N(0,1)}(-\sqrt{\frac{p}{m_2 - m_1}} < Z < -\frac{1}{2}) \cdot \mathbb{P}_{Z \sim N(0,1)}(\frac{1}{2} < Z < \sqrt{\frac{p}{m_3 - m_2}}) \\ &\geq \lambda_2 \sqrt{\frac{p}{8n}} \cdot \mathbb{P}_{Z \sim N(0,1)}(-\sqrt{2} < Z < -\frac{1}{2}) \cdot \mathbb{P}_{Z \sim N(0,1)}(\frac{1}{2} < Z < \sqrt{2}) \geq c_2 \lambda_2 \sqrt{\frac{p}{n}}. \end{aligned}$$

Similar to that in Section 3.7.1, let's denote the probability density function of Z_1 by f . Use central limit theorem again, when $\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_2)$, $p \lesssim n$, and n, p are sufficiently large, $Z_1 \approx N(-4\lambda_1^2 - \frac{4\tilde{\lambda}_1^2}{1+\tilde{\lambda}_2} + \frac{p}{8n}, \lambda_1^2 + \frac{\tilde{\lambda}_1^2}{1+\tilde{\lambda}_2} + \frac{2(m_1-2)\lambda_2^2}{n})$ if $m_1 \rightarrow \infty$. Therefore, there exists constant $c_1(\lambda_1, \tilde{\lambda}_1, \lambda_2, \tilde{\lambda}_2)$, such that $\inf_{|x| < \lambda_2 \sqrt{p/n}} f(x) > c_1(\lambda_1, \tilde{\lambda}_1, \lambda_2, \tilde{\lambda}_2)$, and $c_1(\lambda_1, \tilde{\lambda}_1, \lambda_2, \tilde{\lambda}_2)$ goes to infinity when $\lambda_1, \lambda_2 \rightarrow 0, \tilde{\lambda}_2 \rightarrow \infty$, and $\tilde{\lambda}_1$ is chosen such that $\sqrt{\lambda_1^2 + \tilde{\lambda}_1^2} = \|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_2/2$.

$$\begin{aligned}
& \mathbb{P}_{\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_2)} (Z_2 \leq Z_1 \leq Z_3) \geq \mathbb{P}_{\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_2)} \left(Z_2 \leq Z_1 \leq Z_3, -\lambda_2 \sqrt{\frac{p}{n}} < Z_2 < Z_3 < \lambda_2 \sqrt{\frac{p}{n}} \right) \\
& = \mathbb{E}_{Z_2} \left[\int_{Z_2}^{Z_3} f(z_1) dz_1 \cdot \mathbb{1}\{-\lambda_2 \sqrt{\frac{p}{n}} < Z_2 < Z_3 < \lambda_2 \sqrt{\frac{p}{n}}\} \right] \\
& \geq c_1(\lambda_1, \tilde{\lambda}_1, \lambda_2, \tilde{\lambda}_2) \cdot \mathbb{E}_{Z_2} [(Z_3 - Z_2) \cdot \mathbb{1}\{-\lambda_2 \sqrt{\frac{p}{n}} < Z_2 < Z_3 < \lambda_2 \sqrt{\frac{p}{n}}\}] \\
& \geq c_1(\lambda_1, \tilde{\lambda}_1, \lambda_2, \tilde{\lambda}_2) c_2 \lambda_2 \cdot \sqrt{\frac{p}{n}}.
\end{aligned}$$

Therefore, by choosing sufficiently small λ_1, λ_2 and large $\tilde{\lambda}_2$ (doesn't depend on n, p), we have $c_2 c_1(\lambda_1, \tilde{\lambda}_1, \lambda_2, \tilde{\lambda}_2) \cdot \lambda_2 \sqrt{\frac{p}{n}} \geq \frac{\lambda_2}{4} \sqrt{\frac{p}{n}}$. \square

3.7.2. Proof of the Theorem 3

To prove Theorem 3 we begin by collecting a few important technical lemmas that will be used in the main proofs.

Auxiliary Lemmas

Lemma 5. *Suppose $\mathbf{X}_1, \dots, \mathbf{X}_n$ i.i.d. $\sim N_p(\boldsymbol{\mu}, \Sigma)$, and assume that $\hat{\boldsymbol{\mu}}, \hat{\Sigma}$ are the sample mean and sample covariance matrix respectively. Let $\Gamma(s; p) = \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}\|_2 = 1, \|\mathbf{u}_{S^c}\|_1 \leq \|\mathbf{u}_S\|_1, \text{ for some } S \subset [p] \text{ with } |S| = s\}$, then with probability at least $1 - p^{-1}$,*

$$\begin{aligned}
& \sup_{\mathbf{u} \in \Gamma(s; p)} \mathbf{u}^\top (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \lesssim \sqrt{\frac{s \log p}{n}}; \\
& \sup_{\mathbf{u}, \mathbf{v} \in \Gamma(s; p)} \mathbf{u}^\top (\hat{\Sigma} - \Sigma) \mathbf{v} \lesssim \sqrt{\frac{s \log p}{n}}; \quad \sup_{\mathbf{a} \in \Gamma(s; p^2)} \mathbf{a}^\top \text{vec}(\hat{\Sigma} - \Sigma) \lesssim \sqrt{\frac{s \log p}{n}}.
\end{aligned}$$

Lemma 6. Suppose $\mathbf{X}_1, \dots, \mathbf{X}_{n_1}$ i.i.d. $\sim N_p(\boldsymbol{\mu}_1, \Sigma_1)$, $\mathbf{Y}_1, \dots, \mathbf{Y}_{n_2}$ i.i.d. $\sim N_p(\boldsymbol{\mu}_2, \Sigma_2)$, $n = \min(n_1, n_2)$ and assume that $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\Sigma}_1, \hat{\Sigma}_2$ are the sample means and sample covariance matrices. Denote $V = \frac{1}{2}\Sigma_1 \otimes \Sigma_2 + \frac{1}{2}\Sigma_2 \otimes \Sigma_1$ and $\hat{V} = \frac{1}{2}\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 + \frac{1}{2}\hat{\Sigma}_2 \otimes \hat{\Sigma}_1$. Assume that $\boldsymbol{\beta} = \Omega_2(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ and $\text{vec}(D)$ has bounded ℓ_2 norm, then with probability at least $1 - p^{-1}$,

$$\|\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_\infty \lesssim \sqrt{\frac{\log p}{n}}, \quad \|(\hat{\Sigma}_k - \Sigma_k)\boldsymbol{\beta}\|_\infty \lesssim \sqrt{\frac{\log p}{n}}, \quad k = 1, 2;$$

$$\|\text{vec}(\hat{\Sigma} - \Sigma)\|_\infty \lesssim \sqrt{\frac{\log p}{n}}; \quad \|(\hat{V} - V)\text{vec}(D)\|_\infty \lesssim \sqrt{\frac{\log p}{n}}.$$

Lemma 7. Suppose $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. Let $\mathbf{h} = \mathbf{x} - \mathbf{y}$. Denote $\mathcal{S} = \text{supp}(\mathbf{y})$ and $s = |\mathcal{S}|$. If $\|\mathbf{x}\|_1 \leq \|\mathbf{y}\|_1$, then $\mathbf{h} \in \Gamma(s; p)$, that is,

$$\|\mathbf{h}_{\mathcal{S}^c}\|_1 \leq \|\mathbf{h}_{\mathcal{S}}\|_1.$$

Lemma 8. For any two matrices $A, B \in \mathbb{R}^{p \times p}$ with non-negative eigenvalues,

$$|\log |A| - \log |B|| \leq \max\{|\text{tr}(B^{-1}(A - B))|, |\text{tr}(A^{-1}(B - A))|\}.$$

Main proofs

We prove the consistency of estimation of D first. The consistency of estimating $\boldsymbol{\beta}$ can be derived similarly.

Recall that

$$\hat{D} = \arg \min_{D \in \mathbb{R}^{p \times p}} \left\{ |D|_1 : \left\| \left(\frac{1}{2}\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 + \frac{1}{2}\hat{\Sigma}_2 \otimes \hat{\Sigma}_1 \right) \text{vec}(D) - \text{vec}(\hat{\Sigma}_1) + \text{vec}(\hat{\Sigma}_2) \right\|_\infty \leq \lambda_{1,n} \right\}. \quad (3.12)$$

By Lemma 6, D is a feasible solution to (3.12) with $\lambda_{1,n} = c_1 \sqrt{\frac{\log p}{n}}$ when c_1 is a sufficiently large constant. Then using Lemma 7, we have $\text{vec}(D - \hat{D}) \in \Gamma(s_1; p^2)$.

Denote $V = \frac{1}{2}\Sigma_1 \otimes \Sigma_2 + \frac{1}{2}\Sigma_2 \otimes \Sigma_1$, $\mathbf{v}_\Sigma = \text{vec}(\Sigma_1) - \text{vec}(\Sigma_2)$ and $\hat{V} = \frac{1}{2}\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 + \frac{1}{2}\hat{\Sigma}_2 \otimes \hat{\Sigma}_1$, $\hat{\mathbf{v}}_\Sigma = \text{vec}(\hat{\Sigma}_1) - \text{vec}(\hat{\Sigma}_2)$.

We have

$$\begin{aligned} V \text{vec}(D) &= (\frac{1}{2}\Sigma_1 \otimes \Sigma_2 + \frac{1}{2}\Sigma_2 \otimes \Sigma_1) \text{vec}(D) = \text{vec}(\frac{1}{2}\Sigma_1 D \Sigma_2 + \frac{1}{2}\Sigma_2 D \Sigma_1) \\ &= \text{vec}(\Sigma_1 - \Sigma_2) = \mathbf{v}_\Sigma. \end{aligned}$$

In addition, over the parameter space $\Theta_p(s_1, s_2)$,

$$\|V^{-1}\|_2 = \|\Omega_1 \otimes \Omega_2\|_2 = \|\Omega_1\|_2 \cdot \|\Omega_2\|_2 \leq M_1^2.$$

which is followed by $\lambda_{\min}(V) \geq M_1^{-2}$.

As a consequence, by Lemma 5, with probability at least $1 - 3p^{-1}$,

$$\begin{aligned} & |(\text{vec}(\hat{D}) - \text{vec}(D))^\top V(\text{vec}(\hat{D}) - \text{vec}(D))| \\ & \leq |(\text{vec}(\hat{D}) - \text{vec}(D))^\top (\hat{V} \text{vec}(\hat{D}) - \hat{\mathbf{v}}_\Sigma)| + |(\text{vec}(\hat{D}) - \text{vec}(D))^\top (\hat{V} - V) \text{vec}(\hat{D})| \\ & \quad + |(\text{vec}(\hat{D}) - \text{vec}(D))^\top (\mathbf{v}_\Sigma - \hat{\mathbf{v}}_\Sigma)| \\ & \lesssim \sqrt{s_1} \|\text{vec}(\hat{D}) - \text{vec}(D)\|_2 \cdot \|\hat{V} \text{vec}(\hat{D}) - \hat{\mathbf{v}}_\Sigma\|_\infty \\ & \quad + \|\text{vec}(\hat{D}) - \text{vec}(D)\|_2 \cdot \sqrt{\frac{s_1 \log p}{n}} \cdot \|\text{vec}(D) - \text{vec}(\hat{D})\|_2 \\ & \quad + \|\text{vec}(\hat{D}) - \text{vec}(\hat{D})\|_2 \sqrt{\frac{s_1 \log p}{n}} \cdot \|\text{vec}(D)\|_2 + \|\text{vec}(D) - \text{vec}(\hat{D})\|_2 \sqrt{\frac{s_1 \log p}{n}}. \end{aligned} \tag{3.13}$$

In addition, since $|(\text{vec}(\hat{D}) - \text{vec}(D))^\top V(\text{vec}(\hat{D}) - \text{vec}(D))| \geq \lambda_{\min}(V) \|\text{vec}(\hat{D}) - \text{vec}(D)\|_2^2 \geq M_1^{-2} \|\text{vec}(\hat{D}) - \text{vec}(D)\|_2^2$, we then have

$$\|D - \hat{D}\|_F = \|\text{vec}(\hat{D}) - \text{vec}(D)\|_2 \lesssim \sqrt{\frac{s_1 \log p}{n}}.$$

The estimation error of $\boldsymbol{\beta}$ can be derived similarly. By Lemma 6, $\boldsymbol{\beta}$ is a feasible solution to (3.6) with $\lambda_{2,n} = c_2 \sqrt{\frac{\log p}{n}}$ when c_2 is sufficiently large. Then using Lemma 7, we have $\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \in \Gamma(s_2; p)$.

Then with probability at least $1 - 3p^{-1}$,

$$\begin{aligned}
& |(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma_2 (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})| \\
& \leq |(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\Sigma}_2 \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\delta}})| + |(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\Sigma}_2 - \Sigma_2) \hat{\boldsymbol{\beta}}| + |(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\boldsymbol{\delta} - \hat{\boldsymbol{\delta}})| \\
& \lesssim \sqrt{s_2} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \cdot \|\hat{\Sigma}_2 \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\delta}}\|_\infty + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \cdot \sqrt{\frac{s_2 \log p}{n}} \cdot \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2 \\
& \quad + \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2 \sqrt{\frac{s_2 \log p}{n}} \cdot \|\boldsymbol{\beta}\|_2 + \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2 \sqrt{\frac{s_2 \log p}{n}}.
\end{aligned} \tag{3.14}$$

Similarly, since $\lambda_{\min}(\Sigma_2) \geq M_1^{-1}$, we have with probability at least $1 - p^{-1}$,

$$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2 \lesssim \sqrt{\frac{s_2 \log p}{n}}.$$

3.7.3. Proof of Theorem 4

We note here that the notation c, C denote generic constants and their values might vary line by line. Recall that the QDA rule is

$$1 + \mathbb{1}\{(\mathbf{z} - \boldsymbol{\mu}_1)^\top D(\mathbf{z} - \boldsymbol{\mu}_1) - 2\boldsymbol{\beta}^\top (\mathbf{z} - \bar{\boldsymbol{\mu}}) - \log(|D\Sigma_1 + I_p|) + 2\log\left(\frac{\pi_1}{\pi_2}\right) > 0\}.$$

Let $\bar{\boldsymbol{\mu}} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, $Q(\mathbf{z}) = (\mathbf{z} - \boldsymbol{\mu}_1)^\top D(\mathbf{z} - \boldsymbol{\mu}_1) - 2\boldsymbol{\beta}^\top (\mathbf{z} - \bar{\boldsymbol{\mu}}) - \log(|D\Sigma_1 + I_p|) + 2\log\left(\frac{\pi_1}{\pi_2}\right)$, $\hat{Q}(\mathbf{z}) = (\mathbf{z} - \hat{\boldsymbol{\mu}}_1)^\top \hat{D}(\mathbf{z} - \hat{\boldsymbol{\mu}}_1) - 2\hat{\boldsymbol{\beta}}^\top (\mathbf{z} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}) - \log(|\hat{D}\hat{\Sigma}_1 + I_p|) + \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right)$, and $M(\mathbf{z}) = Q(\mathbf{z}) - \hat{Q}(\mathbf{z})$, we are going to show that there exist some constants $c, C > 0$, such that for any $M > 0$,

$$\mathbb{P}_{\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)} \left(|M(\mathbf{z})| > M \sqrt{\frac{(s_1 + s_2) \log p}{n}} \right) \leq e^{-cM} + Cp^{-1},$$

note that the above probability is taken with respect to the random samples $\mathbf{X}_1, \dots, \mathbf{X}_{n_1}$ *i.i.d.* $\sim N_p(\boldsymbol{\mu}_1, \Sigma_1)$, $\mathbf{Y}_1, \dots, \mathbf{Y}_{n_2}$ *i.i.d.* $\sim N_p(\boldsymbol{\mu}_2, \Sigma_2)$, and $\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)$. We will later see how we reduce the mixed distribution of the test sample to the single distribution when we calculate the classification error.

Rewrite the QDA rule as

$$\mathbb{1}\{(\mathbf{z} - \boldsymbol{\mu}_1)^\top D(\mathbf{z} - \boldsymbol{\mu}_1) - 2\boldsymbol{\beta}^\top(\mathbf{z} - \boldsymbol{\mu}_1) + \boldsymbol{\beta}^\top(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - \log(|D\Sigma_1 + I_p|) + 2\log\left(\frac{\pi_1}{\pi_2}\right) > 0\}.$$

We firstly bound the estimation error of the constant term $\boldsymbol{\beta}^\top(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. We have with probability at least $1 - p^{-1}$,

$$\begin{aligned} |\boldsymbol{\beta}^\top(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - \hat{\boldsymbol{\beta}}^\top(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)| &\leq |\hat{\boldsymbol{\beta}}^\top(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1)| + \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\|_2 \\ &\leq \|\hat{\boldsymbol{\beta}}\|_1 \cdot \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1\|_\infty + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_2 \\ &\leq \|\boldsymbol{\beta}\|_1 \cdot \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1\|_\infty + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_2 \\ &\leq \sqrt{s_2} \|\boldsymbol{\beta}\|_2 \cdot \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1\|_\infty + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_2 \lesssim \sqrt{\frac{s_2 \log p}{n}}. \end{aligned}$$

For $\log |D\Sigma_1 + I_p|$, notice that $D\Sigma_1 + I_p = \Omega_2 \Sigma_1$ and the product of two positive semidefinite and symmetric matrices has non-negative eigenvalues, followed by $(D\Sigma_1 + I_p)^{-1} = \Omega_1 \Sigma_2 = (\Omega_2 - D)\Sigma_2 = I_p - D\Sigma_2$, then

$$\begin{aligned} \log |D\Sigma_1 + I_p| - \log |\hat{D}\hat{\Sigma}_1 + I_p| &\leq \text{tr}((D\Sigma_1 + I_p)^{-1}(D\Sigma_1 - \hat{D}\hat{\Sigma}_1)) \\ &= \text{tr}((-D\Sigma_2 + I_p)(D\Sigma_1 - \hat{D}\hat{\Sigma}_1)) \\ &= \text{tr}((-D\Sigma_2)(D\Sigma_1 - \hat{D}\hat{\Sigma}_1)) + \text{tr}(D\Sigma_1 - \hat{D}\hat{\Sigma}_1) \\ &\leq \|D\Sigma_2\|_F \cdot \|D\Sigma_1 - \hat{D}\hat{\Sigma}_1\|_F + \text{tr}(D\Sigma_1 - \hat{D}\hat{\Sigma}_1) \\ &\leq \|D\|_F \|\Sigma_2\|_2 \cdot \|D\Sigma_1 - \hat{D}\hat{\Sigma}_1\|_F + \text{tr}(D\Sigma_1 - \hat{D}\hat{\Sigma}_1) \\ &\leq \|D\|_F \|\Sigma_2\|_2 \cdot \|D\Sigma_1 - \hat{D}\hat{\Sigma}_1\|_F + |\text{tr}(\hat{D}\hat{\Sigma}_1 - D\Sigma_1)| + \text{tr}(D\Sigma_1 - \hat{D}\hat{\Sigma}_1). \end{aligned} \quad (3.15)$$

In addition, with probability at least $1 - p^{-1}$,

$$\begin{aligned}
& \|D\Sigma_1 - \hat{D}\hat{\Sigma}_1\|_F \leq \|D\Sigma_1 - \hat{D}\Sigma_1\|_F + \|\hat{D}(\Sigma_1 - \hat{\Sigma}_1)\|_F \\
& \leq \|\Sigma_1\|_2 \|D - \hat{D}\|_F + \|\Sigma_1 - \hat{\Sigma}_1\|_{2,s_1} \|\hat{D}\|_F \\
& \lesssim \sqrt{\frac{s_1 \log p}{n}} + \|\Sigma_1 - \hat{\Sigma}_1\|_{2,s_1} (\|D\|_F + \sqrt{\frac{s_1 \log p}{n}}) \\
& \leq \sqrt{\frac{s_1 \log p}{n}} + \sqrt{\frac{s_1 \log p}{n}} (\|D\|_F + \sqrt{\frac{s_1 \log p}{n}}) \lesssim \sqrt{\frac{s_1 \log p}{n}},
\end{aligned}$$

where $\|\Sigma_1 - \hat{\Sigma}_1\|_{2,s_1}$ is defined as

$$\|\Sigma_1 - \hat{\Sigma}_1\|_{2,s_1} := \sup_{\|\mathbf{u}\|_0 \leq s_1, \|\mathbf{u}\|_2 = 1} \|(\Sigma_1 - \hat{\Sigma}_1)\mathbf{u}\|_2 \lesssim \sqrt{\frac{s_1 \log p}{n}},$$

where the last inequality is similarly proved as Lemma 5, by using the packing number argument.

In addition, with probability at least $1 - p^{-1}$,

$$|\text{tr}(\hat{D}\Sigma_1 - \hat{D}\hat{\Sigma}_1)| \leq \sqrt{s_1} \|\Sigma_1 - \hat{\Sigma}_1\|_\infty \|\hat{D}\|_F \lesssim \sqrt{\frac{s_1 \log p}{n}}.$$

There is still a remaining term $\text{tr}(D\Sigma_1 - \hat{D}\Sigma_1)$ in (3.15), we will leave it there and use it when we derive the distribution of the term involving \mathbf{z} . The other direction, the upper bound of $\text{tr}(D\Sigma_1 - \hat{D}\Sigma_1) - (\log |D\Sigma_1 + I_p| - \log |\hat{D}\hat{\Sigma}_1 + I_p|)$, can be derived similarly. Therefore by symmetry, we have with probability at least $1 - p^{-1}$

$$\left| (\log |D\Sigma_1 + I_p| - \log |\hat{D}\hat{\Sigma}_1 + I_p|) - (\text{tr}(D\Sigma_1 - \hat{D}\Sigma_1)) \right| \lesssim \sqrt{\frac{s_1 \log p}{n}}.$$

For the term involving \mathbf{z} , when $\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)$, we have

$$\begin{aligned} & (\mathbf{z} - \boldsymbol{\mu}_1)^\top D(\mathbf{z} - \boldsymbol{\mu}_1) - (\mathbf{z} - \boldsymbol{\mu}_1)^\top \hat{D}(\mathbf{z} - \boldsymbol{\mu}_1) - (\text{tr}(D\Sigma_1 - \hat{D}\Sigma_1)) \\ &= (\mathbf{z} - \boldsymbol{\mu}_1)^\top (\hat{D} - D)(\mathbf{z} - \boldsymbol{\mu}_1) - (\text{tr}(D\Sigma_1 - \hat{D}\Sigma_1)) \\ &\stackrel{d}{=} \mathbf{z}_0^\top \Sigma_1^{1/2} (\hat{D} - D) \Sigma_1^{1/2} \mathbf{z}_0 - \text{tr}(\Sigma_1^{1/2} (\hat{D} - D) \Sigma_1^{1/2}) \stackrel{def}{=} \sum_{i=1}^p \lambda_i (z_{0i}^2 - 1), \end{aligned}$$

where λ_i 's are the eigenvalues of $\Sigma_1^{1/2} (\hat{D} - D) \Sigma_1^{1/2}$.

Since with probability at least $1 - p^{-1}$,

$$\sqrt{\sum_{i=1}^p \lambda_i^2} = \|\Sigma_1^{1/2} (\hat{D} - D) \Sigma_1^{1/2}\|_F \leq \|\Sigma_1\|_2 \|\hat{D} - D\|_F \lesssim \sqrt{\frac{s_1 \log p}{n}},$$

and with probability at least $1 - p^{-1}$,

$$\max_i |\lambda_i| \leq \|\Sigma_1^{1/2} (\hat{D} - D) \Sigma_1^{1/2}\|_2 \leq \|\Sigma_1\|_2 \|\hat{D} - D\|_2 \lesssim \sqrt{\frac{s_1 \log p}{n}},$$

by Bernstein type inequality for sub-exponential random variables, see Vershynin (2011), we have for some $\tilde{c}_1 > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^p \lambda_i (z_{0i}^2 - 1)\right| \geq t\right) \leq 2 \exp\left\{-\tilde{c}_1 \min\left\{\frac{t^2}{s_1 \log p/n}, \frac{t}{\sqrt{s_1 \log p/n}}\right\}\right\},$$

which implies that for some $c_1 > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^p \lambda_i (z_{0i}^2 - 1)\right| \geq M \sqrt{\frac{s_1 \log p}{n}}\right) \leq e^{-c_1 M} + Cp^{-1}.$$

For $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{z}$, when $\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)$, we have

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{z} \sim N((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{\mu}_1, (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma_1 (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})).$$

Since with probability at least $1 - p^{-1}$,

$$|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{\mu}_1| \leq \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \cdot \|\boldsymbol{\mu}_1\|_2 \lesssim \sqrt{\frac{s_2 \log p}{n}},$$

and with probability at least $1 - p^{-1}$,

$$|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma_1 (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})| \leq \|\Sigma_1\|_2 \cdot \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leq \frac{s_2 \log p}{n},$$

we have for some $c_2 > 0$,

$$\mathbb{P}(|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{z}| > M \sqrt{\frac{s_2 \log p}{n}}) \leq e^{-c_2 M^2} + Cp^{-1}.$$

Lastly,

$$|2 \log\left(\frac{\pi_1}{\pi_2}\right) - \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right)| \lesssim |\hat{\pi}_1 - \pi_1| + |\hat{\pi}_2 - \pi_2|.$$

and by Hoeffding inequality, for $k \in [2]$, there are some constant $c_H > 0$, such that

$$\mathbb{P}(|\hat{\pi}_k - \pi_k| > t) \leq \exp(-c_H \cdot nt^2).$$

We have for some constant $c, M_H > 0$,

$$\mathbb{P}(|2 \log\left(\frac{\pi_1}{\pi_2}\right) - \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right)| > M_H \sqrt{\frac{1}{n}}) \leq e^{-cM_H}.$$

Therefore, there exists some $c > 0$, such that for any $M > 0$,

$$\mathbb{P}_{\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)}(M(\mathbf{z}) > M \sqrt{\frac{(s_1 + s_2) \log p}{n}}) \leq e^{-cM} + Cp^{-1}.$$

Then it follows that

$$\begin{aligned}
& R(\hat{G}_{\text{SDAR}}) - R_{\theta}(G_{\theta}^*) \\
&= \frac{1}{2} \int_{Q(\mathbf{z}) > 0} \frac{\pi_1}{(2\pi)^{p/2} |\Sigma_1|^{1/2}} e^{-1/2 \cdot (\mathbf{z} - \boldsymbol{\mu}_1)^\top \Omega_1 (\mathbf{z} - \boldsymbol{\mu}_1)} d\mathbf{z} \\
&\quad + \frac{1}{2} \int_{Q(\mathbf{z}) \leq 0} \frac{\pi_2}{(2\pi)^{p/2} |\Sigma_2|^{1/2}} e^{-1/2 \cdot (\mathbf{z} - \boldsymbol{\mu}_2)^\top \Omega_2 (\mathbf{z} - \boldsymbol{\mu}_2)} d\mathbf{z} \\
&\quad - \frac{1}{2} \int_{\hat{Q}(\mathbf{z}) > 0} \frac{\pi_1}{(2\pi)^{p/2} |\Sigma_1|^{1/2}} e^{-1/2 \cdot (\mathbf{z} - \boldsymbol{\mu}_1)^\top \Omega_1 (\mathbf{z} - \boldsymbol{\mu}_1)} d\mathbf{z} \\
&\quad - \frac{1}{2} \int_{\hat{Q}(\mathbf{z}) \leq 0} \frac{\pi_2}{(2\pi)^{p/2} |\Sigma_2|^{1/2}} e^{-1/2 \cdot (\mathbf{z} - \boldsymbol{\mu}_2)^\top \Omega_2 (\mathbf{z} - \boldsymbol{\mu}_2)} d\mathbf{z}.
\end{aligned}$$

$$\begin{aligned}
& R(\hat{G}_{\text{SDAR}}) - R_{\theta}(G_{\theta}^*) \\
&= \frac{1}{2} \int_{Q(\mathbf{z}) > 0} \frac{1}{(2\pi)^{p/2}} e^{-1/2 \cdot (\mathbf{z} - \boldsymbol{\mu}_1)^\top \Omega_1 (\mathbf{z} - \boldsymbol{\mu}_1) - \log |\Sigma_1|/2 + \log \pi_1} \\
&\quad - \frac{1}{(2\pi)^{p/2}} e^{-1/2 \cdot (\mathbf{z} - \boldsymbol{\mu}_2)^\top \Omega_2 (\mathbf{z} - \boldsymbol{\mu}_2) - \log |\Sigma_2|/2 + \log \pi_2} d\mathbf{z} \\
&\quad - \frac{1}{2} \int_{\hat{Q}(\mathbf{z}) > 0} \frac{1}{(2\pi)^{p/2}} e^{-1/2 \cdot (\mathbf{z} - \boldsymbol{\mu}_1)^\top \Omega_1 (\mathbf{z} - \boldsymbol{\mu}_1) - \log |\Sigma_1|/2 + \log \pi_1} \\
&\quad - \frac{1}{(2\pi)^{p/2}} e^{-1/2 \cdot (\mathbf{z} - \boldsymbol{\mu}_2)^\top \Omega_2 (\mathbf{z} - \boldsymbol{\mu}_2) - \log |\Sigma_2|/2 + \log \pi_2} d\mathbf{z} \\
&= \frac{1}{2} \int_{Q(\mathbf{z}) > 0} \frac{1}{(2\pi)^{p/2}} e^{-1/2 \cdot (\mathbf{z} - \boldsymbol{\mu}_1)^\top \Omega_1 (\mathbf{z} - \boldsymbol{\mu}_1) - \log |\Sigma_1|/2} (1 - e^{-Q(\mathbf{z})}) d\mathbf{z} \\
&\quad - \frac{1}{2} \int_{\hat{Q}(\mathbf{z}) > 0} \frac{1}{(2\pi)^{p/2}} e^{-1/2 \cdot (\mathbf{z} - \boldsymbol{\mu}_1)^\top \Omega_1 (\mathbf{z} - \boldsymbol{\mu}_1) - \log |\Sigma_1|/2} (1 - e^{-Q(\mathbf{z})}) d\mathbf{z}
\end{aligned}$$

Then it follows

$$\begin{aligned}
& R(\hat{G}_{\text{SDAR}}) - R_{\theta}(G_{\theta}^*) \\
& \leq \frac{1}{2} \int_{Q(\mathbf{z}) > 0, \hat{Q}(\mathbf{z}) \leq 0} \frac{1}{(2\pi)^{p/2}} e^{-1/2 \cdot (\mathbf{z} - \boldsymbol{\mu}_1)^\top \Omega_1 (\mathbf{z} - \boldsymbol{\mu}_1) - \log |\Sigma_1|/2} (1 - e^{-Q(\mathbf{z})}) d\mathbf{z} \\
& = \frac{1}{2} \int_{Q(\mathbf{z}) > 0, Q(\mathbf{z}) \leq Q(\mathbf{z}) - \hat{Q}(\mathbf{z})} \frac{1}{(2\pi)^{p/2}} e^{-1/2 \cdot (\mathbf{z} - \boldsymbol{\mu}_1)^\top \Omega_1 (\mathbf{z} - \boldsymbol{\mu}_1) - \log |\Sigma_1|/2} (1 - e^{-Q(\mathbf{z})}) d\mathbf{z} \\
& = \frac{1}{2} \mathbb{E}_{\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)} [(1 - e^{-Q(\mathbf{z})}) \mathbb{1}\{0 < Q(\mathbf{z}) \leq M(\mathbf{z})\}] \\
& = \frac{1}{2} \mathbb{E}_{\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)} \left[(1 - e^{-Q(\mathbf{z})}) \mathbb{1}\{0 < Q(\mathbf{z}) \leq M(\mathbf{z})\} \cdot \mathbb{1}\{M(\mathbf{z}) < M \log n \sqrt{\frac{(s_1 + s_2) \log p}{n}}\} \right] \\
& \quad + \frac{1}{2} \mathbb{E}_{\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)} \left[(1 - e^{-Q(\mathbf{z})}) \mathbb{1}\{0 < Q(\mathbf{z}) \leq M(\mathbf{z})\} \cdot \mathbb{1}\{M(\mathbf{z}) \geq M \log n \sqrt{\frac{(s_1 + s_2) \log p}{n}}\} \right] \\
& \leq \frac{1}{2} \mathbb{E}_{\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)} \left[(1 - e^{-Q(\mathbf{z})}) \mathbb{1}\{0 < Q(\mathbf{z}) \leq M(\mathbf{z})\} \cdot \mathbb{1}\{M(\mathbf{z}) < M \log n \sqrt{\frac{(s_1 + s_2) \log p}{n}}\} \right] \\
& \quad + \mathbb{P}_{\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)} (M(\mathbf{z}) \geq M \log n \sqrt{\frac{(s_1 + s_2) \log p}{n}}) \\
& \lesssim \mathbb{E}_{\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)} \left[(1 - e^{-Q(\mathbf{z})}) \mathbb{1}\{0 < Q(\mathbf{z}) \leq M(\mathbf{z})\} \cdot \mathbb{1}\{M(\mathbf{z}) < M \log n \sqrt{\frac{(s_1 + s_2) \log p}{n}}\} \right] \\
& \quad + n^{-1} + p^{-1} \\
& \lesssim \log n \cdot \sqrt{\frac{(s_1 + s_2) \log p}{n}} \cdot \mathbb{E}_{\mathbf{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)} \left[\mathbb{1}\{0 < Q(\mathbf{z}) \leq M \log n \sqrt{\frac{(s_1 + s_2) \log p}{n}}\} \right] + n^{-1} + p^{-1} \\
& \lesssim \log^2 n \cdot \frac{(s_1 + s_2) \log p}{n},
\end{aligned}$$

where the last inequality uses the assumption that $\sup_{|x| < \delta} f_{Q, \theta}(x) < M_2$.

CHAPTER 4 : CHIME: Clustering of High-Dimensional Gaussian Mixtures with EM Algorithm and Its Optimality

4.1. Introduction

Clustering analysis, which aims to partition unlabeled data into homogeneous groups, is an ubiquitous problem in statistics and machine learning with a broad range of applications, including pattern recognition, disease diagnostics, and information retrieval (see Bishop, 2006; Hastie et al., 2009, and the references therein). A number of clustering algorithms have been proposed in the literature. The well-known k -means and k -medians algorithms Bradley et al. (1999) are centroid-based. Hierarchical clustering Ward Jr (1963) builds a hierarchy of clusters based on the empirical measures of dissimilarity among sets of observations. Clustering algorithms have also been developed and analyzed under the probabilistic mixture model framework Scott and Symons (1971); Duda and Hart (1973). Among the possible probability distributions for the mixture components, the Gaussian distribution is the most commonly used for both theoretical and computational considerations Everitt (1981); Lindsay (1995); Bouveyron and Brunet-Saumard (2014), and has been widely used in a range of applications for clustering and discriminant analysis Fraley and Raftery (2002); Reynolds (2015).

In the present paper, we consider clustering of data generated from Gaussian mixtures with the focus on the high-dimensional setting. We begin with the following mixture of two p -dimensional Gaussian distributions with equal covariance matrices:

$$Y \sim \begin{cases} 1, & \text{with probability } 1 - \omega^* \\ 2, & \text{with probability } \omega^* \end{cases} \quad \text{and } Z | Y = k \sim N_p(\boldsymbol{\mu}_k^*, \Sigma^*), \quad k = 1, 2. \quad (4.1)$$

In clustering, Z is observable and Y is not. For identifiability, we assume $\omega^* \in (0, 1/2]$. Suppose we have n unlabeled observations $\mathbf{z}^{(i)}$ ($i = 1, \dots, n$) generated independently and

identically from the mixture in (4.1), that is,

$$\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \dots, \mathbf{z}^{(n)} \stackrel{i.i.d.}{\sim} (1 - \omega^*)N_p(\boldsymbol{\mu}_1^*, \Sigma^*) + \omega^*N_p(\boldsymbol{\mu}_2^*, \Sigma^*). \quad (4.2)$$

The goal is to cluster $\mathbf{z}^{(i)}$ ($i = 1, \dots, n$) into two groups. Although the conventional low-dimensional setting will also be considered later, we are particularly interested in the high-dimensional setting where the dimension p can be much larger than the sample size n .

Clustering analysis is closely connected to classification analysis where the goal is to construct a classifier for future unlabeled observations based on the observed labeled data. In the ideal case where the parameter $\boldsymbol{\theta}^* = \{\omega^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*, \Sigma^*\}$ is known, the optimal classification procedure is the Fisher's linear discriminant rule

$$G_{\boldsymbol{\theta}^*}(\mathbf{z}) = \begin{cases} 1, & (\mathbf{z} - \frac{\boldsymbol{\mu}_1^* + \boldsymbol{\mu}_2^*}{2})^\top \boldsymbol{\beta}^* \geq \log(\frac{\omega^*}{1 - \omega^*}) \\ 2, & (\mathbf{z} - \frac{\boldsymbol{\mu}_1^* + \boldsymbol{\mu}_2^*}{2})^\top \boldsymbol{\beta}^* < \log(\frac{\omega^*}{1 - \omega^*}), \end{cases} \quad (4.3)$$

where $\boldsymbol{\beta}^* = \Omega^* \boldsymbol{\delta}^*$, $\Omega^* = (\Sigma^*)^{-1}$ and $\boldsymbol{\delta}^* = \boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$. Let Φ be the cumulative distribution function of the standard normal distribution. Fisher's rule given in (4.3) achieves the optimal mis-classification error

$$\begin{aligned} R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) &:= \mathbb{E}[I(G_{\boldsymbol{\theta}^*}(Z) \neq Y)] \\ &= (1 - \omega^*)\Phi\left(\frac{1}{\Delta} \log \frac{\omega^*}{1 - \omega^*} - \frac{1}{2}\Delta\right) + \omega^*\bar{\Phi}\left(\frac{1}{\Delta} \log \frac{\omega^*}{1 - \omega^*} + \frac{1}{2}\Delta\right), \end{aligned} \quad (4.4)$$

where $\Delta = \sqrt{(\boldsymbol{\delta}^*)^\top \Omega^* \boldsymbol{\delta}^*}$ and $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$. See, for example, Anderson (2003).

In practice, the parameters $\omega^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*$ and Σ^* are unknown and a data driven method is needed. In the supervised case where the sample labels of $\mathbf{z}^{(i)}$ are known, a common approach in the low-dimensional setting is to simply plug the sample values in (4.3). Driven by a wide range of applications, recent focus in clustering and classification has shifted to the high-dimensional setting where p can be much larger than n . In this case, the

sample covariance matrix may not even be invertible and it is difficult to estimate the precision matrix Ω^* . Cai and Liu (2011); Mai et al. (2012) proposed to directly estimate the discriminant direction $\beta^* = \Omega^* \delta^*$. More specifically, let $\hat{\mu}_k$ be the sample mean for class k ($k = 1, 2$) and $\hat{\Sigma}$ be the pooled sample covariance matrix. Assuming that β^* is sparse, one can estimate β^* directly through the regularized ℓ_1 minimization

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \beta^\top \hat{\Sigma} \beta - \beta^\top (\hat{\mu}_1 - \hat{\mu}_2) + \lambda_n \|\beta\|_1 \right\}, \quad (4.5)$$

where λ_n is a tuning parameter. The classification rule is obtained by using (4.3) with β^* replaced by $\hat{\beta}$, μ_k^* replaced by $\hat{\mu}_k$ for $k = 1, 2$, and ω^* replaced by the sample proportion. This algorithm is easy to implement and avoids estimation of Ω^* .

For unsupervised learning, the class labels are not observed. Compared with the classification analysis, clustering high-dimensional Gaussian mixtures is significantly more complicated, both in terms of the algorithm and in terms of the theoretical analysis. It is not easy to estimate the parameters $\omega^*, \mu_1^*, \mu_2^*$ and Σ^* in the high-dimensional case. In the classical low-dimensional setting, commonly used methods for estimating the parameters include the method of moments Pearson (1894), spectral method Jin et al. (2017), the maximum likelihood, and the Expectation-Maximization (EM) algorithm Redner and Walker (1984); Balakrishnan et al. (2017).

In this paper, we introduce CHIME, a clustering procedure for high-dimensional Gaussian mixtures based on the EM algorithm together with the direct estimation idea introduced in Cai and Liu (2011). The method uses the posterior probability of $\mathbf{z}^{(i)}$ in class k as the ‘sample label’ of $\mathbf{z}^{(i)}$ and efficiently estimates the parameters via the EM algorithm. A key component of the proposed method is to directly estimate and update the discriminant direction β^* in each iteration through the regularized ℓ_1 minimization algorithm (4.5). The resulting estimates are subsequently used to yield the discriminant rule as in (4.3). Instead of restricting both the mean vectors and the precision matrix to be sparse, CHIME only requires sparsity of the discriminant vector β^* .

Both theoretical and numerical properties of the CHIME algorithm are studied. Our analysis first obtains the rate of convergence for estimating β^* under the ℓ_2 norm loss, and the convergence rate of the expected excess mis-clustering error $R(\hat{G}_{CHIME}) - R_{\text{opt}}(G_{\theta^*})$ (the mis-clustering error is defined later in (4.6)). Furthermore, minimax lower bounds are obtained. The upper and lower bounds together establish the rate optimality of the estimator $\hat{\beta}$ and the CHIME procedure. Specifically, we show that

$$R(\hat{G}_{CHIME}) - R_{\text{opt}}(G_{\theta^*}) \asymp \frac{s \log p}{n},$$

where s is the sparsity of the discriminant vector β^* , and prove that this rate is optimal. To the best of our knowledge, this is the first optimality result for clustering of high-dimensional Gaussian mixtures and the first construction of a rate-optimal clustering procedure.

In addition to its theoretical optimality, CHIME is computationally easy to implement. The updates of $\hat{\omega}$ and $\hat{\mu}_k$ in the M-step of the EM algorithm can be calculated analytically, and the update of $\hat{\beta}$ can be implemented via linear programming. Simulation results show that CHIME outperforms existing clustering methods and achieves performance comparable to that of (4.5), which requires the additional label information. The effectiveness of CHIME is also illustrated through an analysis of a glioblastoma gene expression data set, and CHIME yields the smallest error when clustering heterogeneous patients into two distinct subtypes of glioblastoma.

Although the focus of the present paper is on the high-dimensional setting, we also consider clustering of low-dimensional Gaussian mixtures via the CLOME procedure. The technical tools developed for the high-dimensional setting can be used to establish the optimality for the general low-dimensional setting where the covariance matrix is not necessarily the identity matrix.

Our proposed clustering method together with its theoretical optimality guarantees extends the literature on clustering of high-dimensional Gaussian mixtures. Azizyan et al. (2013)

considered a special case of (4.1) with $\Sigma^* = \sigma^2 \mathbf{I}_p$, $\omega^* = 1/2$, and provided both lower and upper bounds, on the mis-clustering error for sparse $\boldsymbol{\delta}^*$, but the upper bound is not tight. Wang et al. (2014) also focused on the special case $\Sigma^* = \sigma^2 \mathbf{I}_p$ and $\omega^* = 1/2$, studied the performance of the high-dimensional EM algorithm and established the convergence rate for the estimator of the sparse mean vector. Jin et al. (2017) considered the special case where $\Sigma^* = \mathbf{I}_p$ and studied the statistical limits of clustering when the signals are "rare and weak". A phase transition diagram for the IF-PCA method is given in Jin et al. (2016b). Azizyan et al. (2015) extended the results in Azizyan et al. (2013) to allow for a general covariance matrix Σ^* and directly estimated the discriminant vector $\boldsymbol{\beta}^*$ via the LPD rule Cai and Liu (2011). Using the initial estimates of $\boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*$ and Σ^* provided by Hardt and Price (2014), they established an upper bound for the mis-clustering error as well as recovery of the support of sparse $\boldsymbol{\beta}^*$ under regularity conditions. Compared to the procedure in Wang et al. (2014), our proposed CHIME yields a sparse estimate of $\boldsymbol{\beta}^*$ without the need of truncation, nor does it require sample splitting across iterations.

The rest of the paper is organized as follows. The proposed procedure, CHIME, for clustering high-dimensional Gaussian mixtures is described in detail in Section 4.2. The theoretical properties are analyzed in Section 4.3. Both upper and lower bounds are obtained. Together they establish the optimality of CHIME as well as the estimator of discriminant vector $\boldsymbol{\beta}^*$. Section 4.4 considers clustering low-dimensional Gaussian mixtures based on the classical EM algorithm and establishes the optimality of the clustering procedure by modifying our analysis for the high-dimensional setting. A simulation study is given in Section 4.5 where we compare the performance of CHIME to other existing clustering methods in the literature. Section 4.6 uses a real data application to illustrate the merit of CHIME. Section 4.7 discusses extensions to the multi-class setting. The proofs of the main results are given in Section 4.8. Proofs of other results together with additional technical details as well as additional simulations are provided in Cai et al. (2018b).

4.2. Methodology

In this section, we present in detail the clustering procedure CHIME under the two-component Gaussian mixture model (4.2).

We begin with notations. Throughout the paper, X, Y, Z, \dots denote random vectors and $\mathbf{x}, \mathbf{y}, \mathbf{z}, \dots$ denote their realizations. For $a, b \in \mathbb{R}$, we denote by $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, 2, \dots, n\}$. For a vector $\mathbf{x} \in \mathbb{R}^p$, the usual vector ℓ_0, ℓ_1, ℓ_2 and ℓ_∞ norms are denoted respectively by $\|\mathbf{x}\|_0, \|\mathbf{x}\|_1, \|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_\infty$. Here the ℓ_0 norm counts the number of nonzero entries in a vector. We use $\text{supp}(\mathbf{x})$ to denote the support of the vector \mathbf{x} . The Frobenius norm of a matrix $A = (a_{ij})$ is defined as $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$. The matrix ℓ_1 and ℓ_2 norms are defined, respectively, as $\|A\|_1 = \sup_{\|\mathbf{x}\|_1=1} \|A\mathbf{x}\|_1$ and $\|A\|_2 = \sup_{\|\mathbf{x}\|_2=1} \|A\mathbf{x}\|_2$. The matrix ℓ_0 norm is defined similarly to the vector ℓ_0 norm as $\|A\|_0 = |\{(i, j) : a_{ij} \neq 0\}|$, where $|\cdot|$ denotes the cardinality here. The vector ℓ_∞ norm on matrix A is $|A|_\infty = \max_{i,j} |A_{ij}|$. For a symmetric matrix A , we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote respectively the largest and smallest eigenvalue of A . We say $A \succ 0$ if A is positive definite. The inner product between two matrices A and B is defined as $\langle A, B \rangle = \text{tr}(A^\top B)$. For a set \mathcal{A} , we use \mathcal{A}^c to denote its complement, and use $I(\mathcal{A})$ to denote its corresponding indicator function. For a positive integer $s < p$, let $\Gamma(s) = \{\mathbf{u} \in \mathbb{R}^p : 2\|\mathbf{u}_{S^c}\|_1 \leq 4\|\mathbf{u}_S\|_1 + 3\sqrt{s}\|\mathbf{u}\|_2, \text{ for some } S \subset [p] \text{ with } |S| = s\}$. For a vector $\mathbf{x} \in \mathbb{R}^p$ and a matrix $A \in \mathbb{R}^{m \times p}$, we define $\|\mathbf{x}\|_{2,s} = \sup_{\|\mathbf{y}\|_2=1, \mathbf{y} \in \Gamma(s)} |\mathbf{x}^\top \mathbf{y}|$ and $\|A\|_{2,s} = \sup_{\|\mathbf{y}\|_2=1, \mathbf{y} \in \Gamma(s)} \|A\mathbf{y}\|_2$. For two sequences of positive numbers a_n and b_n , $a_n \lesssim b_n$ means that for some constant $c > 0$, $a_n \leq cb_n$ for all n , and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Finally, we use $c_0, c_1, c_2, C_1, C_2, \dots$ to denote generic positive constants that may vary from place to place.

4.2.1. The Gaussian mixture model

Suppose we have n observations $\{\mathbf{z}^{(1)} \dots, \mathbf{z}^{(n)}\}$ generated independently and identically from the p -dimensional Gaussian mixture model in (4.2) without knowing labels (y_1, \dots, y_n) ,

and wish to cluster the observations $\{\mathbf{z}^{(1)} \dots, \mathbf{z}^{(n)}\}$ into two groups. The accuracy of a clustering rule $G : \mathbf{z}^{(i)} \rightarrow \{1, 2\}$, $i = 1, \dots, n$, is measured by the expected mis-clustering error,

$$R(G) = \min_{\pi \in \mathcal{P}_2} \mathbb{E}[I(G(\mathbf{z}) \neq \pi(y))], \quad (4.6)$$

where $\mathcal{P}_2 = \{\pi : [1, 2] \rightarrow [1, 2]\}$ is a set of permutation function, and y is the latent label of a future observation \mathbf{z} .

As mentioned in the introduction, for this clustering problem, it is important to first estimate the parameters ω^* , $\boldsymbol{\mu}_1^*$, $\boldsymbol{\mu}_2^*$ and Σ^* in (4.2). In the classical setting where p is much smaller than n , it has been shown that the maximum likelihood estimator (MLE) performs well under mild conditions Balakrishnan et al. (2017). The joint log-likelihood of the data $\mathbf{z}^{(i)}$ ($i = 1, \dots, n$) can be written as

$$L(\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma; \mathbf{z}) = \frac{1}{n} \sum_{i=1}^n \log \left\{ (1 - \omega) f(\mathbf{z}^{(i)} \mid \boldsymbol{\mu}_1, \Sigma) + \omega f(\mathbf{z}^{(i)} \mid \boldsymbol{\mu}_2, \Sigma) \right\}, \quad (4.7)$$

where $f(\cdot \mid \boldsymbol{\mu}_k, \Sigma)$ represents the density function of $N_p(\boldsymbol{\mu}_k, \Sigma)$. The MLE maximizes the joint log-likelihood function $L(\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma; \mathbf{z})$.

When p is large, direct optimization of the log-likelihood in (4.7) becomes infeasible due to the nonconvexity of the objective function $L(\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma; \mathbf{z})$. Moreover, the MLE does not even exist in the high-dimensional setting where $p \gg n$. In this paper, we propose to explore the sparsity of the discriminant vector $\boldsymbol{\beta}^*$ as in Cai and Liu (2011) for the supervised case by noting that the discriminant rule in (4.3) depends on Σ^* only through $\boldsymbol{\beta}^*$. Further, we adopt the EM algorithm Dempster et al. (1977) to address the nonconvexity of the joint log-likelihood.

4.2.2. A clustering procedure based on the EM algorithm

To simplify the notation, under the mixture model (4.2), we denote $\boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$, and let $\boldsymbol{\beta} = \Omega(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ with $\Omega = \Sigma^{-1}$. For a given $\boldsymbol{\theta}$, we use $\mathbb{E}_{\boldsymbol{\theta}}$ and $\mathbb{P}_{\boldsymbol{\theta}}$ to denote the

expectation and probability under the model (4.2) with respect to the parameter $\boldsymbol{\theta}$. In addition, sometimes we write $\mathbb{E}_{\boldsymbol{\theta}^*}$, $\mathbb{P}_{\boldsymbol{\theta}^*}$ as \mathbb{E} and Pr when there is no ambiguity.

Note that if the true labels $\mathbf{y} = \{y_i\}_{i=1}^n \in \{1, 2\}^n$ were observed together with $\mathbf{z} = \{\mathbf{z}^{(i)}\}_{i=1}^n$, the log-likelihood of the complete data is given by

$$L_C(\boldsymbol{\theta}; \mathbf{y}, \mathbf{z}) = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^2 I(y_i = k) \left\{ \log f(\mathbf{z}^{(i)} \mid \boldsymbol{\mu}_k, \Sigma) + \log \mathbb{P}_{\boldsymbol{\theta}}(y_i = k) \right\}.$$

To address the nonconvexity of the joint log-likelihood, we use the EM algorithm, which iterates between two goals: classification given the parameters, and estimation given the labels. In the t -th iteration, given the estimated $\hat{\boldsymbol{\theta}}^{(t)} = (\hat{\omega}^{(t)}, \hat{\boldsymbol{\mu}}_1^{(t)}, \hat{\boldsymbol{\mu}}_2^{(t)}, \hat{\Sigma}^{(t)})$ from the previous step, the E-step can be interpreted as classifying the observed data $\mathbf{z}^{(i)}$ by assuming the true parameter is $\hat{\boldsymbol{\theta}}^{(t)}$. The posterior probability of the i -th sample in class 2 given the observed data $\mathbf{z}^{(i)}$ can be calculated as

$$\begin{aligned} \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)}) &= \mathbb{P}_{\hat{\boldsymbol{\theta}}^{(t)}}(y_i = 2 \mid \mathbf{z}^{(i)}) \\ &= \frac{\hat{\omega}^{(t)}}{\hat{\omega}^{(t)} + (1 - \hat{\omega}^{(t)}) \exp \left\{ (\hat{\Omega}^{(t)}(\hat{\boldsymbol{\mu}}_2^{(t)} - \hat{\boldsymbol{\mu}}_1^{(t)}))^\top (\mathbf{z}^{(i)} - \frac{\hat{\boldsymbol{\mu}}_1^{(t)} + \hat{\boldsymbol{\mu}}_2^{(t)}}{2}) \right\}}. \end{aligned} \quad (4.8)$$

We then calculate the expectation of the log-likelihood, with respect to the conditional distribution of y given \mathbf{z} under the current estimate of the parameters $\hat{\boldsymbol{\theta}}^{(t)}$, as

$$\begin{aligned} Q_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)}) &= \mathbb{E}_{\hat{\boldsymbol{\theta}}^{(t)}}[\log L_C(\boldsymbol{\theta}; \mathbf{y}, \mathbf{z}) \mid \mathbf{z}] \\ &= -\frac{1}{2n} \sum_{i=1}^n \left\{ (1 - \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)})) (\mathbf{z}^{(i)} - \boldsymbol{\mu}_1)^\top \Omega (\mathbf{z}^{(i)} - \boldsymbol{\mu}_1) \right. \\ &\quad \left. + \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)}) (\mathbf{z}^{(i)} - \boldsymbol{\mu}_2)^\top \Omega (\mathbf{z}^{(i)} - \boldsymbol{\mu}_2) \right\} \\ &\quad + \frac{1}{n} \sum_{i=1}^n \left\{ (1 - \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)})) \log(1 - \omega) + \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)}) \log \omega \right\} + \frac{1}{2} \log |\Omega|. \end{aligned}$$

The M-step proceeds by maximizing $Q_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$ given $\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)})$, and is interpreted as

parameter estimation given the labels. The maximizer,

$$\hat{\boldsymbol{\theta}}^{(t+1)} = M_n(\hat{\boldsymbol{\theta}}^{(t)}) = \arg \max_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)}), \quad (4.9)$$

can be calculated analytically. We now derive the exact analytic form for the M-step in the t -th iteration, which is used to obtain updates of ω , $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ . It is straightforward to define and calculate

$$\hat{\omega}^{(t+1)} = \hat{\omega}(\hat{\boldsymbol{\theta}}^{(t)}) = \frac{1}{n} \sum_{i=1}^n \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)}), \quad (4.10)$$

$$\hat{\boldsymbol{\mu}}_1^{(t+1)} = \hat{\boldsymbol{\mu}}_1(\hat{\boldsymbol{\theta}}^{(t)}) = \left\{ n - \sum_{i=1}^n \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)}) \right\}^{-1} \left\{ \sum_{i=1}^n (1 - \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)})) \mathbf{z}^{(i)} \right\}, \quad (4.11)$$

$$\hat{\boldsymbol{\mu}}_2^{(t+1)} = \hat{\boldsymbol{\mu}}_2(\hat{\boldsymbol{\theta}}^{(t)}) = \left\{ \sum_{i=1}^n \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)}) \right\}^{-1} \left\{ \sum_{i=1}^n \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)}) \mathbf{z}^{(i)} \right\}. \quad (4.12)$$

This leads to a solution for $\hat{\Sigma}^{(t+1)}$ given by

$$\begin{aligned} \hat{\Sigma}^{(t+1)} = \hat{\Sigma}(\hat{\boldsymbol{\theta}}^{(t)}) = \frac{1}{n} \sum_{i=1}^n \left\{ (1 - \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)})) (\mathbf{z}^{(i)} - \hat{\boldsymbol{\mu}}_1^{(t+1)}) (\mathbf{z}^{(i)} - \hat{\boldsymbol{\mu}}_1^{(t+1)})^\top + \right. \\ \left. \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)}) (\mathbf{z}^{(i)} - \hat{\boldsymbol{\mu}}_2^{(t+1)}) (\mathbf{z}^{(i)} - \hat{\boldsymbol{\mu}}_2^{(t+1)})^\top \right\}. \end{aligned} \quad (4.13)$$

Note that in the high-dimensional setting where $p \gg n$, $\hat{\Sigma}^{(t+1)}$ is singular and cannot be used directly in (4.3) and (4.8) to obtain a clustering rule and $\gamma(\mathbf{z}^{(i)})$. Instead of estimating the covariance matrix Σ^* , we estimate the discriminant vector $\boldsymbol{\beta}^*$ directly. The update $\hat{\boldsymbol{\beta}}^{(t+1)}$ can be obtained through the regularized ℓ_1 minimization

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{2} \boldsymbol{\beta}^\top \hat{\Sigma}^{(t+1)} \boldsymbol{\beta} - \boldsymbol{\beta}^\top (\hat{\boldsymbol{\mu}}_1^{(t+1)} - \hat{\boldsymbol{\mu}}_2^{(t+1)}) + \lambda_n^{(t+1)} \|\boldsymbol{\beta}\|_1 \right\}, \quad (4.14)$$

where $\lambda_n^{(t+1)}$ is the tuning parameter. It is shown in the supplement Cai et al. (2018b) that the sequence $\lambda_n^{(t+1)} = \kappa^t \cdot C_1 \frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*)}{\sqrt{s}} + (\frac{1-\kappa^{t+1}}{1-\kappa}) C_\lambda \sqrt{\frac{\log p}{n}}$, for some constants $C_1, C_\lambda > 0$, $d_{2,s}$ is defined later in (4.17) and $\kappa \in (0, 1/2)$ is an appropriate choice for tuning parameters.

In practice, $\lambda_n^{(t+1)}$ can be chosen by cross validation.

As a result, in the high-dimensional setting, the update of $\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)})$ in the E-step is different from (4.8) and proposed to be

$$\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)}) := \mathbb{P}_{\hat{\boldsymbol{\theta}}^{(t)}}(y_i = 2 | \mathbf{z}^{(i)}) = \frac{\hat{\omega}^{(t)}}{\hat{\omega}^{(t)} + (1 - \hat{\omega}^{(t)}) \exp \left\{ (\hat{\boldsymbol{\beta}}^{(t)})^\top \left(\mathbf{z}^{(i)} - \frac{\hat{\boldsymbol{\mu}}_1^{(t)} + \hat{\boldsymbol{\mu}}_2^{(t)}}{2} \right) \right\}}. \quad (4.15)$$

Given a suitable initialization, the EM algorithm iterates between the E-step and M-Step as described above, and terminates in, say T_0 , steps. Once the final estimates of $\boldsymbol{\theta}^*$ and $\boldsymbol{\beta}^*$ are obtained, the clustering rule can be constructed by plugging them into the Fisher's rule (4.3). We call this procedure CHIME for **C**lustering of **H**igh-dimensional Gaussian **M**ixtures with the **E**M, which is summarized in Algorithm 1.

Remark 5. CHIME requires the initialization $\hat{\boldsymbol{\theta}}^{(0)}$ to be reasonably good to ensure the convergence of $\hat{\boldsymbol{\theta}}^{(t)}$ to an optimum near the true parameters $\boldsymbol{\theta}^*$. We address the issue of initialization in Section 4.3. The total number of iterations T_0 needs to be specified. It is shown in Section 4.3 that $T_0 \asymp \log n$ is sufficient to yield the optimal convergence rate for $\hat{\boldsymbol{\beta}}^{(T_0)}$. In practice, it is recommended to run Algorithm 1 until the distance between $\hat{\boldsymbol{\theta}}^{(t+1)}$ and $\hat{\boldsymbol{\theta}}^{(t)}$ is less than a pre-specified tolerance level. In addition, Algorithm 1 requires specifying the contraction constant κ as well as constants C_1 and C_λ . The choice of the tuning parameter in the form of $\lambda_n^{(0)}$ and (4.16) is necessary for establishing convergence of $\hat{\boldsymbol{\beta}}^{(T_0)}$ to the true parameter $\boldsymbol{\beta}^*$, and will be discussed in detail in Section 4.3.

4.3. Theoretical Analysis

In this section, we study the properties of the estimator $\hat{\boldsymbol{\beta}}^{(T_0)}$ and the performance of the CHIME clustering rule \hat{G}_{CHIME} proposed in Algorithm 1. We first establish the rates of convergence for the estimation and mis-clustering error and then provide matching minimax lower bounds. These results together show the optimality of CHIME as well as the proposed estimator of the discriminant vector $\boldsymbol{\beta}^*$.

Algorithm 1 Clustering of **HI**gh-dimensional Gaussian **MIX**tures with the **EM** (CHIME)

- 1: **Inputs:** Initializations $\hat{\omega}^{(0)}, \hat{\boldsymbol{\mu}}_1^{(0)}, \hat{\boldsymbol{\mu}}_2^{(0)}$ and $\hat{\Sigma}^{(0)}$, maximum number of iterations T_0 , and a constant $\kappa \in (0, 1)$. Set

$$\hat{\boldsymbol{\beta}}^{(0)} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{2} \boldsymbol{\beta}^\top \hat{\Sigma}^{(0)} \boldsymbol{\beta} - \boldsymbol{\beta}^\top (\hat{\boldsymbol{\mu}}_1^{(0)} - \hat{\boldsymbol{\mu}}_2^{(0)}) + \lambda_n^{(0)} \|\boldsymbol{\beta}\|_1 \right\},$$

where the tuning parameter $\lambda_n^{(0)} = C_\lambda \cdot (|\hat{\omega}| \vee \|\hat{\boldsymbol{\mu}}_1^{(0)} - \hat{\boldsymbol{\mu}}_2^{(0)}\|_{2,s} \vee \|\hat{\Sigma}^{(0)}\|_{2,s}) / \sqrt{s} + C_\lambda \sqrt{\log p/n}$.

- 2: **for** $t = 0, 1, \dots, T_0 - 1$ **do**
3: **E-Step:** Evaluate $Q_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$ with $\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)})$ defined in (4.15).
4: **M-Step:** Update $\hat{\omega}^{(t+1)}, \hat{\boldsymbol{\mu}}_1^{(t+1)}, \hat{\boldsymbol{\mu}}_2^{(t+1)}$, and $\hat{\Sigma}^{(t+1)}$ via (4.10), (4.11), (4.12) and (4.13), and $\hat{\boldsymbol{\beta}}^{(t+1)}$ via

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{2} \boldsymbol{\beta}^\top \hat{\Sigma}^{(t+1)} \boldsymbol{\beta} - \boldsymbol{\beta}^\top (\hat{\boldsymbol{\mu}}_1^{(t+1)} - \hat{\boldsymbol{\mu}}_2^{(t+1)}) + \lambda_n^{(t+1)} \|\boldsymbol{\beta}\|_1 \right\},$$

with

$$\lambda_n^{(t+1)} = \kappa \lambda_n^{(t)} + C_\lambda \sqrt{\frac{\log p}{n}}. \quad (4.16)$$

- 5: **end for**
6: Output $\hat{\omega}^{(T_0)}, \hat{\boldsymbol{\mu}}_1^{(T_0)}, \hat{\boldsymbol{\mu}}_2^{(T_0)}$ and $\hat{\boldsymbol{\beta}}^{(T_0)}$.
7: Construct the clustering rule

$$\hat{G}_{CHIME}(\mathbf{z}) = \begin{cases} 1, & \left\{ \mathbf{z} - \frac{\hat{\boldsymbol{\mu}}_1^{(T_0)} + \hat{\boldsymbol{\mu}}_2^{(T_0)}}{2} \right\}^\top \hat{\boldsymbol{\beta}}^{(T_0)} \geq \log\left(\frac{\hat{\omega}^{(T_0)}}{1 - \hat{\omega}^{(T_0)}}\right), \\ 2, & \left\{ \mathbf{z} - \frac{\hat{\boldsymbol{\mu}}_1^{(T_0)} + \hat{\boldsymbol{\mu}}_2^{(T_0)}}{2} \right\}^\top \hat{\boldsymbol{\beta}}^{(T_0)} < \log\left(\frac{\hat{\omega}^{(T_0)}}{1 - \hat{\omega}^{(T_0)}}\right). \end{cases}$$

We first introduce the parameter space. For parameters $\boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$ and $\tilde{\boldsymbol{\theta}} = (\tilde{\omega}, \tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\mu}}_2, \tilde{\Sigma})$, define their $\ell_{2,s}$ distance by

$$d_{2,s}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = |\omega - \tilde{\omega}| \vee \|\boldsymbol{\mu}_1 - \tilde{\boldsymbol{\mu}}_1\|_{2,s} \vee \|\boldsymbol{\mu}_2 - \tilde{\boldsymbol{\mu}}_2\|_{2,s} \vee \|(\Sigma - \tilde{\Sigma})\tilde{\boldsymbol{\beta}}\|_{2,s}. \quad (4.17)$$

We shall write $d_{2,s}(\boldsymbol{\theta})$ for $d_{2,s}(\boldsymbol{\theta}, \mathbf{0})$, and consider the following parameter space

$$\Theta_p(s, c_\omega, M, M_b) = \{ \boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p \times p}, \Sigma \succ 0, \|\boldsymbol{\beta}\|_0 \leq s, \\ \omega \in (c_\omega, 1 - c_\omega), M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M, \|\boldsymbol{\beta}\|_1 \leq M_b \}. \quad (4.18)$$

This is a natural parameter space to consider. The condition on the eigenvalues of Σ is standard. For example, it has been used in Cai et al. (2011), Bickel and Levina (2008) and Cai and Zhang (2017) for estimation of precision matrices, covariance matrices, and regression coefficients, respectively. Condition on $\|\beta\|_1$ were also similarly used in Neykov et al. (2015) and Tian and Gu (2017) for discriminant analysis.

4.3.1. Upper bounds

We need two technical conditions before stating the properties of the clustering algorithm.

Recall that in (4.4), $\Delta = \sqrt{(\mu_1^* - \mu_2^*)^\top (\Sigma^*)^{-1} (\mu_1^* - \mu_2^*)}$ is the Mahalanobis distance between μ_1^* and μ_2^* with covariance matrix Σ^* , and can be interpreted as the Signal-to-Noise Ratio. For constants $c_0, c_1, C_b > 0$ and $c_0 \leq c_\omega, c_1 < 1$, the contraction basin $B_{con}(\theta^*; c_0, c_1, C_b, s)$ is defined as

$$\begin{aligned} B_{con}(\theta^*; c_0, c_1, C_b, s) = \{ & \theta = (\omega, \mu_1, \mu_2, \Sigma) : \mu_1, \mu_2 \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p \times p}, \Sigma \succ 0, \\ & \omega \in (c_0, 1 - c_0), (1 - c_1)\Delta^2 < |\delta_1(\beta)|, |\delta_2(\beta)|, \sigma^2(\beta) < (1 + c_1)\Delta^2, \\ & \beta - \beta^* \in \Gamma(s), \|\beta - \beta^*\|_1 \leq C_b\Delta, \|\mu_1 - \mu_1^*\|_{2,s}, \|\mu_2 - \mu_2^*\|_{2,s} \leq C_b\Delta \}, \end{aligned} \quad (4.19)$$

where $\delta_1(\beta) = \beta^\top (\mu_1^* - \frac{\mu_1 + \mu_2}{2})$, $\delta_2(\beta) = \beta^\top (\mu_2^* - \frac{\mu_1 + \mu_2}{2})$, and $\sigma(\beta) = \sqrt{\beta^\top \Sigma^* \beta}$.

The following conditions are needed to establish the convergence of $\hat{\beta}^{(T_0)}$.

(C1) The initial value $\hat{\theta}^{(0)}$ satisfies that

$$d_{2,s}(\hat{\theta}^{(0)}, \theta^*) \vee \|\hat{\beta}^{(0)} - \beta^*\|_2 \leq r\Delta, \quad \hat{\beta}^{(0)} - \beta^* \in \Gamma(s)$$

with $r < \frac{|c_0 - c_\omega|}{\Delta} \wedge \frac{\sqrt{9M + 16c_1} - \sqrt{9M}}{4} \wedge \sqrt{\frac{c_1}{M}} \wedge \frac{C_b}{5\sqrt{s}}$.

In fact, condition **(C1)** guarantees that $\theta^{(t)} \in B_{con}(\theta^*; c_0, c_1, C_b, s)$ for $t \geq 0$ in Algorithm 1, which is shown in Lemma A.2 and proved in the supplement Cai et al. (2018b). We will discuss in Section 4.3.2 an initialization algorithm whose output satisfies Condition **(C1)**.

(C2) The Signal-to-Noise Ratio Δ satisfies that

$$\Delta > C(c_0, c_1, M, C_b), \quad (4.20)$$

where $C(c_0, c_1, M, C_b)$ is a constant that only depends on the c_0, c_1, M , and C_b , and is given in (C.24) in the supplement Cai et al. (2018b).

Before we state the main results, we introduce two technical lemmas that characterize the properties of the population version of the proposed CHIME algorithm under Conditions (C1) and (C2). We define the respective population version of M-step as follows.

Let $M(\boldsymbol{\theta}) = (\omega(\boldsymbol{\theta}), \boldsymbol{\mu}_1(\boldsymbol{\theta}), \boldsymbol{\mu}_2(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta}))$ denote the population version of $M_n(\boldsymbol{\theta})$, the estimator evaluated in (4.9). More specifically,

$$M(\boldsymbol{\theta}) = \arg \max_{\tilde{\boldsymbol{\theta}}} Q(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta}) := \arg \max_{\tilde{\boldsymbol{\theta}}} \mathbb{E}_{\boldsymbol{\theta}^*} [Q_n(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta})]. \quad (4.21)$$

By definition, $M(\boldsymbol{\theta})$ can be analytically expressed as

$$\omega(\boldsymbol{\theta}) = \mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)], \quad \boldsymbol{\mu}_1(\boldsymbol{\theta}) = \frac{\mathbb{E}[(1 - \gamma_{\boldsymbol{\theta}}(Z))Z]}{\mathbb{E}[1 - \gamma_{\boldsymbol{\theta}}(Z)]}, \quad \boldsymbol{\mu}_2(\boldsymbol{\theta}) = \frac{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)Z]}{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]}, \quad (4.22)$$

$$\Sigma(\boldsymbol{\theta}) = \mathbb{E}[(1 - \gamma_{\boldsymbol{\theta}}(Z))(Z - \boldsymbol{\mu}_1(\boldsymbol{\theta}))(Z - \boldsymbol{\mu}_1(\boldsymbol{\theta}))^\top + \gamma_{\boldsymbol{\theta}}(Z)(Z - \boldsymbol{\mu}_2(\boldsymbol{\theta}))(Z - \boldsymbol{\mu}_2(\boldsymbol{\theta}))^\top]. \quad (4.23)$$

Using the above definition of the population version updates, we then introduce the following two lemmas, Lemma 9 characterizes the linear convergence of the population EM updates, and Lemma 10 captures the distance between the sample and population version estimates. These two lemmas are the key steps in the proof of the main result Theorem 6.

Lemma 9 (Contraction on the population iteration). *Suppose $\boldsymbol{\theta}^* \in \Theta_p(s, c_\omega, M, M_b)$. If $\Delta > C(c_0, c_1, M, C_b)$, where $C(c_0, c_1, M, C_b)$ is given in (C.24) in the supplement Cai et al. (2018b). Then there exists $0 < \kappa_0 < \frac{1}{2\sqrt{16M}}$, such that for $\boldsymbol{\theta} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)$,*

$$d_2(M(\boldsymbol{\theta}), \boldsymbol{\theta}^*) \leq \kappa_0 \cdot (d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2), \quad (4.24)$$

where $d_2(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = |\omega - \tilde{\omega}| \vee \|\boldsymbol{\mu}_1 - \tilde{\boldsymbol{\mu}}_1\|_2 \vee \|\boldsymbol{\mu}_2 - \tilde{\boldsymbol{\mu}}_2\|_2 \vee \|(\Sigma - \tilde{\Sigma})\tilde{\boldsymbol{\beta}}\|_2$.

Remark 6. *This theorem implies that*

$$d_{2,s}(M(\boldsymbol{\theta}), \boldsymbol{\theta}^*) \leq \kappa_0 \cdot (d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2).$$

Lemma 10 (Uniform concentration inequality). *Suppose $\boldsymbol{\theta}^* \in \Theta_p(s, c_\omega, M, M_b)$ with $c_\omega \in (0, 1)$ and M, M_b universally bounded. Under the condition **(C1)**, there exists a constant $C_{con} > 0$, such that with probability at least $1 - o(1)$,*

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)} d_{2,s}(M_n(\boldsymbol{\theta}), M(\boldsymbol{\theta})) &\leq C_{con} \sqrt{\frac{s \log p}{n}}; \\ \sup_{\boldsymbol{\theta} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)} \|(\hat{\Sigma}(\boldsymbol{\theta}) - \Sigma(\boldsymbol{\theta}))\boldsymbol{\beta}^*\|_\infty &\leq C_{con} \sqrt{\frac{\log p}{n}}. \end{aligned}$$

The above two lemmas imply that at each iteration, $\hat{\boldsymbol{\theta}}^{(t)}$ converges geometrically to the truth $\boldsymbol{\theta}^*$, until their distance is indistinguishable with the statistical error, whose rate is characterized by Lemma 10.

In addition, we point out that the inequality in (4.24) quantifies the contraction w.r.t the $\ell_{2,s}$ -norm of the distance between the population EM update and the true parameter $\boldsymbol{\theta}^*$. This contraction property is different from the ones used in Balakrishnan et al. (2017); Wang et al. (2014); Yi and Caramanis (2015). Consequently, our subsequent analysis differs from theirs. Indeed, existing works use the ℓ_2 or ℓ_∞ -norm of the distance between the EM update and the true parameter to define the contraction. The advantage with the $\ell_{2,s}$ -norm is that it characterizes a more refined sparsity-based difference, which converges at the rate $\sqrt{s \log p/n}$ by Lemma 10. The ℓ_2 or ℓ_∞ -norm used in previous works is not suitable for our purpose and requires stronger assumptions to obtain the same convergence rate in Theorem 6. Lastly, the establishment of Lemma 9 is based on the key observation that each term of $M(\hat{\boldsymbol{\theta}})$ and its corresponding Taylor expansion around the truth $\boldsymbol{\theta}^*$ involves either $\hat{\boldsymbol{\beta}}$ or $\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}$, both of whom lie in the region $\Gamma(s)$, leading to a sharper Cauchy-Schwartz

inequality by using the $\ell_{2,s}$ norm.

We are now ready to state the first main result. The following theorem shows that under Conditions **(C1)** and **(C2)**, the estimate $\hat{\boldsymbol{\beta}}^{(T_0)}$ provided by Algorithm 1 converges to the true parameter $\boldsymbol{\beta}^*$.

Theorem 6. *Suppose we observe n i.i.d. samples $\{\mathbf{z}^{(1)} \dots, \mathbf{z}^{(n)}\}$ from model (4.2) with the true parameter $\boldsymbol{\theta}^* \in \Theta_p(s, c_\omega, M, M_b)$, for some constant $c_\omega \in (0, 1)$ and universally bounded constants $M, M_b > 0$ and $s = o(\sqrt{n/\log p})$. Assume that conditions **(C1)** and **(C2)** hold with r satisfying $\sqrt{s \log p/n} = o(r)$. Then there exist constants $C_d, C_\lambda > 0$, $\kappa \in (0, 1/2)$, such that the output $\hat{\boldsymbol{\beta}}^{(T_0)}$ of Algorithm 1 with tuning parameters C_d, C_λ, κ satisfies, with probability $1 - o(1)$,*

$$\|\hat{\boldsymbol{\beta}}^{(T_0)} - \boldsymbol{\beta}^*\|_2 \lesssim \kappa^{T_0} d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) + \sqrt{\frac{s \log p}{n}}.$$

Consequently, if $T_0 \gtrsim (-\log(\kappa))^{-1} \log(n \cdot d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*))$, then

$$\|\hat{\boldsymbol{\beta}}^{(T_0)} - \boldsymbol{\beta}^*\|_2 \lesssim \sqrt{\frac{s \log p}{n}}. \quad (4.25)$$

The proof of Theorem 6 relies on Lemmas 9 and 10. The idea of proving Theorem 6 by establishing the contraction and uniform concentration properties is similar to that in Balakrishnan et al. (2017) for the conventional low-dimensional setting. However, establishing such results in the high-dimensional setting is quite challenging. The proof of Lemmas 9 and 10 are involved and are given in the supplement Cai et al. (2018b).

Remark 7. In comparison with the results in Wang et al. (2014); Yi and Caramanis (2015), which consider the high-dimensional EM algorithm under the special Gaussian mixture model $\frac{1}{2}N_p(-\boldsymbol{\mu}^*, \sigma^2 \mathbf{I}_p) + \frac{1}{2}N_p(\boldsymbol{\mu}^*, \sigma^2 \mathbf{I}_p)$, Theorem 6 establishes a faster convergence rate under a more general model. In fact, Wang et al. (2014) and Yi and Caramanis (2015) show the convergence rate $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|_2 \lesssim \sqrt{s \log p \log n/n}$ for their estimator $\hat{\boldsymbol{\mu}}$ and require sample splitting. In the present paper, we remove the $\log n$ factor and establish that $\|\hat{\boldsymbol{\beta}}^{(T_0)} - \boldsymbol{\beta}^*\|_2 \lesssim$

$\sqrt{s \log p/n}$ by using a uniform concentration inequality (Lemma 10) and thus avoid the need for sample splitting. The idea of using uniform concentration results is similar to that in Balakrishnan et al. (2017), but the techniques to prove this uniform concentration is much more involved in the high-dimensional setting.

We now turn to the performance of the clustering rule given by Algorithm 1. For ease of presentation, we denote the final output $\hat{\boldsymbol{\theta}}^{(T_0)}$ and $\hat{\boldsymbol{\beta}}^{(T_0)}$ of Algorithm 1 by $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ respectively. Recall that in Algorithm 1, after obtaining the final estimates $\hat{\omega}$, $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$ and $\hat{\boldsymbol{\beta}}$, we construct the following clustering rule

$$\hat{G}_{CHIME}(\mathbf{z}) = \begin{cases} 1, & (\mathbf{z} - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\beta}} \geq \log\left(\frac{\hat{\omega}}{1-\hat{\omega}}\right), \\ 2, & (\mathbf{z} - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\beta}} < \log\left(\frac{\hat{\omega}}{1-\hat{\omega}}\right), \end{cases} \quad (4.26)$$

where $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2$. By (4.6), we obtain

$$R(\hat{G}_{CHIME}) = (1 - \omega^*) \Phi\left(\frac{\log\left(\frac{\hat{\omega}}{1-\hat{\omega}}\right) + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1^*)^\top \hat{\boldsymbol{\beta}}}{\sqrt{\hat{\boldsymbol{\beta}}^\top \Sigma^* \hat{\boldsymbol{\beta}}}}\right) + \omega^* \bar{\Phi}\left(\frac{\log\left(\frac{\hat{\omega}}{1-\hat{\omega}}\right) + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2^*)^\top \hat{\boldsymbol{\beta}}}{\sqrt{\hat{\boldsymbol{\beta}}^\top \Sigma^* \hat{\boldsymbol{\beta}}}}\right).$$

The following theorem shows the convergence rate of $R(\hat{G}_{CHIME})$ to $R_{\text{opt}}(G_{\boldsymbol{\theta}^*})$, where $R_{\text{opt}}(G_{\boldsymbol{\theta}^*})$ is defined in (4.4).

Theorem 7. *Under the conditions of Theorem 6, if $T_0 \geq (-\log(\kappa))^{-1} \cdot \log(n \cdot d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*))$, the mis-clustering error $R(\hat{G}_{CHIME})$ for the classifier $\hat{G}_{CHIME}(\mathbf{z})$ defined in (4.26) satisfies*

$$R(\hat{G}_{CHIME}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) \lesssim \frac{s \log p}{n},$$

with probability at least $1 - o(1)$.

The result in Theorem 7 pushes forward the convergence rate of the mis-classification error of the LPD rule Cai and Liu (2011). In fact, Theorem 3 in Cai and Liu (2011) implies that the convergence rate is $R(\hat{G}_{LPD}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) = O((s \log p/n)^{1/2})$, over the parameter space $\Theta_p(s, c_\omega, M_1, M_2)$. Theorem 7 shows a faster rate and later in Section 4.3.3 we will show

that this convergence rate in the order of $(s \log p)/n$ is indeed optimal.

4.3.2. Initialization

As mentioned earlier, CHIME requires a good initialization $\hat{\boldsymbol{\theta}}^{(0)}$ that lies in the contraction basin $B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)$, defined in (4.19). This contraction basin forces the two inner products, $\boldsymbol{\delta}^\top \boldsymbol{\beta}^*$ and $(\boldsymbol{\delta}^*)^\top \boldsymbol{\beta}$ to be of the same order as $\Delta^2 = (\boldsymbol{\delta}^*)^\top \boldsymbol{\beta}^*$. In the special case where $\Sigma^* = \mathbf{I}_p$, this constraint reduces to the boundedness condition on the relative error of $\boldsymbol{\delta}$. The latter condition was used in Balakrishnan et al. (2017); Wang et al. (2014); Yi and Caramanis (2015), where they focused on the specialized mixture model $\frac{1}{2}N_p(-\boldsymbol{\mu}^*, \sigma^2 \mathbf{I}_p) + \frac{1}{2}N_p(\boldsymbol{\mu}^*, \sigma^2 \mathbf{I}_p)$. From a theoretical perspective, this condition guarantees that the weights $\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)})$ assigned in the E-step are close to the truth.

In the following, we introduce the initialization condition **(IC)**, which ensures that $\hat{\boldsymbol{\theta}}^{(0)} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)$.

(IC) For some permutation $\pi : \{1, 2\} \rightarrow \{1, 2\}$

$$\max_{k=1,2} \{ \|\hat{\boldsymbol{\mu}}_k^{(0)} - \boldsymbol{\mu}_{\pi(k)}^*\|_\infty \} \lesssim \frac{1}{s}, \quad |\hat{\Sigma}^{(0)} - \Sigma^*|_\infty \lesssim \frac{1}{s}.$$

The estimator $\hat{\boldsymbol{\theta}}^{(0)}$ satisfying **(IC)** can be obtained by the Hardt-Price algorithm. The Hardt-Price algorithm was proposed by (Hardt and Price, 2014, see algorithm B), which first established tight bounds for learning the parameters of a mixture of two univariate Gaussians using a variant of the method of moments Pearson (1894). They then extended the univariate result to the multivariate Gaussian mixture model and obtained the following theorem.

Proposition 1 (Hardt and Price (2014)). *Suppose we observe n i.i.d. samples $\mathbf{z}^{(i)}$ from model (4.2). Given $\epsilon, \nu > 0$, if $n = \Omega(\frac{1}{\epsilon^8} \log(\frac{p}{\nu} \log(\frac{1}{\epsilon})))$, then with probability at least $1 - \nu$, the Hardt-Price algorithm produces estimates $\hat{\boldsymbol{\mu}}_1^{(0)}$, $\hat{\boldsymbol{\mu}}_2^{(0)}$ and $\hat{\Sigma}^{(0)}$ such that for some*

permutation $\pi : \{1, 2\} \rightarrow \{1, 2\}$,

$$\max \left\{ \|\hat{\boldsymbol{\mu}}_1^{(0)} - \boldsymbol{\mu}_{\pi(1)}^*\|_\infty^2, \|\hat{\boldsymbol{\mu}}_2^{(0)} - \boldsymbol{\mu}_{\pi(2)}^*\|_\infty^2, |\hat{\Sigma}^{(0)} - \Sigma^*|_\infty \right\} \leq \epsilon \left(\frac{1}{4} \|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_\infty^2 + |\Sigma^*|_\infty \right).$$

Using Proposition 1, the following lemma shows that $\hat{\boldsymbol{\theta}}^{(0)}$ given by the Hardt-Price algorithm satisfies **(IC)**, and thus guarantees that the subsequent estimators $\hat{\boldsymbol{\theta}}^{(t)}$ in Algorithm 1 are contained in the contraction basin.

Lemma 11. *Let $\hat{\boldsymbol{\theta}}^{(0)}$ be the estimator constructed by the Hardt-Price algorithm. Under the conditions of Theorem 6, if $s(\frac{\log p}{n})^{1/12} = o(1)$, then for sufficiently large n , with probability $1 - o(1)$, $\hat{\boldsymbol{\theta}}^{(0)}$ satisfies **(IC)** and thus **(C1)** holds.*

Remark 8. *The conditions in Lemma 11 implies that the sample size $n \gtrsim s^{12} \log p$. To the best of our knowledge, the rate $n \gtrsim s^{12} \log p$ is so far the best for general Gaussian mixture models (without assuming spherical covariance matrix) in the literature (see, e.g., Hardt and Price (2014)). The optimality for the required sample size is an interesting problem for future work.*

4.3.3. Lower bounds

We now turn to the minimax lower bounds for the estimation of $\boldsymbol{\beta}^*$ and the mis-clustering error. Our results show that CHIME yields optimal results in the minimax sense, both for estimating the discriminating direction $\boldsymbol{\beta}^*$ and for clustering.

Theorem 8. *Under the conditions of Theorem 6, let \mathcal{C} be the set of all clustering rules based on n i.i.d. samples $\{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(n)}\}$ from model (4.2) with the true parameter $\boldsymbol{\theta}^* \in \Theta_p(s, c_\omega, M_1, M_2)$, for some constants $c_\omega, M_1, M_2 > 0$. If $\log p = O(\log(p/s))$, then*

(1).

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\theta}^* \in \Theta_p(s, c_\omega, M, M_b)} \mathbb{E} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \gtrsim \sqrt{\frac{s \log p}{n}},$$

(2).

$$\inf_{\hat{G} \in \mathcal{C}} \sup_{\boldsymbol{\theta}^* \in \Theta_p(s, c_\omega, M, M_b)} \mathbb{E}[R(\hat{G}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*})] \gtrsim \frac{s \log p}{n}.$$

Theorems, 6, 7 and 8 together show that our proposed estimator of $\boldsymbol{\beta}^*$ and the clustering rule attain the optimal rates of convergence.

Remark 9. Although a sparsity assumption on $\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$ seems to be more appealing in the Gaussian mixture model (4.2), Theorem 8 demonstrates that sparsity on $\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$ alone is not sufficient as the precision matrix Ω^* also plays an important role. Indeed, Theorem 8 shows that the difficulty of the problem depends on the sparsity of the product $\boldsymbol{\beta}^* = \Omega^*(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*)$. Therefore, a structural assumption directly on $\boldsymbol{\beta}^*$ is the most natural.

In the proof of Theorem 8, while the construction of the lower bound for the estimation of $\boldsymbol{\beta}^*$ is standard, that of the mis-clustering error is not straightforward. This is partially due to the fact that the risk function $R(\hat{G}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*})$ does not satisfy the triangle inequality. A key step is to reduce the above loss to an alternative risk function.

Let $G_{\boldsymbol{\theta}}$ be the optimal Fisher's classification rule defined with the parameter $\boldsymbol{\theta}$. For some generic classification rule G , we rewrite the risk function $R(G) - R(G_{\boldsymbol{\theta}}) = \mathbb{P}_{\boldsymbol{\theta}}(G(Z) \neq Y) - \mathbb{P}_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}(Z) \neq Y)$ and define $L_{\boldsymbol{\theta}}(G)$ by

$$L_{\boldsymbol{\theta}}(G) = \min_{\pi \in \mathcal{P}_2} \mathbb{P}_{\boldsymbol{\theta}}(G(Z) \neq \pi(G_{\boldsymbol{\theta}}(Z))).$$

The following lemma enables us to reduce the loss $R(\hat{G}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*})$ to the risk function $L_{\boldsymbol{\theta}^*}(\hat{G})$.

Lemma 12. *Let $Z \sim \frac{1}{2}N_p(\boldsymbol{\mu}_1, \Sigma) + \frac{1}{2}N_p(\boldsymbol{\mu}_2, \Sigma)$ with parameter $\boldsymbol{\theta} = (1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$. Suppose $\boldsymbol{\theta}$ satisfies $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq c_L$ for some $c_L > 0$. Then there exists some constant $m > 0$, such that if $L_{\boldsymbol{\theta}}(G) \leq 1/m$ for some classifier G , then*

$$\frac{1}{2m} L_{\boldsymbol{\theta}}^2(G) \leq \mathbb{P}_{\boldsymbol{\theta}}(G(Z) \neq Y) - \mathbb{P}_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}(Z) \neq Y).$$

Lemma 12 shows the relationship between the risk function $R(\hat{G}) - R_{\text{opt}}(G_{\theta^*})$ and a more ‘standard’ risk function $L_{\theta^*}(\hat{G})$. With Lemma 12, Theorem 8 can be proved by providing a lower bound for $L_{\theta^*}(\hat{G})$. The risk function $L_{\theta^*}(\hat{G})$ has been studied in Azizyan et al. (2013) for a specialized model $\frac{1}{2}N(\boldsymbol{\mu}_1, \sigma^2 \mathbf{I}_p) + \frac{1}{2}N(\boldsymbol{\mu}_2, \sigma^2 \mathbf{I}_p)$. Although no matching upper and lower bounds were provided, the following lemma in Azizyan et al. (2013) is crucial to our analysis, which shows the triangle inequality property of the risk function $L_{\theta^*}(\hat{G})$. For two probability density functions \mathbb{P}_{θ_1} and \mathbb{P}_{θ_2} , denote their KL divergence by

$$\text{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) = \int \mathbb{P}_{\theta_1}(z) \log \frac{\mathbb{P}_{\theta_1}(z)}{\mathbb{P}_{\theta_2}(z)} dz.$$

Lemma 13 (Azizyan et al. (2013)). *For any $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta_p(s, c_\omega, M, M_b)$ and any clustering \hat{G} , if $L_{\boldsymbol{\theta}}(G_{\tilde{\boldsymbol{\theta}}}) + L_{\boldsymbol{\theta}}(\hat{G}) + \sqrt{\frac{\text{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})}{2}} \leq 1/2$, then*

$$L_{\boldsymbol{\theta}}(G_{\tilde{\boldsymbol{\theta}}}) - L_{\boldsymbol{\theta}}(\hat{G}) - \sqrt{\frac{\text{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})}{2}} \leq L_{\tilde{\boldsymbol{\theta}}}(\hat{G}) \leq L_{\boldsymbol{\theta}}(G_{\tilde{\boldsymbol{\theta}}}) + L_{\boldsymbol{\theta}}(\hat{G}) + \sqrt{\frac{\text{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})}{2}}.$$

After applying Lemmas 12 and 13, we then use Fano’s inequality to complete the proof of Theorem 8. The details are shown in Section 4.8.

4.4. Low-dimensional Gaussian Mixtures

Although the focus of the present paper is on the high-dimensional setting, our analysis can be modified to establish the optimality of the clustering procedure for the low-dimensional Gaussian mixtures that is based on the classical EM algorithm. In the general low-dimensional setting, we consider the model

$$\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \dots, \mathbf{z}^{(n)} \stackrel{i.i.d.}{\sim} (1 - \omega^*)N_p(\boldsymbol{\mu}_1^*, \Sigma^*) + \omega^*N_p(\boldsymbol{\mu}_2^*, \Sigma^*), \quad (4.27)$$

without imposing any assumption on the sparsity of the discriminant direction. In such case, direct estimation of $\boldsymbol{\beta}^*$ is not needed. The clustering procedure under model (4.27), which uses the classical EM algorithm to estimate ω^* , $\boldsymbol{\mu}_1^*$, $\boldsymbol{\mu}_2^*$ and Σ^* , is summarized in

Algorithm 2. We call it CLOME for **C**lustering of **L**Ow-dimensional **G**aussian **M**ixtures with the **E**M.

Algorithm 2 Clustering of **L**Ow-dimensional **G**aussian **M**ixtures with the **E**M (CLOME)

- 1: **Inputs:** Initializations $\hat{\omega}^{(0)}, \hat{\boldsymbol{\mu}}_1^{(0)}, \hat{\boldsymbol{\mu}}_2^{(0)}$ and $\hat{\Sigma}^{(0)}$, maximum number of iterations T_0 .
- 2: **for** $t = 0, 1, \dots, T_0 - 1$ **do**
- 3: **E-Step:** Evaluate $Q_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$ with $\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\mathbf{z}^{(i)})$ defined in (4.8).
- 4: **M-Step:** Update $\hat{\omega}^{(t+1)}, \hat{\boldsymbol{\mu}}_1^{(t+1)}, \hat{\boldsymbol{\mu}}_2^{(t+1)}$, and $\hat{\Sigma}^{(t+1)}$ via (4.10), (4.11), (4.12) and (4.13).
- 5: **end for**
- 6: Output $\hat{\omega}^{(T_0)}, \hat{\boldsymbol{\mu}}_1^{(T_0)}, \hat{\boldsymbol{\mu}}_2^{(T_0)}$ and $\hat{\Sigma}^{(T_0)}$.
- 7: Construct the clustering rule

$$\hat{G}_{EM}(\mathbf{z}) = \begin{cases} 1, & \left\{ \mathbf{z} - \frac{\hat{\boldsymbol{\mu}}_1^{(T_0)} + \hat{\boldsymbol{\mu}}_2^{(T_0)}}{2} \right\}^\top (\hat{\Sigma}^{(T_0)})^{-1} (\hat{\boldsymbol{\mu}}_1^{(T_0)} - \hat{\boldsymbol{\mu}}_2^{(T_0)}) \geq \log\left(\frac{\hat{\omega}^{(T_0)}}{1 - \hat{\omega}^{(T_0)}}\right), \\ 2, & \left\{ \mathbf{z} - \frac{\hat{\boldsymbol{\mu}}_1^{(T_0)} + \hat{\boldsymbol{\mu}}_2^{(T_0)}}{2} \right\}^\top (\hat{\Sigma}^{(T_0)})^{-1} (\hat{\boldsymbol{\mu}}_1^{(T_0)} - \hat{\boldsymbol{\mu}}_2^{(T_0)}) < \log\left(\frac{\hat{\omega}^{(T_0)}}{1 - \hat{\omega}^{(T_0)}}\right). \end{cases}$$

The technical tools developed for the proofs of Theorems 6, 7 and 8 can be used to establish the optimality of CLOME in Algorithm 2. We consider the theoretical performance of estimation and the CLOME clustering procedure over the parameter space $\Theta_p(c_\omega, M_1, M_2)$, defined by

$$\Theta_p(c_\omega, M_1, M_2) = \{ \boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p \times p}, \Sigma \succ 0, \\ \omega \in (c_\omega, 1 - c_\omega), \|\Sigma\|_2 \leq M_1, \|\boldsymbol{\mu}_k\|_2 \leq M_2, k = 1, 2 \}.$$

Similar to the high-dimensional setting, CLOME requires a good initialization. The initial value $\hat{\boldsymbol{\theta}}^{(0)}$ should lie in the contraction basin $\tilde{B}_{con}(\boldsymbol{\theta}^*; c_0)$,

$$\tilde{B}_{con}(\boldsymbol{\theta}^*; c_0) = \{ \boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \omega \in (c_0, 1 - c_0), \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p \times p}, \\ \Sigma \succ 0, \|\Sigma - \Sigma^*\|_2 \leq \frac{1}{4} \phi_{\min}(\Sigma^*), \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\|_2 \leq \frac{1}{4 \|\Sigma\|_2} \|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_2, k = 1, 2 \}.$$

Indeed, in the low-dimensional regime, the algorithm proposed by Ge et al. (2015), which is based on the method of moments, was proved to satisfy the above condition (see Theorem

3.4 of Ge et al., 2015).

We are ready to provide the upper bound results of CLOME under the low-dimensional Gaussian mixture model (4.27). For any $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta_p(c_\omega, M_1, M_2)$, define the ℓ_2 distance between $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ by

$$d_2(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = |\omega - \tilde{\omega}| + \|\boldsymbol{\mu}_1 - \tilde{\boldsymbol{\mu}}_1\|_2 + \|\boldsymbol{\mu}_2 - \tilde{\boldsymbol{\mu}}_2\|_2 + \|\Sigma - \tilde{\Sigma}\|_2.$$

Theorem 9. *Consider the model (4.27) over the parameter space*

$\Theta_p(c_\omega, M_1, M_2)$ *where* $p = o(n)$. *Suppose the initialization* $\hat{\boldsymbol{\theta}}^{(0)} \in \tilde{B}_{con}(\boldsymbol{\theta}^*; c_0)$ *and* $\Delta^2 > \log(16M_1/3 + 64M_2/3)$. *Then there exist constants* $\kappa \in (0, 1), C_1, C_2 > 0$, *such that with probability at least* $1 - n^{-1}$, *the outputs* $\hat{\boldsymbol{\mu}}_1^{(T_0)}, \hat{\boldsymbol{\mu}}_2^{(T_0)}$ *and* $\hat{\Sigma}^{(T_0)}$ *of Algorithm 2 satisfy*

$$\begin{aligned} \|\hat{\boldsymbol{\mu}}_k^{(T_0)} - \boldsymbol{\mu}_k^*\|_2 &\leq \kappa^{T_0} d_2(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(0)}) + C_1 \sqrt{\frac{p}{n}}, \quad k = 1, 2; \\ \|\hat{\Sigma}^{(T_0)} - \Sigma^*\|_2 &\leq \kappa^{T_0} d_2(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(0)}) + C_2 \sqrt{\frac{p}{n}}. \end{aligned}$$

In particular, if $T_0 \geq 2(-\log(\kappa))^{-1} \log(nd_2(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(0)})/p)$, *then there exists a constant* $C_3 > 0$, *such that*

$$d_2(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(T_0)}) \leq C_3 \sqrt{\frac{p}{n}} \quad \text{and} \quad R(\hat{G}_{EM}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) \leq C_3 \frac{p}{n}.$$

Remark 10. Theorem 9 provides upper bound results for the estimators given in Algorithm 2 under a general Gaussian mixture model in (4.27), and shows that CLOME is consistent if the initialization $\hat{\boldsymbol{\theta}}^{(0)}$ lies in the contraction basin $\tilde{B}_{con}(\boldsymbol{\theta}^*; c_0)$. Applying Theorem 9 to the special case $\frac{1}{2}N_p(-\boldsymbol{\mu}^*, \sigma^2 \mathbf{I}_p) + \frac{1}{2}N_p(\boldsymbol{\mu}^*, \sigma^2 \mathbf{I}_p)$ leads to the same result as that in Balakrishnan et al. (2017).

We establish the optimality of Algorithm 2 for both the estimators and the clustering rule by providing the following lower bound results.

Theorem 10. *Under the conditions of Theorem 9, we have*

$$\begin{aligned} \inf_{\hat{\boldsymbol{\mu}}_k} \sup_{\boldsymbol{\theta}^* \in \Theta_p(c_\omega, M_1, M_2)} \mathbb{E} \|\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k^*\|_2 &\gtrsim \sqrt{\frac{p}{n}}, \quad k = 1, 2; \\ \inf_{\hat{\Sigma}} \sup_{\boldsymbol{\theta}^* \in \Theta_p(c_\omega, M_1, M_2)} \mathbb{E} \|\hat{\Sigma} - \Sigma^*\|_2 &\gtrsim \sqrt{\frac{p}{n}}, \\ \inf_{\hat{G} \in \mathcal{C}} \sup_{\boldsymbol{\theta}^* \in \Theta_p(c_\omega, M_1, M_2)} R(\hat{G}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) &\gtrsim \frac{p}{n}. \end{aligned}$$

Theorems 9 and 10 together characterize the optimality of CLOME. Note that in the low-dimensional case the estimators $\hat{\boldsymbol{\mu}}_k^{(T_0)}$ and $\hat{\Sigma}^{(T_0)}$ achieve the same convergence rate as the MLE obtained with known sample labels.

4.5. Simulations

The proposed CHIME procedure is easily implementable. In this section we conduct simulation studies to investigate the numerical properties of CHIME under various settings.

We compare the performance of CHIME with the k -means (KM), sparse k -means (SKM, Witten and Tibshirani, 2010), Influential Feature PCA (IF-PCA, Jin et al., 2016b), penalized model-based clustering with common covariance matrices (PCCM, Zhou et al., 2009), sparse clustering via HardtPrice (SHP, Azizyan et al., 2015), the linear programming discriminant rule (LPD, Cai and Liu, 2011) and the oracle Fisher’s rule obtained by plugging in the true parameters (Oracle) on a suite of three simulated examples. Three methods including SKM, PCCM and SHP were implemented in **R**, whereas the others were implemented in **MATLAB**. We refer readers to Cai et al. (2018b) for additional simulation scenarios—including unequal mixing proportion case and settings with discriminant vectors of different sparsity levels—and subsequent discussion.

In all simulations, the sample size is $n = 200$ while the number of variables p varies from 100, 200, 500 to 800. The probability of being in either of the two classes is equal, i.e. $\omega^* = 0.5$. The discriminant vector $\boldsymbol{\beta}^* \propto (1, \dots, 1, 0, \dots, 0)^\top$ is sparse such that only the

first $s = 10$ entries are nonzero. We consider the following three models for the inverse covariance matrix Ω^* .

Model 1 Erdős-Rényi random graph: Let $\tilde{\Omega} = (\tilde{\omega}_{ij})$ where $\tilde{\omega}_{ij} = u_{ij}\delta_{ij}$, $\delta_{ij} \sim \text{Ber}(1, 0.05)$ being the Bernoulli random variable with success probability 0.05 and $u_{ij} \sim \text{Unif}[0.5, 1] \cup [-1, -0.5]$. After symmetrizing $\tilde{\Omega}$, set $\Omega^* = \tilde{\Omega} + \{\max(-\phi_{\min}(\tilde{\Omega}), 0) + 0.05\}\mathbf{I}_p$ to ensure the positive definiteness. Finally, Ω^* is standardized to have unit diagonals.

Model 2 Block sparse model: $\Omega^* = (\mathbf{B} + \delta\mathbf{I}_p)/(1 + \delta)$ where $b_{ij} = b_{ji} = 0.5 * \text{Ber}(1, 0.3)$ for $1 \leq i \leq s, i < j \leq p$; $b_{ij} = b_{ji} = 0.5$ for $s + 1 \leq i < j \leq p$; $b_{ii} = 1$ for $1 \leq i \leq p$. In other words, only the first s rows and columns of Ω^* are sparse, whereas the rest of the matrix is not sparse. Here $\delta = \max(-\phi_{\min}(\mathbf{B}), 0) + 0.05$. The matrix Ω^* is also standardized to have unit diagonals.

Model 3 AR(1) model: $\Omega^* = (\Omega_{ij}^*)_{p \times p}$ with $\Omega_{ij}^* = 0.8^{|i-j|}$.

In both Model 1 and Model 2, the vector $\beta^* = (1, \dots, 1, 0, \dots, 0)^\top$. To ensure sufficiently strong signals in Model 3, we increase the magnitude of nonzero entries in β^* such that $\beta^* = 2.5 \cdot (1, \dots, 1, 0, \dots, 0)^\top$. Given the inverse covariance matrix Ω^* , the mean of class 1 is $\mu_1^* = \mathbf{0}$ and mean of class 2 is $\mu_2^* = \mu_1^* - (\Omega^*)^{-1}\beta^*$.

To find initializations for use in CHIME, we first run the k -means algorithm to find the initial class labels, and calculate $\mu_1^{(0)}$ and $\mu_2^{(0)}$. The pooled sample covariance matrices $\hat{\Sigma}^{(0)}$ is used as the initial value for the covariance matrix. We recommend running CHIME with multiple random initial class labels to obtain the best possible clustering and estimation results. In the case of Model 3, class labels estimated from SKM are sometimes more accurate than those from the k -means algorithm, and are thus used as candidates for initializing the parameters needed in CHIME.

As with any other penalization-based methods, CHIME, SKM, SHP, PCCM and LPD all require selecting a tuning parameter. To this end, we generated independently training

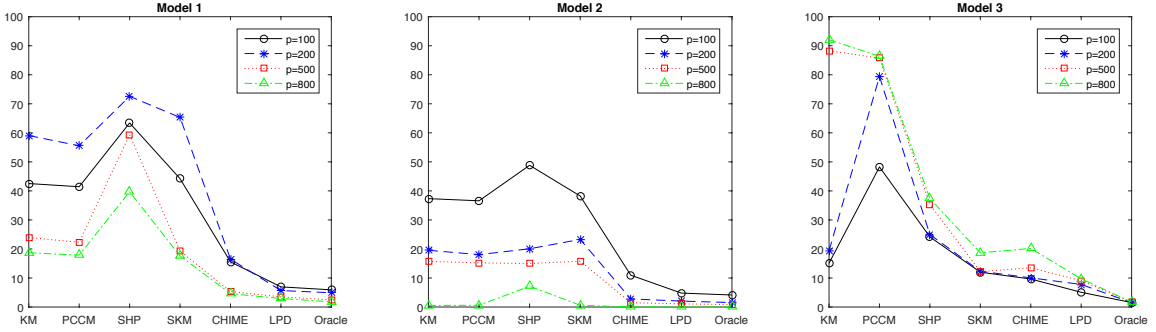


Figure 1: Average mis-clustering errors based on $n = 200$ test samples from 100 replications under Model 1 (left), Model 2 (middle) and Model 3 (right). CHIME performs well in all three models.

data and test data from the same distribution. For a given λ , the training data were first used to estimate the parameters, with mis-clustering error evaluated based on the test data. The optimal λ was selected as the one that minimizes the mis-clustering errors over the test data. If there are multiple λ 's that yield the same mis-clustering error, then the largest one will be selected. The tuning for SKM follows a slightly different procedure as the penalty parameter is specified in terms of an upper bound for a sequence of weights. The training data were first used to find the optimal upper bound, with mis-clustering error further evaluated on the test data under the optimal upper bound.

Figure 1 summarizes the average mis-clustering errors for different methods under the three aforementioned settings, with respective standard errors (s.e.) presented in Table 14. All comparisons were evaluated from 100 replications based on $n = 200$ test samples. Note the LPD rule is a supervised method for classification and is included as a benchmark comparison with the proposed method CHIME.

CHIME outperforms all other unsupervised clustering methods in both Models 1 and 2. Moreover, the mis-clustering errors from CHIME are comparable to those from LPD for $p = 500, 800$ in Model 1 and $p = 200, 500, 800$ in Model 2. In comparison, KM, SKM and PCCM yield rather similar performances, with IF-PCA showing the worst performances in all three models, since IF-PCA is designed for the case of “rare and weak” signal Jin et al. (2016b)

Table 14: Average mis-clustering errors (s.e.) based on $n = 200$ test samples from 100 replications under three different models

| | p | 100 | 200 | 500 | 800 |
|---------|--------|--------------|--------------|--------------|---------------|
| Model 1 | KM | 42.53(6.81) | 59.07(7.67) | 23.94(6.53) | 18.72(4.32) |
| | PCCM | 41.43(5.63) | 55.53(7.41) | 22.31(5.60) | 17.87(3.93) |
| | SHP | 64.33(16.38) | 72.34(13.91) | 58.28(18.79) | 51.33(16.78) |
| | SKM | 44.20(6.02) | 65.30(8.87) | 19.29(6.42) | 17.59(3.91) |
| | IF-PCA | 92.73(5.87) | 94.50(4.58) | 94.98(4.59) | 94.03(3.94) |
| | CHIME | 16.21(6.21) | 15.37(9.97) | 5.21(3.03) | 4.79(1.99) |
| | LPD | 6.94(2.49) | 5.67(2.22) | 3.51(2.02) | 2.94(1.58) |
| | Oracle | 5.92(2.46) | 4.92(2.13) | 2.44(1.64) | 1.79(1.24) |
| Model 2 | KM | 37.33(5.82) | 19.54(4.33) | 15.71(3.57) | 0.54(0.72) |
| | PCCM | 36.59(6.26) | 18.05(4.23) | 15.20(3.34) | 0.60(0.75) |
| | SHP | 51.54(20.14) | 20.07(16.71) | 14.98(9.84) | 7.16(6.75) |
| | SKM | 38.23(6.18) | 23.28(4.89) | 15.78(3.67) | 0.60(0.72) |
| | IF-PCA | 75.25(22.16) | 82.65(15.99) | 87.55(10.40) | 91.15(8.94) |
| | CHIME | 9.62(4.92) | 3.35(2.18) | 2.07(1.46) | 0.03(0.21) |
| | LPD | 4.80(2.42) | 2.04(1.39) | 1.09(0.96) | 0.03(0.17) |
| | Oracle | 4.14(2.20) | 1.53(1.23) | 0.81(0.86) | 0.01(0.10) |
| Model 3 | KM | 15.08(4.49) | 19.39(9.47) | 47.68(22.73) | 65.19(20.57) |
| | PCCM | 48.52(36.81) | 79.38(18.67) | 85.72(3.47) | 86.37(3.64) |
| | SHP | 24.13(20.92) | 24.96(19.90) | 35.37(22.17) | 37.34(24.64) |
| | SKM | 12.00(3.17) | 12.32(3.28) | 12.21(3.28) | 18.66(20.99) |
| | IF-PCA | 88.17(11.19) | 93.00(6.50) | 92.98(7.22) | 93.83(5.2456) |
| | CHIME | 8.96(2.89) | 9.75(2.87) | 12.94(3.76) | 19.97(20.14) |
| | LPD | 5.08(2.41) | 7.77(2.69) | 8.98(2.85) | 9.65(2.76) |
| | Oracle | 1.53(1.34) | 1.52(1.29) | 1.73(1.24) | 1.69(1.19) |

and requires a diagonal covariance matrix. Clustering with SHP generally returns large mis-clustering errors and large standard errors, including in Model 3. This is due to its use of the moment-based estimator from the Hardt-Price algorithm for parameter initializations. The Hardt-Price algorithm requires a good pivot, i.e. one out of the p variables that shows the largest difference between the two cluster centers, to get a reasonable initialization. Such a pivot might be especially difficult to find in Model 1 as a majority of entries in $\boldsymbol{\mu}_2^*$ are randomly distributed around zero.

Clustering with Model 3 is more challenging due to the special structure of the inverse covariance matrix. Indeed, Ω^* in Model 3 is not sparse. Nonetheless CHIME maintains

its good performance and achieves mis-clustering errors that are comparable to those from SKM and smaller than those from other clustering methods. On the other hand, since μ_2^* is exactly sparse with $s + 1$ nonzero entries by construction, SKM shows significant improvement over KM, especially for large p , by taking advantage of sparsity in the true mean parameters. PCCM performs poorly in Model 3 and worse than KM for $p = 500, 800$, mainly because of its poor performance in estimating the non-sparse precision matrix. In the case of large p , it also suffers from poor initializations with the k -means algorithm.

4.6. Applications to Glioblastoma Gene Expression Data

To illustrate the proposed CHIME procedure, we consider in this section an application based on glioblastoma gene expression data. Glioblastoma (GBM) is the most common and aggressive form of brain cancer in adults. In order to provide the best treatments for patients with glioblastoma, an important question is classification of GBM subtypes, as different subtypes may respond to treatments differently. In a well-known paper, Verhaak et al. (2010) introduced a robust gene expression-based molecular classification of GBM into Proneural, Neural, Classical and Mesenchymal subtypes. The data are available at https://tcga-data.nci.nih.gov/docs/publications/gbm_exp/. In this study, 200 GBM and two normal brain samples assayed across three gene expression platforms were first integrated into a single unified dataset. After further filtering, there remain 1740 genes with consistent but highly variable expression across the platforms. The 202 samples were hierarchically clustered using the consensus average linkage. Based on the silhouette width, 173 of the 202 samples were selected as the “core samples” for being most representative of the clusters. Thus our following analysis was based solely on the core samples.

To validate the performance of CHIME in recovering the labels of a two-component Gaussian mixture, we focused on two of the four identified GBM subtypes: Mesenchymal and Neural, yielding a total of 82 samples among which 56 are from the Mesenchymal group. For the purpose of clustering, one can use the full set of 1740 genes, or select a subset of them. As the samples are pre-selected, direct application of any clustering methods on the full

set yields almost perfect match between the estimated clusters and the true ones. We thus followed the latter approach and chose $p = 200$ genes from the full set of 1740 genes. In particular, we considered gene selection as follows. First we calculated the variances of all the genes and ranked them in a decreasing order. The top 20 genes with the largest variances and the last 180 genes with the smallest variances were then selected as the training set. We anticipate that the high variance genes are more informative than low variance genes, although this is not always true as the results below show.

Since we do not have a separate test data with labels, we propose to select the tuning parameter required in CHIME via a stability approach, motivated by Tibshirani and Walther (2005). The idea is to first randomly split the data into the training set and the test set. For a given λ , we run CHIME on the test data and obtain class labels of the test data, run CHIME on the training data, and finally measure how well the parameters estimated from the training data predict the class labels of the test data. Formally, let $f(X)$ be a clustering operation learned from data X and $G[f(\cdot), X]$ be the class labels estimated on X based on the clustering operation $f(\cdot)$. The prediction strength is then defined as the average adjusted random index when comparing $G[f(X_{train}), X_{test}]$ to $G[f(X_{test}), X_{test}]$ over B replications:

$$ps(\lambda) = \frac{1}{B} \sum_{i=1}^B \text{ARI}(G[f(X_{train}^i), X_{test}^i], G[f(X_{test}^i), X_{test}^i]). \quad (4.28)$$

The optimal λ^* is selected as $\arg \max_{\lambda} ps(\lambda)$. Note the adjusted rand index is preferred over the rand index as the former has the advantage of being corrected-for-chance. This is especially important since if $\hat{\beta} = 0$ due to the large penalty, $G[f(X_{train}), X_{test}]$ can randomly coincide with $G[f(X_{test}), X_{test}]$, resulting in a large value in rand index, but not in terms of the adjusted rand index. In addition, we define the prediction strength in terms of the adjusted rand index rather than the original one proposed in Tibshirani and Walther (2005), as the former favors a larger penalty parameter and thus returns a sparser estimate that is more interpretable.

To apply CHIME, SHP and PCCM, we first selected the tuning parameters by maximizing the prediction strength defined in (4.28). The tuning parameter required in SKM was selected via criteria defined in Witten and Tibshirani (2010). Sparse clustering of the 200 genes with CHIME at the optimal λ yields 2 errors. A comparison with other clustering methods reveals that CHIME performs the best in recovering the correct sample labels, as shown in Table 15. Among all other methods, SHP yields the largest error, possibly due to incorrect parameter initializations with the Hardt-Price algorithm.

Table 15: Clustering results for the GBM gene expression data with $p = 200$ genes and 82 samples

| Class | CHIME | | KM | | PCCM | | SHP | | SKM | |
|-------------|-------|----|----|----|------|----|-----|----|-----|----|
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| Neural | 26 | 0 | 26 | 0 | 26 | 0 | 12 | 14 | 25 | 1 |
| Mesenchymal | 2 | 54 | 7 | 49 | 5 | 51 | 10 | 46 | 6 | 50 |

To understand the performance of CHIME better, we also looked at the selected informative variables, i.e. genes with nonzero coefficients in $\hat{\beta}$. Figure 2 shows that large marginal variances do not necessarily imply large coefficients in $|\hat{\beta}|$. In fact, a significant number of the low variance genes (59 of 180) turn out to be informative for the clustering. This again confirms that direct estimation of the discriminant vector with CHIME yields a better characterization of the clustering boundary than estimating separately the cluster mean differences and (partial) correlations among variables.

4.7. Extensions to Multi-class Gaussian Mixtures

The proposed method can be readily extended to Gaussian mixtures with K ($K \geq 2$) components. Consider the model

$$\Pr(Y = k) = \omega_k^*, \quad Z | Y = k \sim N_p(\boldsymbol{\mu}_k^*, \Sigma^*), \quad k = 1, \dots, K.$$

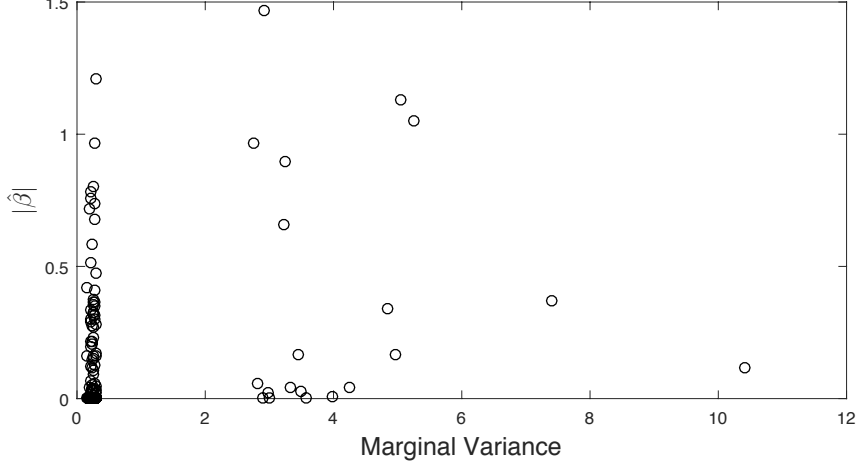


Figure 2: The discriminant vector $|\hat{\beta}|$ is plotted against the marginal variances.

Here $\sum_{k=1}^K \omega_k^* = 1$. We assume K is fixed and known. In the ideal case where the parameters are known, the oracle Bayes rule yields the label assignment

$$\hat{Y} = \arg \max_{k=1, \dots, K} \left\{ \beta_k^{*\top} (Z - (\mu_k^* + \mu_1^*)/2) + \log \omega_k^* \right\}, \quad (4.29)$$

where $\beta_k^* = (\Sigma^*)^{-1}(\mu_k^* - \mu_1^*)$ ($k = 1, 2, \dots, K$) are the discriminant directions. By definition, the vector β_1^* is trivial.

When neither the parameters nor the sample labels are known, under the assumption that the discriminant directions β_k^* ($k = 2, \dots, K$) are sparse, CHIME can be generalized for clustering multi-class Gaussian mixtures. Specifically, denote the posterior probability of the i -th sample in class k by

$$\hat{\gamma}_{ik}^{(t)} := \Pr(y_i = k | \mathbf{z}^{(i)}, \hat{\theta}^{(t)}) = \frac{\hat{\omega}_k^{(t)} f(\mathbf{z}^{(i)} | \hat{\mu}_k^{(t)}, \hat{\beta}^{(t)})}{\sum_{\ell=1}^K \hat{\omega}_\ell^{(t)} f(\mathbf{z}^{(i)} | \hat{\mu}_\ell^{(t)}, \hat{\beta}^{(t)})}.$$

The conditional log-likelihood at the t -th step becomes

$$Q_n(\theta | \hat{\theta}^{(t)}) = -\frac{1}{2n} \sum_{\substack{i \in [n] \\ k \in [K]}} \hat{\gamma}_{ik}^{(t)} (\mathbf{z}^{(i)} - \hat{\mu}_k^{(t)})^T \Omega (\mathbf{z}^{(i)} - \hat{\mu}_k^{(t)}) + \frac{1}{n} \sum_{\substack{i \in [n] \\ k \in [K]}} \hat{\gamma}_{ik}^{(t)} \log \hat{\omega}_k^{(t)} + \frac{1}{2} \log |\Omega|.$$

The updates of ω_k , μ_k and Σ in the M-step are respectively,

$$\begin{aligned}\hat{\omega}_k^{(t+1)} &= \frac{1}{n} \sum_{i=1}^n \hat{\gamma}_{ik}^{(t)}, \hat{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^n \hat{\gamma}_{ik}^{(t)} \mathbf{z}^{(i)}}{\sum_{i'=1}^n \hat{\gamma}_{i'k}^{(t)}}, \\ \hat{\Sigma}^{(t+1)} &= \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \hat{\gamma}_{ik}^{(t)} (\mathbf{z}^{(i)} - \hat{\mu}_k^{(t+1)}) (\mathbf{z}^{(i)} - \hat{\mu}_k^{(t+1)})^\top.\end{aligned}$$

Finally, $\hat{\beta}_k$'s are updated by solving the following optimizations:

$$\hat{\beta}^{(t+1)} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \beta^\top \hat{\Sigma}^{(t+1)} \beta - \beta^\top (\hat{\mu}_k^{(t+1)} - \hat{\mu}_1^{(t+1)}) + \lambda_n^{(t+1)} \|\beta\|_1 \right\}, \quad k = 2, \dots, K.$$

This algorithm assumes sparsity of each discriminant direction β_k^* ($k = 2, \dots, K$), but no conditions on their joint support. If it is believed that the discriminant vectors have similar support, one might impose a group lasso penalty for their estimation, as done in Mai et al. (2015).

The final clustering rule is constructed by plugging the estimates ω^* , μ_k^* ($k = 1, \dots, K$) and β_k^* ($k = 2, \dots, K$) into the optimal rule (4.29). Provided with a good initialization, similar techniques introduced in previous sections can be used to establish the convergence rate of $\hat{\beta}_k$ as well as the upper and lower bounds of the mis-clustering error under suitable regularity conditions. The initialization for clustering multi-class Gaussian mixtures can be obtained by algorithms in Moitra and Valiant (2010) or Ge et al. (2015). It was shown that the estimate lies in B_{con} with probability at least $1 - \delta$ when $n > \text{poly}(p, \frac{1}{\delta}, \frac{1}{\Delta})$, where $\text{poly}(\cdot)$ denotes the polynomial dependence. We also note here that the initialization step is of much importance in the multi-class setting, since it has been shown in Jin et al. (2016a) that the EM algorithm could stuck at a local optimum without a good initialization.

4.8. Proofs

In this section, we prove the optimality for the mis-clustering error, i.e. Theorem 7 and the part (2) of Theorem 8. The proof of the optimality for the estimation error, Theorem 6 and part (1) of Theorem 8, is given in the supplement Cai et al. (2018b). A few technical

lemmas are needed for the proof of the main results. These technical lemmas as well as some other minor results are proved in the supplement Cai et al. (2018b).

4.8.1. Proof of Theorem 7

We start with the following lemma.

Lemma 14. *For two vectors $\boldsymbol{\gamma}^*$ and $\hat{\boldsymbol{\gamma}}$, if $\|\boldsymbol{\gamma}^* - \hat{\boldsymbol{\gamma}}\|_2 \leq \|\boldsymbol{\gamma}^*\|_2$, and $\|\boldsymbol{\gamma}^*\|_2 \geq c$ for some constant $c > 0$, then*

$$(\boldsymbol{\gamma}^*)^\top \hat{\boldsymbol{\gamma}} - \|\boldsymbol{\gamma}^*\|_2 \cdot \|\hat{\boldsymbol{\gamma}}\|_2 \asymp \|\boldsymbol{\gamma}^* - \hat{\boldsymbol{\gamma}}\|_2^2.$$

Consider the model (4.2). Given the estimators $\hat{\omega}$, $\hat{\boldsymbol{\mu}}_k$, and $\hat{\boldsymbol{\beta}}$, the sample \mathbf{z} is classified as

$$\hat{G}(\mathbf{z}) = \begin{cases} 1, & (\mathbf{z} - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2)^\top \hat{\boldsymbol{\beta}} \geq \log(\frac{\hat{\omega}}{1-\hat{\omega}}) \\ 2, & (\mathbf{z} - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2)^\top \hat{\boldsymbol{\beta}} < \log(\frac{\hat{\omega}}{1-\hat{\omega}}). \end{cases}$$

Let $\tau^* = \frac{\omega^*}{1-\omega^*}$, $\hat{\tau} = \frac{\hat{\omega}}{1-\hat{\omega}}$ and $\hat{\Delta} = \sqrt{\hat{\boldsymbol{\beta}}^\top \Sigma^* \hat{\boldsymbol{\beta}}}$. The mis-clustering error is

$$R(\hat{G}) = (1 - \omega^*) \Phi\left(\frac{\log \hat{\tau} + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1^*)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}}\right) + \omega^* \bar{\Phi}\left(\frac{\log \hat{\tau} + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2^*)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}}\right),$$

with $R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) = (1 - \omega^*) \Phi\left(\frac{\log \tau^* - \Delta^2/2}{\Delta}\right) + \omega^* \bar{\Phi}\left(\frac{\log \tau^* + \Delta^2/2}{\Delta}\right)$. Define an intermediate quantity

$$R^* = (1 - \omega^*) \Phi\left(\frac{\log \tau^* - (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right) + \omega^* \bar{\Phi}\left(\frac{\log \tau^* + (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right).$$

We first show that $R^* - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) \lesssim \frac{s \log p}{n}$. Applying Taylor's expansion to the two terms in R^* at $\frac{\log \tau^*}{\hat{\Delta}} - \frac{\Delta}{2}$ and $\frac{\log \tau^*}{\hat{\Delta}} + \frac{\Delta}{2}$ respectively, we obtain

$$\begin{aligned} R^* - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) &= (1 - \omega^*) \left(\frac{\log \tau^*}{\hat{\Delta}} - \frac{(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\log \tau^*}{\Delta} + \frac{\Delta}{2} \right) \Phi' \left(\frac{\log \tau^*}{\hat{\Delta}} - \frac{\Delta}{2} \right) \\ &\quad - \omega^* \left(\frac{\log \tau^*}{\hat{\Delta}} + \frac{(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\log \tau^*}{\Delta} - \frac{\Delta}{2} \right) \Phi' \left(\frac{\log \tau^*}{\hat{\Delta}} + \frac{\Delta}{2} \right) + O_P\left(\frac{s \log p}{n}\right), \end{aligned} \quad (4.30)$$

where the remaining term is bounded by using the facts that

$$\left(\frac{\log \tau^*}{\hat{\Delta}} + \frac{(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\log \tau^*}{\Delta} - \frac{\Delta}{2}\right)^2 = O_p\left(\frac{s \log p}{n}\right), \text{ and } \Phi'' = O(1).$$

In fact,

$$\begin{aligned} & \left| \frac{\log \tau^*}{\hat{\Delta}} + \frac{(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\log \tau^*}{\Delta} - \frac{\Delta}{2} \right| \leq \left| \frac{\log \tau^*}{\hat{\Delta}} - \frac{\log \tau^*}{\Delta} \right| + \left| \frac{(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\Delta}{2} \right| \\ & \leq \left| \frac{\log \tau^*}{\hat{\Delta}} - \frac{\log \tau^*}{\Delta} \right| + \left| \frac{(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\Delta^2}{2\hat{\Delta}} \right| + \left| \frac{\Delta^2}{2\hat{\Delta}} - \frac{\Delta}{2} \right| \lesssim |\hat{\Delta} - \Delta| + |(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}} - (\boldsymbol{\delta}^*)^\top \boldsymbol{\beta}^*| \\ & \leq \sqrt{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \Sigma^* (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)} + |(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}} - (\boldsymbol{\delta}^*)^\top \boldsymbol{\beta}^*| \lesssim \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \lesssim \sqrt{\frac{s \log p}{n}}. \end{aligned} \quad (4.31)$$

Recall that $\tau^* = \frac{\omega^*}{1-\omega^*}$, (4.30) can be further expanded such that

$$\begin{aligned} \frac{R^* - R_{\text{opt}}(G_{\boldsymbol{\theta}^*})}{\sqrt{(1-\omega^*)\omega^*}} & \asymp \left(\frac{\log \tau^*}{\hat{\Delta}} - \frac{(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\log \tau^*}{\Delta} + \frac{\Delta}{2} \right) e^{-\frac{1}{2} \left(\frac{\log \tau^*}{\hat{\Delta}} - \frac{\Delta}{2} \right)^2 - \frac{\log \tau^*}{2}} \\ & \quad - \left(\frac{\log \tau^*}{\hat{\Delta}} + \frac{(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\log \tau^*}{\Delta} - \frac{\Delta}{2} \right) e^{-\frac{1}{2} \left(\frac{\log \tau^*}{\hat{\Delta}} + \frac{\Delta}{2} \right)^2 + \frac{\log \tau^*}{2}} \\ & = \exp\left(-\frac{\log^2 \tau^*}{2\Delta^2} - \frac{\Delta^2}{8}\right) \cdot \left(\Delta - \frac{(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} \right) \lesssim \left| \Delta - \frac{(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} \right| \\ & \lesssim \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2^2. \end{aligned}$$

In fact, for the last step, we can obtain this inequality by letting $\boldsymbol{\gamma} = (\Sigma^*)^{1/2} \boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\gamma}} = (\Sigma^*)^{1/2} \hat{\boldsymbol{\beta}}$. Then

$$\left| \Delta - \frac{(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} \right| = \left| \|\boldsymbol{\gamma}\|_2 - \frac{\boldsymbol{\gamma}^\top \hat{\boldsymbol{\gamma}}}{\|\hat{\boldsymbol{\gamma}}\|_2} \right| = \left| \frac{\|\boldsymbol{\gamma}\|_2 \|\hat{\boldsymbol{\gamma}}\|_2 - \boldsymbol{\gamma}^\top \hat{\boldsymbol{\gamma}}}{\|\hat{\boldsymbol{\gamma}}\|_2} \right|.$$

By Lemma 14, eventually we obtain $R^* - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) \lesssim \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2^2$.

To upper bound $R(\hat{G}) - R^*$, applying Taylor's expansion to $R(\hat{G})$,

$$\begin{aligned}
R(\hat{G}) &= (1 - \omega^*) \left\{ \Phi \left(\frac{\log \tau^* - (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right) \right. \\
&\quad \left. + \frac{\log \hat{\tau} - \log \tau^* + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1^*)^\top \hat{\boldsymbol{\beta}} + (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \Phi' \left(\frac{\log \tau^* - (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right) + O_P \left(\frac{s \log p}{n} \right) \right\} \\
&\quad + \omega^* \left\{ \bar{\Phi} \left(\frac{\log \tau^* + (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right) \right. \\
&\quad \left. - \frac{\log \hat{\tau} - \log \tau^* + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2^*)^\top \hat{\boldsymbol{\beta}} - (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \Phi' \left(\frac{\log \tau^* + (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right) + O_P \left(\frac{s \log p}{n} \right) \right\},
\end{aligned}$$

where the remaining term $O_P(\frac{s \log p}{n})$ can be obtained similarly as (4.30).

This leads to

$$\begin{aligned}
&\frac{R^* - R(\hat{G})}{\sqrt{(1 - \omega^*)\omega^*}} \\
&\lesssim \sqrt{\frac{1 - \omega^*}{\omega^*}} \cdot \frac{\log \tau^* - \log \hat{\tau} - (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1^*)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} \Phi' \left(\frac{\log \tau^* - (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right) \\
&\quad - \sqrt{\frac{\omega^*}{1 - \omega^*}} \cdot \frac{\log \tau^* - \log \hat{\tau} + (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2^*)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} \Phi' \left(\frac{\log \tau^* + (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right) \\
&= \frac{\log \tau^* - \log \hat{\tau} - (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1^*)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} e^{-\frac{1}{2} \left\{ \frac{\log \tau^* - (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right\}^2} - \frac{\log \tau^*}{2} \\
&\quad - \frac{\log \tau^* - \log \hat{\tau} + (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2^*)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} e^{-\frac{1}{2} \left\{ \frac{\log \tau^* + (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \right\}^2} + \frac{\log \tau^*}{2}.
\end{aligned}$$

Then it follows that

$$\begin{aligned}
\frac{R^* - R(\hat{G})}{\sqrt{(1 - \omega^*)\omega^*}} &\lesssim \left| \frac{\log \tau^* - \log \hat{\tau} - (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1^*)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} \right| \\
&\quad \cdot \left| e^{-\frac{(\log \tau^* - (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2)^2}{2\hat{\Delta}^2} - \frac{\log \tau^*}{2}} - e^{-\frac{(\log \tau^* + (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2)^2}{2\hat{\Delta}^2} + \frac{\log \tau^*}{2}} \right| \\
&= \underbrace{\left| \frac{\log \tau^* - \log \hat{\tau} - (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1^*)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}} \right|}_{(i)} \cdot \underbrace{e^{-\frac{\log^2 \tau^* + (\boldsymbol{\delta}^{*\top} \hat{\boldsymbol{\beta}}/2)^2}{2\hat{\Delta}^2}}}_{(ii)} \\
&\quad \cdot \underbrace{\left| e^{\frac{\log \tau^* \cdot (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}^2} - \frac{\log \tau^*}{2}} - e^{-\frac{\log \tau^* \cdot (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}^2} + \frac{\log \tau^*}{2}} \right|}_{(iii)} \\
&\lesssim \sqrt{\frac{s \log p}{n}} \cdot O_p(1) \cdot \sqrt{\frac{s \log p}{n}} \lesssim \frac{s \log p}{n},
\end{aligned}$$

where the last inequality uses the following facts

$$(i) \lesssim \sqrt{\frac{s \log p}{n}}, \quad (ii) = O_P(1), \quad \text{and} \quad (iii) \lesssim \sqrt{\frac{s \log p}{n}}.$$

In fact, the bound on (i) follows the same idea of (4.31). (ii) uses the fact that $e^{-x} \leq 1$ when $x \geq 0$. (iii) uses the fact that $|e^x - e^{-x}| \lesssim x$ when $x = o(1)$, and thus can be bounded as

$$\left| e^{\frac{\log \tau^* \cdot (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}^2} - \frac{\log \tau^*}{2}} - e^{-\frac{\log \tau^* \cdot (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}^2} + \frac{\log \tau^*}{2}} \right| \lesssim \left| \frac{(\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}^2} - 1 \right| \lesssim \sqrt{\frac{s \log p}{n}},$$

where the last inequality also follows the same idea as (4.31). Combining the pieces, we obtain

$$R(\hat{G}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) \lesssim \frac{s \log p}{n}. \quad \square$$

4.8.2. Proof of Theorem 8

We focus on mis-classification error. Consider the model $\frac{1}{2}N_p(\boldsymbol{\mu}_1, \Sigma) + \frac{1}{2}N_p(\boldsymbol{\mu}_2, \Sigma)$ with $\boldsymbol{\theta} = (1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) \in \Theta_p(s, c_\omega, M, M_b)$. Let $G_{\boldsymbol{\theta}}$ be the Fisher's rule defined in (4.3) with parameter $\boldsymbol{\theta}$, and the risk function for a generic parameter $\boldsymbol{\theta}$ and classification rule G is

defined as

$$L_{\boldsymbol{\theta}}(G) = \Pr(G \neq G_{\boldsymbol{\theta}}). \quad (4.32)$$

The proof of lower bound requires the generalized version of Fano's lemma.

Lemma 15 (Tsybakov (2009)). *Let $M \geq 0$ and $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_M \in \Theta_p(s, c_\omega, M, M_b)$. For some constants $\alpha \in (0, 1/8), \gamma > 0$, and any classifier \hat{G} , if $\text{KL}(\Pr_{\boldsymbol{\theta}_i}, \Pr_{\boldsymbol{\theta}_0}) \leq \alpha \log M/n$ for all $1 \leq i \leq M$, and $L_{\boldsymbol{\theta}_i}(\hat{G}) < \gamma$ implies $L_{\boldsymbol{\theta}_j}(\hat{G}) \geq \gamma$ for all $0 \leq i \neq j \leq M$, then*

$$\inf_{\hat{G}} \sup_{i \in [M]} \mathbb{E}_{\boldsymbol{\theta}_i} [L_{\boldsymbol{\theta}_i}(\hat{G})] \gtrsim \gamma.$$

Lemma 16 (Tsybakov (2009)). *Let $\mathcal{A}_s = \{\mathbf{u} : \mathbf{u} \in \{0, 1\}^p, \|\mathbf{u}\|_0 \leq s\}$. If $p \geq 4s$, then there exists $\{\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_M\} \subset \mathcal{A}_s$ such that $\mathbf{u}_0 = \{0, \dots, 0\}^\top$, $\rho_H(\mathbf{u}_i, \mathbf{u}_j) \geq s/2$ and $\log(M+1) \geq \frac{s}{5} \log(\frac{p}{s})$, where ρ_H is the Hamming distance.*

Lemma 17. *For any $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta_p(s, c_\omega, M, M_b)$, let $\Pr_{\boldsymbol{\theta}} = (1-\omega)N_p(-\boldsymbol{\mu}/2, \mathbf{I}_p) + \omega N_p(\boldsymbol{\mu}/2, \mathbf{I}_p)$ and $\Pr_{\tilde{\boldsymbol{\theta}}} = (1-\omega)N_p(-\tilde{\boldsymbol{\mu}}/2, \mathbf{I}_p) + \omega N_p(\tilde{\boldsymbol{\mu}}/2, \mathbf{I}_p)$ with $\|\boldsymbol{\mu}\|_2 = \|\tilde{\boldsymbol{\mu}}\|_2$. Then $\text{KL}(\Pr_{\boldsymbol{\theta}}, \Pr_{\tilde{\boldsymbol{\theta}}}) \leq (\|\boldsymbol{\mu}\|_2^2 + \log \tau/2)(\|\boldsymbol{\mu}\|_2^2 - |\boldsymbol{\mu}^\top \tilde{\boldsymbol{\mu}}|)$, where $\tau = \frac{\omega}{1-\omega}$. In particular, if $\omega = 1/2$, we have*

$$\text{KL}_{\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}}(\Pr, \Pr) \leq \|\boldsymbol{\mu}\|_2^4 \cdot \left(1 - \frac{|\boldsymbol{\mu}^\top \tilde{\boldsymbol{\mu}}|}{\|\boldsymbol{\mu}\|_2}\right).$$

Define the function $g(x) = \phi(x)\{\phi(x) - x\Phi(-x)\}$, where $\phi(x)$ is the probability density function of the standard normal distribution, i.e. $\phi(x) = \Phi'(x)$.

Lemma 18 (Azizyan et al. (2013)). *For any $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta_p(s, c_\omega, M, M_b)$ and $\cos \psi = |\boldsymbol{\mu}^\top \tilde{\boldsymbol{\mu}}|/\|\boldsymbol{\mu}\|_2$, we have*

$$2g\left(\frac{\|\boldsymbol{\mu}\|}{2\sigma}\right) \sin \psi \cos \psi \leq L_{\boldsymbol{\theta}}(G_{\tilde{\boldsymbol{\theta}}}).$$

Proof of Theorem 8. First we construct a subset of the parameter space Θ that characterizes the hardness of the problem. Let $\mathbf{e}_1 = \{1, 0, \dots, 0\}^\top \in \mathbb{R}^p$. By Lemma 16, there exist $\mathbf{u}_1, \dots, \mathbf{u}_M \in \tilde{\mathcal{A}}_s = \{\mathbf{u} \in \{0, 1\}^p : \mathbf{u}^\top \mathbf{e}_1 = 0, \|\mathbf{u}\|_0 = s\}$, such that $\rho_H(\mathbf{u}_i, \mathbf{u}_j) > s/2$ and $\log(M+1) \geq \frac{s}{5} \log(\frac{p-1}{s})$. Note the first entry in \mathbf{u}_j is 0 for all $j = 1, \dots, M$.

Define the parameter space

$$\Theta_1 = \{\boldsymbol{\theta} = (1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1 = \epsilon \mathbf{u} + \lambda \mathbf{e}_1, \boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1, \Sigma = \sigma^2 \mathbf{I}_p; \mathbf{u} \in \tilde{\mathcal{A}}_s\}.$$

Here $\epsilon = \sigma \sqrt{\log p/n}$, $\sigma^2 = O(1)$ and $\lambda = O(1)$ are chosen to ensure $\boldsymbol{\theta} \in \Theta_p(s, c_\omega, M, M_b)$ and $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{4\|\epsilon \mathbf{u} + \lambda \mathbf{e}_1\|_2^2}{\sigma^2} \geq c_1$, as required in Lemma 12. To apply Lemma 15, we need to verify two conditions: (i) the upper bound on the KL divergence between $\Pr_{\boldsymbol{\theta}_u}$ and $\Pr_{\boldsymbol{\theta}_v}$, and (ii) the lower bound of $L_{\boldsymbol{\theta}_u}(\hat{G}) + L_{\boldsymbol{\theta}_v}(\hat{G})$ for $\mathbf{u} \neq \mathbf{v}$.

We calculate the KL divergence first. For $\mathbf{u} \in \tilde{\mathcal{A}}_s$, denote $\boldsymbol{\mu}_u = \epsilon \mathbf{u} + \lambda \mathbf{e}_1$. For $\boldsymbol{\theta}_u = (1/2, \boldsymbol{\mu}_u, -\boldsymbol{\mu}_u, \sigma^2 \mathbf{I}_p) \in \Theta_1$, the model parameterized by $\boldsymbol{\mu}_u$ is $\frac{1}{2}N_p(\boldsymbol{\mu}_u, \sigma^2 \mathbf{I}_p) + \frac{1}{2}N_p(-\boldsymbol{\mu}_u, \sigma^2 \mathbf{I}_p)$. For $\mathbf{u}, \mathbf{v} \in \tilde{\mathcal{A}}_s$, since

$$\epsilon^2 \cdot \rho_H(\mathbf{u}, \mathbf{v}) = \langle \boldsymbol{\mu}_u - \boldsymbol{\mu}_v, \boldsymbol{\mu}_u - \boldsymbol{\mu}_v \rangle = \|\boldsymbol{\mu}_u\|_2^2 + \|\boldsymbol{\mu}_v\|_2^2 - 2\boldsymbol{\mu}_u^\top \boldsymbol{\mu}_v = 2\|\boldsymbol{\mu}_u\|_2^2 - 2\boldsymbol{\mu}_u^\top \boldsymbol{\mu}_v,$$

we have $\|\boldsymbol{\mu}_u\|_2^2 - \boldsymbol{\mu}_u^\top \boldsymbol{\mu}_v = \frac{1}{2}\epsilon^2 \cdot \rho_H(\mathbf{u}, \mathbf{v}) \asymp \frac{s \log p}{n}$. Lemma 17 then yields

$$\text{KL}(\Pr_{\boldsymbol{\theta}_u}, \Pr_{\boldsymbol{\theta}_v}) \leq \|\boldsymbol{\mu}_u\|_2^2 (\|\boldsymbol{\mu}_u\|_2^2 - \boldsymbol{\mu}_u^\top \boldsymbol{\mu}_v) \lesssim \frac{s \log p}{n}. \quad (4.33)$$

Consider $L_{\boldsymbol{\theta}}(G)$ defined in (4.32). Recall that in Lemma 18, $\cos \psi = \boldsymbol{\mu}_u^\top \boldsymbol{\mu}_v / \|\boldsymbol{\mu}_u\|_2^2$. For the choice of ϵ and $\boldsymbol{\mu}_u$, we have $\frac{\|\boldsymbol{\mu}_u\|_2}{2\sigma} = O(1)$, which implies that $2g(\frac{\|\boldsymbol{\mu}_u\|_2}{2\sigma}) = O(1)$ under the condition $s = o(n/\log p)$. Also,

$$1 - \cos \psi = 1 - \frac{\boldsymbol{\mu}_u^\top \boldsymbol{\mu}_v}{\|\boldsymbol{\mu}_u\|_2^2} = \frac{\|\boldsymbol{\mu}_u\|_2^2 - \boldsymbol{\mu}_u^\top \boldsymbol{\mu}_v}{\|\boldsymbol{\mu}_u\|_2^2} = \frac{\rho_H(\mathbf{u}, \mathbf{v})\epsilon^2}{2(\lambda^2 + s\epsilon^2)} \asymp \frac{s \log p}{n}.$$

Therefore, by Lemma 18,

$$L_{\boldsymbol{\theta}_u}(G_{\boldsymbol{\theta}_v}) \geq 2g\left(\frac{\|\boldsymbol{\mu}\|_2}{2\sigma}\right) \sin \psi \cos \psi \geq g\left(\frac{\|\boldsymbol{\mu}\|_2}{2\sigma}\right) \sqrt{1 + \cos \psi} \sqrt{1 - \cos \psi} \geq \sqrt{\frac{s \log p}{n}}.$$

Applying Lemma 13 with a proper choice of ϵ , we have, for any $\mathbf{u}, \mathbf{v} \in \tilde{\mathcal{A}}_s$,

$$L_{\boldsymbol{\theta}_u}(\hat{G}) + L_{\boldsymbol{\theta}_v}(\hat{G}) \geq L_{\boldsymbol{\theta}_u}(G_{\boldsymbol{\theta}_v}) - \sqrt{\frac{\text{KL}(\text{Pr}_{\boldsymbol{\theta}_u}, \text{Pr}_{\boldsymbol{\theta}_v})}{2}} \gtrsim \sqrt{\frac{s \log p}{n}}.$$

So far we have verified the aforementioned conditions (i) and (ii). Lemma 15 immediately implies that

$$\inf_{\hat{G} \in \mathcal{C}} \sup_{\boldsymbol{\theta} \in \Theta_p(s, c_\omega, M, M_b)} L_{\boldsymbol{\theta}}(\hat{G}) \gtrsim \sqrt{\frac{s \log p}{n}}. \quad (4.34)$$

Finally combining (4.34) with Lemma 12, we obtain the desired lower bound for the mis-clustering error. \square

CHAPTER 5 : The Cost of Privacy: Optimal Rates of Convergence for Parameter Estimation with Differential Privacy

5.1. Introduction

With the unprecedented availability of datasets containing sensitive personal information, there are increasing concerns that statistical analysis of such datasets may compromise individual privacy. These concerns give rise to statistical methods that provide privacy guarantees at the cost of statistical accuracy, but there has been very limited understanding of the optimal tradeoff between statistical accuracy and privacy cost.

A rigorous definition of privacy is a prerequisite for such an understanding. Differential privacy, introduced in Dwork et al. (2006), is arguably the most widely adopted definition of privacy in statistical data analysis. The promise of a differentially private algorithm is protection of any individual's privacy from an adversary who has access to the algorithm output and even sometimes the rest of the data. Differential privacy has gained significant attention in the machine learning communities over the past few years (Dwork et al., 2014a; Abadi et al., 2016; Dwork et al., 2017; Dwork and Feldman, 2018) and found its way into real world applications developed by Google (Erlingsson et al., 2014), Apple (Differential Privacy Team, 2017), Microsoft (Ding et al., 2017), and the U.S. Census Bureau (Abowd, 2016).

A usual approach to developing differentially private algorithms is perturbing the output of non-private algorithms by random noise. When the observations are continuous, differential privacy can be guaranteed by adding Laplace/Gaussian noise to the non-private output (Dwork et al., 2014a). For discrete data, differential privacy can be achieved by adding Gumbel noise to utility score functions (also known as the exponential mechanism). Naturally, the processed output suffers from some loss of accuracy, which has been observed and studied in the literature, see, for example, Wasserman and Zhou (2010); Smith (2011); Lei (2011); Bassily et al. (2014); Dwork et al. (2014b). However, given a certain privacy

constraint, it is still unclear what the best achievable statistical accuracy is, or in other words, what the optimal tradeoff between privacy cost and statistical accuracy is.

The goal of this paper is to provide a quantitative characterization of the tradeoff between privacy cost and statistical accuracy, under the statistical minimax framework. Specifically, we consider this problem for mean estimation and linear regression models in both classical and high-dimensional settings with (ε, δ) -differential privacy constraint, which is formally defined as follows.

Definition 1 (Differential Privacy (Dwork et al., 2006)). *A randomized algorithm M is (ε, δ) -differentially private if and only if for every pair of adjacent datasets $X_{1:n}$ and $X'_{1:n}$, and for any set S ,*

$$\mathbb{P}(M(X_{1:n}) \in S) \leq e^\varepsilon \cdot \mathbb{P}(M(X'_{1:n}) \in S) + \delta,$$

where we say two datasets $X_{1:n} = \{\mathbf{x}_i\}_{i=1}^n$ and $X'_{1:n} = \{\mathbf{x}'_i\}_{i=1}^n$ are adjacent if and only if $\sum_{i=1}^n \mathbb{1}(\mathbf{x}_i \neq \mathbf{x}'_i) = 1$.

According to the definition, the two parameters ε and δ control the level of privacy against an adversary who attempts to detect the presence of a certain subject in the sample. Roughly speaking, ε is an upper bound on the amount of influence an individual's record has on the information released and δ is the probability that this bound fails to hold, so the privacy constraint becomes more stringent as ε, δ tend to 0.

We establish the necessary cost of privacy by first providing minimax lower bounds for the estimation accuracy under this (ε, δ) -differential privacy constraint. The results show that the estimators with privacy guarantees generally exhibit very different rates of convergence compared to their non-private counterparts. As a first example, we consider the d -dimensional mean estimation under the ℓ_2 loss: Theorem 12 in Section 5.2 shows that, when the sample size is n , for any (ε, δ) -differentially private algorithm, in addition to the standard $\sqrt{d/n}$ statistical error, there must be an extra error of at least the order of $d\sqrt{\log(1/\delta)}/n\varepsilon$. This lower bound is established by using a general technique presented in

Theorem 11, which reduces the establishing minimax risk lower bounds to designing and analyzing a tracing adversary that aims to detect the presence of an individual in a dataset via the output of a differentially private procedure that is applied to the dataset. The design and analysis of tracing adversary makes use of a novel generalization of the fingerprinting lemma, a concept from cryptography (Boneh and Shaw, 1998). The connections between tracing adversaries, the fingerprinting lemma and differential privacy have been observed in Tardos (2008), Bun et al. (2014) and Dwork et al. (2015), but their discussions are primarily concerned with discrete distributions. In this paper, we provide a continuous version of the fingerprinting lemma that enables us to establish minimax lower bounds for a greater variety of statistical problems; more discussions are given in Section 5.2 as well as the Supplementary Material (Cai et al., 2019c).

Further, we argue that these necessary costs of privacy, as shown by lower bounds for the minimax rates, are in fact sharp in both mean estimation and linear regression problems. We construct efficient algorithms and establish matching upper bounds up to logarithmic factors. These algorithms are based on several differentially private subroutines, such as the Gaussian mechanism, reporting noisy top- k , and their modifications. In particular, for the high-dimensional linear regression, we propose a novel private iterative hard thresholding pursuit algorithm, based on a privately truncated version of stochastic gradient descent. Such a private truncation step effectively enforces the sparsity of the resulting estimator and leads to optimal control of the privacy cost (see more details in Section 5.4.2). To the best of our knowledge, these algorithms are the first results achieving the minimax optimal rates of convergence in high-dimensional statistical estimation problems with the (ϵ, δ) -differential privacy guarantee. Our Theorems 13, 14, 16, and 18 together provide matching upper and lower bounds for both mean estimation and linear regression problems in high-dimensional and classical settings, up to logarithmic factors.

Related literature

There are previous works studying how the privacy constraints compromise estimation accuracy. In theoretical computer science, Smith (2011) showed that under strong conditions on privacy parameters, some point estimators attain the statistical convergence rates and hence privacy can be gained for free. Bassily et al. (2014); Dwork et al. (2014b); Talwar et al. (2015) proposed differentially private algorithms for convex empirical risk minimization, principal component analysis, and high-dimensional regression, and investigated the convergence rates of excess risk. In addition, Bun et al. (2014); Ullman (2016); Bafna and Ullman (2017) considered the optimal estimation of sample quantities such as k -way marginals and top- k selection with privacy constrain. Unlike most prior works that focused on excess risks or the release of sample quantities, our focus is the population parameter estimation. Theoretical properties of excess risks or sample quantities can be very different from those of population parameters; see more discussions in Duchi et al. (2013).

More recent works aimed to study differential privacy in the context of statistical estimation. Wasserman and Zhou (2010) observed that, (ϵ, δ) -local differentially private schemes seem to yield slower convergence rates than the optimal minimax rates in general; Duchi et al. (2018) developed a framework for statistical minimax rates with the α -local privacy constraint; in addition, Rohde and Steinberger (2018) showed minimax optimal rates of convergence under α -local differential privacy and exhibited a mechanism that is minimax optimal for nearly linear functionals based on randomized response. However, α -local privacy is a much stronger notion of privacy than (ϵ, δ) -differential privacy that is hardly compatible with high-dimensional problems (Duchi et al., 2018). As we shall see in this paper, the cost of (ϵ, δ) -differential privacy in statistical estimation behaves quite differently compared to that of α -local privacy.

Organization of the paper

The rest of the paper is organized as follows. Section 5.2 introduces a general technical tool for deriving lower bounds of the minimax risk with differential privacy constraint. The new technical tool is then applied in Section 3 to the high-dimensional mean estimation problem. Both minimax lower bound results and algorithms with matching upper bounds are obtained. Section 4 further applies the general lower bound technique to investigate the minimax lower bounds of the linear regression problem with differential privacy constraint, in both low-dimensional and high-dimensional settings. The upper bounds are also obtained by providing novel differentially private algorithms and analyzing their risks. The results together show that our bounds are rate-optimal up to logarithmic factors. Simulation studies are carried out in Section 5 to show the advantages of our proposed algorithms. Section 6 applies our algorithms to real data sets with potentially sensitive information that warrants privacy-preserving methods. Section 7 discusses extensions to other statistical estimation problems with privacy constraints. The proofs are given in Section 8.

Definitions and notation

We conclude this section by introducing notations that will be used in the rest of the paper. For a positive integer n , $[n]$ denotes the set $\{1, 2, \dots, n\}$. For a vector $\mathbf{x} \in \mathbb{R}^d$, we use $\|\mathbf{x}\|_0, \|\mathbf{x}\|_p = (\sum_{j \in [d]} x_j^p)^{1/p}$ and $\|\mathbf{x}\|_\infty = \max_{j \in [d]} |x_j|$ to denote the usual vector ℓ_0, ℓ_p and ℓ_∞ norm, respectively, where the ℓ_0 norm counts the number of nonzero entries in a vector. For any set $A \subseteq [d]$ and $\mathbf{v} \in \mathbb{R}^d$, let \mathbf{v}_A denote the $|A|$ -dimensional vector consisting of v_i such that $i \in A$. The Frobenius norm of a matrix $\Omega = (\omega_{ij})$ is denoted by $\|\Omega\|_F = \sqrt{\sum_{i,j} \omega_{ij}^2}$, and the spectral norm of Ω is $\|\Omega\|_2$. In addition, we use $\lambda_{\min}, \lambda_{\max}$ to denote the smallest and the largest eigenvalues of Ω . The matrix ℓ_0 norm is defined similarly as the vector ℓ_0 norm, i.e. $\|\Omega\|_0 = \#\{(i, j) : \omega_{ij} \neq 0\}$. In addition, $|\Omega|$ denotes the determinant of Ω . The empirical measure is denoted by $\mathbb{E}_n = n^{-1} \sum_{i=1}^n \delta_{\mathbf{x}_i}$ for a sample $\mathbf{x}_1, \dots, \mathbf{x}_n$. For a set A , we use A^c to denote its complement, and $\mathbb{1}(A)$ denotes the indicator function on A . We use C, C_1, C_2, \dots , and c_1, c_2, \dots to denote generic constants which may vary line by line.

5.2. A General Lower Bound for Minimax Risk with Differential Privacy

This section presents a general minimax lower bound technique for statistical estimation problems with differential privacy constraint. As an application, we use this technique to establish a tight lower bound for differentially private mean estimation in this section.

Our lower bound technique is based on a tracing adversary that attempts to detect the presence of an individual data entry in a data set with the knowledge of an estimator computed from the data set. If one can construct a tracing adversary that is effective at this task given an accurate estimator, an argument by contradiction leads to a lower bound of the accuracy of differentially private estimators: suppose a differentially private estimator from a data set is sufficiently accurate, the tracing adversary will be able to determine the presence of an individual data entry in the data set, thus contradicting with the differential privacy guarantee. In other words, the privacy guarantee and the tracing adversary together ensure that a differentially private estimator cannot be “too accurate”.

5.2.1. Background and problem formulation

Let \mathcal{P} denote a family of distributions supported on a set \mathcal{X} , and let $\boldsymbol{\theta} : \mathcal{P} \rightarrow \Theta \subset \mathbb{R}^d$ denote a population quantity of interest. The statistician has access to a data set of n i.i.d. samples, $X = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathcal{X}^n$, drawn from a statistical model $P \in \mathcal{P}$.

With the data, our goal is to estimate a population parameter $\boldsymbol{\theta}(P)$ by an estimator $M(X) : \mathcal{X}^n \rightarrow \Theta$ that belongs to $\mathcal{M}_{\varepsilon, \delta}$, the collection of all (ε, δ) -differentially private procedures. The performance of $M(X)$ is measured by its distance to the truth $\boldsymbol{\theta}(P)$: formally, let $\rho : \Theta \times \Theta \rightarrow \mathbb{R}^+$ be a metric induced by a norm $\|\cdot\|$ on Θ , namely $\rho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$, and let $l : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be a loss function that is monotonically increasing on \mathbb{R}^+ , this paper studies the minimax risk for differentially-private estimation of the population parameter $\boldsymbol{\theta}(P)$:

$$\inf_{M \in \mathcal{M}_{\varepsilon, \delta}} \sup_{P \in \mathcal{P}} \mathbb{E} [l(\rho(M(X), \boldsymbol{\theta}(P)))].$$

In this paper, our setting of the privacy parameters are $\varepsilon = O(1)$ and $\delta = o(1/n)$. This is essentially the most-permissive setting under which (ε, δ) -differential privacy is a nontrivial guarantee: Steinke and Ullman (2017) shows that $\delta < 1/n$ is essentially the weakest privacy guarantee that is still meaningful.

5.2.2. Lower bound by tracing

Consider a tracing adversary $\mathbf{a}_P(\mathbf{x}, M(X)) : \mathcal{X} \times \Theta \rightarrow \{\text{IN}, \text{OUT}\}$ that outputs IN if it determines a certain sample \mathbf{x} is in the data set X after seeing $M(X)$, and outputs OUT otherwise. We define $\mathcal{TR}(X, M(X)) := \{i \in [n] : \mathbf{a}_P(\mathbf{x}_i, M(X)) = \text{IN}\}$, the index set of samples that are determined as IN by the adversary \mathbf{a}_P . A survey of tracing adversaries and their relationship with differential privacy can be found in Dwork et al. (2017) and the reference therein.

Our general lower bound technique requires some regularity conditions for \mathcal{P} and $\theta : \mathcal{P} \rightarrow \Theta \subset \mathbb{R}^d$: for every $P \in \mathcal{P}$, we assume that there exists a $P_0 \in \mathcal{P}$ such that for every $\alpha \in [0, 1]$, $(1 - \alpha)P_0 + \alpha P \in \mathcal{P}$, and $\theta((1 - \alpha)P_0 + \alpha P) = \alpha\theta(P)$. The two statistical problems investigated in this paper, mean estimation and linear regression, satisfy the property.

The following theorem shows that minimax lower bounds for statistical estimation problems with privacy constraint can be constructed if there exist effective tracing adversaries:

Theorem 11. *Suppose $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is an i.i.d. sample from a distribution $P \in \mathcal{P}$, and assume that \mathcal{P} and θ satisfy the regularity conditions described above. Given a tracing adversary $\mathbf{a}_P(\mathbf{x}, M(X))$ that satisfies the following two properties when $n \lesssim \psi(\mathcal{P}, \delta)$,*

1. *completeness: $\mathbb{P}(\{\mathcal{TR}(X, M(X)) = \emptyset\} \cap \{\rho(M(X), \theta(P)) \lesssim \lambda(\mathcal{P}, \delta)\}) \leq \delta$,*
2. *soundness: $\mathbb{P}(\mathbf{a}_P(\mathbf{x}_i, M(X'_i)) = \text{IN}) \leq \delta$, where X'_i is an adjacent dataset of X with \mathbf{x}_i replaced by $\mathbf{x}'_i \sim P$,*

then if $\varepsilon = O(1)$, $n^{-1}e^{-3\varepsilon n/2} \leq \delta \leq n^{-(1+\tau)}$ for some $\tau > 0$, and $n \gtrsim \psi(\mathcal{P}, \delta) \log(1/\delta)/\varepsilon$,

we have

$$\inf_{M \in \mathcal{M}_{\varepsilon, \delta}} \sup_{P \in \mathcal{P}} \mathbb{E} [l(\rho(M(X), \boldsymbol{\theta}(P)))] \gtrsim l \left(\frac{\psi(\mathcal{P}, \delta) \cdot \lambda(\mathcal{P}, \delta) \cdot \log(1/\delta)}{n\varepsilon} \right).$$

Completeness and soundness roughly correspond to “true positive” and “false positive” in classification: completeness requires the adversary to return some nontrivial result when its input $M(X)$ is accurate; soundness guarantees that an individual is unlikely to be identified as IN if the estimator that \mathbf{a}_P used is independent of the individual. When a tracing adversary satisfies these properties, Theorem 11 conveniently leads to a minimax risk lower bound; that is, Theorem 11 is a reduction from constructing minimax risk lower bounds to finding complete and sound tracing adversaries.

In the next section, we illustrate this technique by designing a complete and sound tracing adversary for the classical mean estimation problem.

5.2.3. A first application: private mean estimation in the classical setting

Consider the d -dimensional sub-Gaussian distribution family $\mathcal{P}(\sigma, d)$, defined as

$$\mathcal{P}(\sigma, d) = \left\{ P \mid \mathbb{E}_{\mathbf{x} \sim P} \left[e^{\lambda \mathbf{e}_k^\top (\mathbf{x} - \boldsymbol{\mu}_P)} \right] \leq e^{\sigma^2 \lambda^2 / 2}, \forall \lambda \in \mathbb{R}, k \in [d] \right\},$$

where $\boldsymbol{\mu}_P = \mathbb{E}_P[\mathbf{x}] \in \mathbb{R}^d$ is the mean of P , and \mathbf{e}_k denotes the k th standard basis vector of \mathbb{R}^d .

Following the notation introduced in Section 5.2.1, $\mathcal{X} = \mathbb{R}^d$ and $\boldsymbol{\theta}(P) = \boldsymbol{\mu}_P$. Further we take $l(t) = t$ and $\rho(\boldsymbol{\theta}, \boldsymbol{\theta}') = \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$, so that our risk function is simply the ℓ_2 error. The minimax risk is then denoted by

$$\inf_{M(X) \in \mathcal{M}_{\varepsilon, \delta}} \sup_{P \in \mathcal{P}(\sigma, d)} \mathbb{E} [\|M(X) - \boldsymbol{\mu}_P\|_2].$$

We propose a tracing adversary:

$$\mathbf{a}_P(\mathbf{x}, M(X)) = \begin{cases} \text{IN} & \langle \mathbf{x} - \tilde{\mathbf{x}}, M(X) \rangle > \sigma^2 \sqrt{8d \log(1/\delta)}, \\ \text{OUT} & \text{otherwise,} \end{cases}$$

where $\tilde{\mathbf{x}}$ is a fresh independent draw from P . The adversary is indeed complete and sound, as desired:

Lemma 12. *If $n \lesssim \sqrt{d/\log(1/\delta)}$, there is a distribution $P \in \mathcal{P}(\sigma, d)$, such that*

1. $\mathbb{P}(\{\mathcal{TR}(X, M(X)) = \emptyset\} \cap \{\|M(X) - \boldsymbol{\mu}_P\|_2 \lesssim \sigma\sqrt{d}\}) \leq \delta$,
2. $\mathbb{P}(A(\mathbf{x}_i, M(X'_i)) = \text{IN}) \leq \delta$, where X'_i is an adjacent dataset of X with \mathbf{x}_i replaced by $\mathbf{x}'_i \sim P$.

Intuitively, this adversary is constructed as follows. Without privacy constraints, a natural estimator for $\boldsymbol{\mu}_P$ is the sample mean $M(X) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$. On one hand, when \mathbf{x} does not belong to $X = \{\mathbf{x}_i\}_{i=1}^n$, $\langle \mathbf{x} - \tilde{\mathbf{x}}, M(X) \rangle$ is a sum of d independent zero-mean random variables and we have $\mathbb{E}[\langle \mathbf{x} - \tilde{\mathbf{x}}, M(X) \rangle] = 0$. On the other hand, when \mathbf{x} belongs to X , we will have $\mathbb{E}[\langle \mathbf{x} - \tilde{\mathbf{x}}, M(X) \rangle] = \mathbb{E}[\frac{1}{n} \|\mathbf{x}\|_2^2] > 0$, and $\mathbf{a}_P(\mathbf{x}, M(X))$ is more likely to output IN than OUT.

In view of Theorem 11, $\lambda(\mathcal{P}, \delta) = \sigma\sqrt{d}$ and $\psi(\mathcal{P}, \delta) = \sqrt{d/\log(1/\delta)}$; it follows that

$$\inf_{M \in \mathcal{M}_{\varepsilon, \delta}} \sup_{P \in \mathcal{P}(\sigma, d)} \mathbb{E}_P [\|M(X) - \boldsymbol{\mu}_P\|_2] = \sigma \frac{d\sqrt{\log(1/\delta)}}{n\varepsilon}.$$

Combining with the well-known statistical minimax lower bound, see for example, Lehmann and Casella (2006), namely

$$\inf_M \sup_{P \in \mathcal{P}(\sigma, d)} \mathbb{E} [\|M(X) - \boldsymbol{\mu}_P\|_2] \gtrsim \sigma \sqrt{\frac{d}{n}},$$

we arrive at the minimax lower bound result for differentially private mean estimation.

Theorem 12. Let $\mathcal{M}_{\varepsilon, \delta}$ denote the collection of all (ε, δ) -differentially private algorithms, and let $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ be an i.i.d. sample drawn from $P \in \mathcal{P}(\sigma, d)$. Suppose that $\varepsilon = O(1)$, $n^{-1}e^{-3\varepsilon n/2} \leq \delta \leq n^{-(1+\tau)}$ for some $\tau > 0$ and $\frac{\sqrt{d \log(1/\delta)}}{n\varepsilon} = O(1)$, then

$$\inf_{M \in \mathcal{M}_{\varepsilon, \delta}} \sup_{P \in \mathcal{P}(\sigma, d)} \mathbb{E} [\|M(X) - \boldsymbol{\mu}_P\|_2] \gtrsim \sigma \left(\sqrt{\frac{d}{n}} + \frac{d\sqrt{\log(1/\delta)}}{n\varepsilon} \right).$$

Remark 11. In comparison, applying Barber and Duchi (2014)'s lower bound argument to our current model yields

$$\inf_{M \in \mathcal{M}_{\varepsilon, \delta}} \sup_{P \in \mathcal{P}(\sigma, d)} \mathbb{E} [\|M(X) - \boldsymbol{\mu}_P\|_2] \gtrsim \sigma \left(\sqrt{\frac{d}{n}} + \frac{\sqrt{d} \cdot (d \wedge \log(1/\delta))}{n\varepsilon} \right).$$

Remark 12. The minimax lower bound characterizes the cost of privacy in the mean estimation problem: the cost of privacy dominates the statistical risk when $\sqrt{d \log(1/\delta)}/\sqrt{n\varepsilon} \gtrsim 1$.

5.3. Privacy Cost of High-dimensional Mean Estimation

In this section and the subsequent Section 5.4, we consider the high-dimensional setting where $d \gtrsim n$ and the population parameters of interest, such as the mean vector $\boldsymbol{\mu}_P$ or the regression coefficient $\boldsymbol{\beta}$, are sparse. In each statistical problem investigated, we present a minimax risk lower bound with differential privacy constraint, as well as a procedure with differential privacy guarantee that attains the lower bound up to factor(s) of $\log n$.

5.3.1. Private high-dimensional mean estimation

We first consider the problem of estimating the sparse mean vector $\boldsymbol{\mu}_P$ of a d -dimensional sub-Gaussian distribution, where d can possibly be much larger than the sample size n . We denote the parameter space of interest by

$$\mathcal{P}(\sigma, d, s) = \left\{ P \mid \mathbb{E}_P \left[e^{\lambda \mathbf{e}_k^\top (\mathbf{x} - \boldsymbol{\mu}_P)} \right] \leq e^{\sigma^2 \lambda^2 / 2}, \forall \lambda \in \mathbb{R}, k \in [d]; \|\boldsymbol{\mu}_P\|_0 \leq s \right\},$$

where the sparsity level is controlled by the parameter s .

The tracing adversary for this problem is given by

$$\mathbf{a}_P(\mathbf{x}, M(X)) = \begin{cases} \text{IN} & \sum_{j:j \in S(M(X))} (x_j - \tilde{x}_j) > \sigma \sqrt{8s \log(1/\delta)}, \\ \text{OUT} & \text{otherwise,} \end{cases}$$

where $\tilde{\mathbf{x}}$ is an independent draw from P , and

$$S(M(X)) := \{j \in [d] : M(X)_j \text{ is among the top } s \text{ largest coordinates of } M(X)\}.$$

Given $M(X)$ computed from a data set X , the tracing adversary attempts to identify whether an individual \mathbf{x} belongs to X , by calculating the difference of $\sum_j x_j$ and $\sum_j \tilde{x}_j$ over those coordinates j where $M(X)$ has a large value. If \mathbf{x} belongs to X , the former should be correlated with $M(X)$ and is likely to be larger than the latter.

Formally, the tracing adversary is complete and sound under appropriate sample size constraint:

Lemma 13. *If $n \lesssim \sqrt{s/\log(1/\delta)} \log(d/s)$, there is a distribution $P \in \mathcal{P}(\sigma, d, s)$ such that*

1. $\mathbb{P}(\{\mathcal{TR}(X, M(X)) = \emptyset\} \cap \{\|M(X) - \boldsymbol{\mu}_P\|_2 \lesssim \sigma\sqrt{s}\}) \leq \delta$,
2. $\mathbb{P}(\mathbf{a}_P(\mathbf{x}_i, M(X'_i)) = \text{IN}) \leq \delta$, where X'_i is an adjacent data set of X with \mathbf{x}_i replaced by $\mathbf{x}'_i \sim P$.

In conjunction with our general lower bound result Theorem 11, we have

Theorem 13. *Let $\mathcal{M}_{\varepsilon, \delta}$ denote the collection of all (ε, δ) -differentially private algorithms, and let $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ be an i.i.d. sample drawn from $P \in \mathcal{P}(\sigma, d, s)$. Suppose that $\varepsilon = O(1)$, $n^{-1}e^{-3\varepsilon n/2} \leq \delta \leq n^{-(1+\tau)}$ for some $\tau > 0$, and $\frac{\sqrt{s \log(1/\delta)} \log(d/s)}{n\varepsilon} = O(1)$, then*

$$\inf_{M \in \mathcal{M}_{\varepsilon, \delta}} \sup_{P \in \mathcal{P}(\sigma, d, s)} \mathbb{E} [\|M(X) - \boldsymbol{\mu}_P\|_2] \gtrsim \sigma \left(\sqrt{\frac{s \log d}{n}} + \frac{s \log(d/s) \sqrt{\log(1/\delta)}}{n\varepsilon} \right).$$

The first term is the statistical minimax lower bound of sparse mean estimation (see, for example, Johnstone (1994)), and the second term is due to the privacy constraint. Comparing the two terms shows that, in high-dimensional sparse mean estimation, the cost of differential privacy is significant when

$$\frac{\sqrt{s \log(1/\delta)/\log d} \log(d/s)}{\sqrt{n}\epsilon} \gtrsim 1.$$

In the next section, we present a differentially private procedure that attains this convergence rate up to a logarithmic factor.

5.3.2. Rate-optimal procedures

The rate-optimal algorithms in this paper utilize some classical subroutines in the differential privacy literature, such as the Laplace and Gaussian mechanisms and reporting the noisy maximum of a vector. Before describing our rate-optimal algorithms in detail, it is helpful to review some relevant results, which will also serve as the building blocks of the differentially private linear regression methods in Section 4.

Basic differentially private procedures

It is frequently the case that differential privacy can be attained by adding properly scaled noises to the output of a non-private algorithm. Among the most prominent examples are the Laplace and Gaussian mechanisms.

The Laplace and Gaussian mechanisms

As the name suggests, the Laplace and Gaussian mechanisms achieve differential privacy by perturbing an algorithm with Laplace and Gaussian noises respectively. The scale of such noises is determined by the sensitivity of the algorithm:

Definition 2. For any algorithm f mapping a dataset X to \mathbb{R}^d , The L^p -sensitivity of f is

$$\Delta_p(f) = \sup_{X, X' \text{ adjacent}} \|f(X) - f(X')\|_p.$$

For algorithms with finite L^1 -sensitivity, the differential privacy guarantee can be attained by adding noises sampled from a Laplace distribution.

Lemma 14 (The Laplace mechanism (Dwork et al., 2014a)). For any algorithm f mapping a dataset to \mathbb{R}^d such that $\Delta_1(f) < \infty$, the Laplace mechanism, given by

$$M_1(X, f, \varepsilon) := f(X) + (\xi_1, \xi_2, \dots, \xi_d)$$

where $\xi_1, \xi_2, \dots, \xi_d$ is an i.i.d. sample drawn from $\text{Laplace}(\Delta_1 f / \varepsilon)$, achieves $(\varepsilon, 0)$ -differential privacy.

Similarly, adding Gaussian noises to algorithms with finite L^2 -sensitivity guarantees differential privacy.

Lemma 15 (The Gaussian mechanism (Dwork et al., 2014a)). For any algorithm f mapping a dataset to \mathbb{R}^d such that $\Delta_2(f) < \infty$, the Gaussian mechanism, given by

$$M_2(X, f, \varepsilon) := f(X) + (\xi_1, \xi_2, \dots, \xi_d)$$

where $\xi_1, \xi_2, \dots, \xi_k$ is an i.i.d. sample drawn from $N\{0, 2(\Delta_2 f / \varepsilon)^2 \log(1.25/\delta)\}$, achieves (ε, δ) -differential privacy.

An important application of these mechanisms is differentially private selection of the maximum/minimum, which also plays a crucial role in our high-dimensional mean estimation algorithm. Next we review some algorithms for differentially private selection, to provide some concrete examples and prepare us for stating the main algorithms.

Differentially private selection

Selecting the maximum (in absolute value) coordinate of $\mathbf{f}(\mathbf{x}) := \mathbf{f}(x_1, x_2, \dots, x_n) \in \mathbb{R}^d$ is a straightforward application of the Laplace mechanism, as follows:

Algorithm 1: PrivateMax($\mathbf{f}(\mathbf{x}), B, \varepsilon$):

- 1: Sample $\xi_1, \dots, \xi_d \stackrel{\text{i.i.d.}}{\sim} \text{Laplace}(2B/\varepsilon)$.
- 2: For $i \in [d]$, compute the noisy version $|f_i(\mathbf{x})| + \xi_i$.
- 3: Return $i_{\max} = \arg \max_j |f_j(\mathbf{x}) + \xi_j|$ and $f_{i_{\max}}(\mathbf{x}) + w$, where w is an independent draw from $\text{Laplace}(2B/\varepsilon)$.

Lemma 16 ((Dwork et al., 2018)). *If $\sup_{\mathbf{x}, \mathbf{x}' \text{ adjacent}} \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}')\|_{\infty} \leq B$, then PrivateMax($\mathbf{f}(\mathbf{x}), B, \varepsilon$) is $(\varepsilon, 0)$ -differentially private.*

In applications, we are often interested in finding the top- k numbers with $k > 1$. There are two methods for this task: an iterative ‘‘Peeling’’ algorithm that runs the PrivateMax algorithm k times, with appropriately chosen privacy parameters in each iteration.

Algorithm 2: Peeling($\mathbf{f}(\mathbf{x}), k, B, \varepsilon, \delta$):

- 1: Set $\mathbf{z} = \mathbf{f}(\mathbf{x})$.
- 2: **for** $j = 1$ to k **do**
- 3: Run PrivateMax $\left(\mathbf{z}, B, \frac{\varepsilon}{2\sqrt{3k \log(1/\delta)}}\right)$ to obtain $(i_j, f_{i_j}(\mathbf{x}) + w'_j)$.
- 4: Remove $f_{i_j}(\mathbf{x})$ from \mathbf{z} .
- 5: **end for**
- 6: Report the k selected pairs.

Lemma 17 ((Dwork et al., 2018)). *If $\sup_{\mathbf{x}, \mathbf{x}' \text{ adjacent}} \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}')\|_{\infty} \leq B$, then Peeling($\mathbf{x}, k, B, \varepsilon, \delta$) is (ε, δ) -differentially private.*

With these differentially private selection subroutine, we are ready to present the high-dimensional mean estimation algorithm in the next section.

Differentially-private mean estimation in high dimensions

Let $f_T(\cdot)$ denote projection onto the ℓ_∞ ball of radius $T > 0$ in \mathbb{R}^d , where T is a tuning parameter for the truncation level. With suitably chosen T , the following algorithm attains the minimax lower bound in Theorem 13, up to at most a logarithmic factor in n .

Algorithm 3: Private High-dimensional Mean Estimation

- 1: Compute $\hat{\boldsymbol{\mu}}_T = \frac{1}{n} \sum_{i=1}^n f_T(X_i)$
- 2: Find the top \hat{s} components of $\hat{\boldsymbol{\mu}}_T$ by running Peeling $(\boldsymbol{\mu}_T, \hat{s}, 2T/n, \varepsilon, \delta)$ and set the remaining components to 0. Denote the resulting vector by $\hat{\boldsymbol{\mu}}_{T, \hat{s}}$.
- 3: Return $\hat{\boldsymbol{\mu}}_{T, \hat{s}}$.

In view of Theorem 13, the theorem below shows that the high-dimensional mean estimation algorithm is rate-optimal up to a factor of $\sqrt{\log n}$.

Theorem 14. *For $X_1, X_2, \dots, X_n \sim P \in \mathcal{P}(\sigma, d, s)$ with $\mathbb{E}_P X = \boldsymbol{\mu}$, if $\|\boldsymbol{\mu}\|_\infty = O(1)$ $\hat{s} \asymp s$ and $\hat{s} > s$, then Algorithm 3 is (ε, δ) -differentially private, and*

1. *if there exists a constant $M < \infty$ such that $\mathbb{P}(\|X\|_\infty < M) = 1$, when $T \geq M$,*

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}_{T, \hat{s}} - \boldsymbol{\mu}\|_2 \lesssim \sigma \left(\sqrt{\frac{s \log d}{n}} + \frac{s \log d \sqrt{\log(1/\delta)}}{n\varepsilon} \right);$$

2. *otherwise, with the choice of $T \geq C\sigma\sqrt{\log n}$ for a sufficiently large constant $C > 0$,*

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}_{T, \hat{s}} - \boldsymbol{\mu}\|_2 \lesssim \sigma \left(\sqrt{\frac{s \log d}{n}} + \frac{s \log d \sqrt{\log(1/\delta) \log n}}{n\varepsilon} \right).$$

Remark 13. *Duchi et al. (2018) introduced the notion of α -local privacy and shows that high-dimensional estimation is effectively impossible with α -local privacy constraint. In contrast, Theorem 14 shows that sparse mean estimation is still possible with (ε, δ) -differential privacy constraint.*

Remark 14. *The role of the truncation parameter T is to control the sensitivity of the sample mean so that the Laplace/Gaussian mechanisms are applicable. T can be replaced by*

a differentially-private estimator that consistently estimates the sample's range. Examples of such an estimator can be found in Lei (2011). This remark is applicable to all truncation tuning parameters in algorithms that appear in Sections 3 and 4.

Differentially private algorithms in the classical setting

In the classical setting of $d \ll n$, the optimal rate of convergence of the mean estimation problem can be achieved simply by a noisy, truncated sample mean: given an i.i.d. sample X_1, X_2, \dots, X_n , the estimator is defined as

$$\hat{\boldsymbol{\mu}}_T := \frac{1}{n} \sum_{i=1}^n f_T(X_i) + W,$$

where $f_T(\cdot)$ denotes projection onto the L^∞ ball of radius $T > 0$ in \mathbb{R}^d , and W is an independent draw from $N_d\left(0, \frac{dT^2 \log(1.25/\delta)}{n^2 \varepsilon^2} \mathbf{I}_d\right)$. The theoretical guarantees for this estimator are summarized in the theorem below.

Theorem 15. *For an i.i.d. sample $X_1, X_2, \dots, X_n \sim P \in \mathcal{P}(\sigma, d)$ with $\mathbb{E}_P X = \boldsymbol{\mu}$ satisfying $\|\boldsymbol{\mu}\|_\infty = O(1)$, $\hat{\boldsymbol{\mu}}_T$ is an (ε, δ) -differentially private procedure, and:*

1. *if there exists a constant $M < \infty$ such that $\mathbb{P}(\|X\|_\infty < M) = 1$, when $T \geq M$,*

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}_T - \boldsymbol{\mu}\|_2 \lesssim \sigma \left(\sqrt{\frac{d}{n}} + \frac{d\sqrt{\log(1/\delta)}}{n\varepsilon} \right);$$

2. *otherwise, with the choice of $T \geq C\sigma\sqrt{\log n}$ for a sufficiently large constant $C > 0$,*

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}_T - \boldsymbol{\mu}\|_2 \lesssim \sigma \left(\sqrt{\frac{d}{n}} + \frac{d\sqrt{\log(1/\delta) \log n}}{n\varepsilon} \right).$$

By comparing with Theorem 12, we see that the noisy truncated sample mean achieves the optimal rate of convergence up to a factor of $\sqrt{\log n}$.

5.4. Privacy Cost of Linear Regression

In this section, we investigate the cost of differential privacy in linear regression problems, with primary focus on the high-dimensional setting where $d \gtrsim n$ and the regression coefficient β is assumed to be sparse; the classical, low-dimensional case ($d \ll n$) will also be covered. Through the general lower bound technique described in Section 5.2, we are able to establish minimax lower bounds that match the minimax rate of our differentially private procedures up to factor(s) of $\log n$.

5.4.1. Lower bound of high-dimensional linear regression

For high-dimensional sparse linear regression, we consider the following distribution space

$$\begin{aligned} \mathcal{P}_{X,Y}(\sigma, d, s) \\ = \{P(\mathbf{x}, y) \mid \|\mathbf{x}\|_\infty \lesssim 1, \epsilon := y - \mathbf{x}^\top \beta \sim P_\epsilon \in \mathcal{P}(\sigma, 1), \|\beta\|_0 \leq s, \|\beta\|_2 \leq C\}, \end{aligned}$$

where the parameter of interest is $\beta = \mathbb{E}[\mathbf{x}^\top \mathbf{x}]^{-1} \mathbb{E}[\mathbf{x}^\top y] \in \mathbb{R}^d$ is defined such that $X\beta$ is the best linear approximation of \mathbf{y} , and C is a generic constant. For brevity, we use P to denote $P(\mathbf{x}, y)$.

Let $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote an i.i.d. sample drawn from some $P \in \mathcal{P}_{X,Y}(\sigma, d, s)$, we propose the tracing adversary

$$\mathbf{a}_P((\mathbf{x}, y), M(D)) = \begin{cases} \text{IN} & \sum_{j \in S(M(D))} x_j(y - \tilde{y}) > \sigma \sqrt{2s \log(1/\delta)}, \\ \text{OUT} & \text{otherwise,} \end{cases}$$

where $S(M(D)) = \{j : M(D)_j \text{ is among the top } s \text{ largest coordinates of } M(D)\}$, and \tilde{y} is a fresh independent sample with covariates \mathbf{x} .

This adversary satisfies the following properties:

Lemma 18. *Suppose that $\|M(D) - \beta\|_\infty < \sigma/2$, then when $n \lesssim \sqrt{s/\log(1/\delta)} \log(d/s)$, there is a distribution $P \in \mathcal{P}_{X,Y}(\sigma, d, s)$ such that*

1. $\mathbb{P}(\{\mathcal{TR}(D, M(D)) = \emptyset\} \cap \{\|M(D) - \beta\|_2 \lesssim \sigma\sqrt{s}\}) \leq \delta$,
2. $\mathbb{P}(\mathbf{a}_P((\mathbf{x}_i, y_i), M(D'_i)) = \text{IN}) \lesssim \delta$, where D'_i is an adjacent dataset of D with (\mathbf{x}_i, y_i) replaced by $(\mathbf{x}'_i, y'_i) \sim P$.

The proof of this lemma, which appears in the supplementary material, includes a novel generalization of the fingerprinting lemma (see Tardos (2008), Bun et al. (2014), and Dwork et al. (2015)) to Gaussian random variables, which may be of independent interest.

We note that the extra assumption in Lemma 18 that $\|M(D) - \beta\|_\infty < \sigma/2$ can be gained “for free”: when it fails to hold, there would be an automatic lower bound that $\mathbb{E}\|M(D) - \beta\|_2 \gtrsim \sigma$. On the other hand, when $\|M(D) - \beta\|_\infty < \sigma/2$, the general lower bound result in Theorem 11 is applicable, and we obtain the following lower bound result.

Theorem 16. *Let $\mathcal{M}_{\varepsilon, \delta}$ denote the collection of all (ε, δ) -differentially private algorithms, and suppose the dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ consists of i.i.d. entries drawn from $\mathcal{P}_{X,Y}(\sigma, d, s)$. Suppose that $\varepsilon = O(1)$, $n^{-1}e^{-3\varepsilon n/2} \leq \delta \leq n^{-(1+\tau)}$ for some $\tau > 0$, and $\frac{\sqrt{s \log(1/\delta)} \log(d/s)}{n\varepsilon} = O(1)$, then*

$$\inf_{M \in \mathcal{M}_{\varepsilon, \delta}} \sup_{P \in \mathcal{P}_{X,Y}(\sigma, d, s)} \mathbb{E} [\|M(D) - \beta\|_2] \gtrsim \sigma \left(\sqrt{\frac{s \log d}{n}} + \frac{s \log(d/s) \sqrt{\log(1/\delta)}}{n\varepsilon} \wedge 1 \right).$$

Specifically, the second term in the lower bound is a consequence of Lemma 18 and Theorem 11. The first term is due to the statistical minimax lower bound for high-dimensional linear regression (see, for instance, Raskutti et al. (2009) and Ye and Zhang (2010)).

5.4.2. Upper bound of high-dimensional linear regression

For high-dimensional sparse linear regression, we propose the following differentially private LASSO algorithm, which splits the sample of size n into subsamples of size $O(\log n)$ and iterates through the subsamples by a truncated gradient descent with random perturbation.

Algorithm 4: Differentially Private LASSO

- 1: **Inputs:** privacy parameters δ, ε , design matrix X , response vector \mathbf{y} , step size η , sparsity tuning parameter \hat{s} , truncation tuning parameter T and the number of iterations N_0 .
- 2: Randomly split (X, \mathbf{y}) into N_0 subsets $(X_{(0)}, \mathbf{y}_{(0)}), (X_{(1)}, \mathbf{y}_{(1)}), \dots, (X_{(N_0-1)}, \mathbf{y}_{(N_0-1)})$ of size n/N_0 each.
- 3: Initialize the algorithm with an \hat{s} -sparse vector $\hat{\boldsymbol{\beta}}^{(0)}$.
- 4: **for** $t = 0, 1, 2, \dots, N_0 - 1$ **do**
- 5: $\hat{\boldsymbol{\beta}}^{(t+0.5)} = \hat{\boldsymbol{\beta}}^{(t)} - \eta \cdot \frac{1}{n/N_0} (X_{(t)}^\top X_{(t)} \hat{\boldsymbol{\beta}}^{(t)} - X_{(t)}^\top f_T(\mathbf{y}_{(t)}))$, where $f_T(\cdot)$ denotes projection onto the ℓ_∞ ball of radius $T > 0$ in \mathbb{R}^d .
- 6: $\hat{\boldsymbol{\beta}}^{(t+1)} = \text{Peeling} \left(\hat{\boldsymbol{\beta}}^{(t+0.5)}, \hat{s}, 4T/(n/N_0), \varepsilon, \delta \right)$.
- 7: **end for**
- 8: Output $\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}^{(N_0)}$.

Theorem 17. Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)$ be an i.i.d. sample drawn from $P(\mathbf{x}, y) \in \mathcal{P}_{X,Y}(\sigma, d, s)$. If we have that

- Σ_X , the covariance matrix of \mathbf{x} , satisfies $0 < 1/\Lambda < \lambda_{\min}(\Sigma_X) \leq \lambda_{\max}(\Sigma_X) < \Lambda$ for some constant $\Lambda > 0$,
- $\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}\|_2 \leq \kappa \|\boldsymbol{\beta}\|_2$ for some $\kappa \in (0, 1)$, and
- the tuning parameters satisfy $T \geq K\sigma\sqrt{\log n}$ for a sufficiently large constant $K > 0$, $N_0 \asymp \log n$, $\hat{s} \asymp s$ and for $\rho := \frac{\lambda_{\max}(\Sigma_X) - \lambda_{\min}(\Sigma_X)}{\lambda_{\max}(\Sigma_X) + \lambda_{\min}(\Sigma_X)}$, it holds that

$$\hat{s} \geq \max \left\{ \frac{4(1 + \kappa)^2}{(1 - \kappa)^2}, \left(\frac{4\rho}{1 - \rho} \right)^2 \right\} s$$

then $\hat{\boldsymbol{\beta}}$ is (ε, δ) -differentially private, and it holds with high probability that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \lesssim \sigma \left(\sqrt{\frac{s \log d \log n}{n}} + \frac{s \sqrt{\log(1/\delta)}}{n\varepsilon} \log d \cdot \log^{3/2} n \right).$$

To the best of our knowledge, this is the first differentially private LASSO algorithm with parameter estimation consistency guarantees. In addition, in view of Theorem 16, we see that the proposed algorithm achieves the optimal convergence rate up to a logarithm term $\log^{3/2} n$.

5.4.3. Linear regression in the classical setting

In the classical linear regression problem, we have i.i.d. observations $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ drawn from some P that belongs to the distribution space

$$\mathcal{P}_{X,Y}(\sigma, d) = \{P(\mathbf{x}, y) \mid \|\mathbf{x}\|_\infty \lesssim 1, \epsilon := y - \mathbf{x}^\top \boldsymbol{\beta} \sim P_\epsilon \in \mathcal{P}(\sigma, 1), \|\boldsymbol{\beta}\|_2 \leq C\},$$

where the parameter of interest is $\boldsymbol{\beta} = \mathbb{E}[\mathbf{x}^\top \mathbf{x}]^{-1} \mathbb{E}[\mathbf{x}^\top y] \in \mathbb{R}^d$ is defined such that $X\boldsymbol{\beta}$ is the best linear approximation of \mathbf{y} , and C is a generic constant.

To apply Theorem 11 to deriving the lower bound for the linear regression model, we consider the following tracing adversary:

$$\mathbf{a}_P((\mathbf{x}, y), M(D)) = \begin{cases} \text{IN} & \langle \mathbf{x}y - \mathbf{x}\tilde{y}, M(D) \rangle > \sigma^2 \sqrt{8d \log(1/\delta)}, \\ \text{OUT} & \text{otherwise,} \end{cases}$$

where \tilde{y} is a fresh independent draw with the same covariates \mathbf{x} as y .

The next lemma summarizes the soundness and completeness properties of the tracing adversary.

Lemma 19. *If $n \lesssim \sqrt{d/\log(1/\delta)}$ and $\|M(D) - \boldsymbol{\beta}\|_\infty \leq \sigma/2$, there is a distribution $P \in \mathcal{P}_{X,Y}(\sigma, d)$, such that*

1. $\mathbb{P}(\{\mathcal{TR}(D, M(D)) = \emptyset\} \cap \{\|M(D) - \boldsymbol{\beta}\|_2 \lesssim \sigma\sqrt{d}\}) \leq \delta$,
2. $\mathbb{P}(\mathbf{a}_P((\mathbf{x}_i, y_i), M(D'_i)) = \text{IN}) \leq \delta$, where D'_i is an adjacent dataset of D with (\mathbf{x}_i, y_i) replaced by $(\mathbf{x}'_i, y'_i) \sim P$.

As in the high-dimensional setting, the extra assumption in this lemma that $\|M(D) - \beta\|_\infty < \sigma/2$ can be gained “for free”.

Our minimax lower bound for private linear regression in the classical setting is presented in the theorem below:

Theorem 18. *Let $\mathcal{M}_{\varepsilon, \delta}$ denote the collection of all (ε, δ) -differentially private algorithms, and suppose that $\varepsilon = O(1)$, $n^{-1}e^{-3\varepsilon n/2} \leq \delta \leq n^{-(1+\tau)}$ for some $\tau > 0$ and $\frac{\sqrt{d \log(1/\delta)}}{n\varepsilon} = O(1)$, then*

$$\inf_{M \in \mathcal{M}_{\varepsilon, \delta}} \sup_{P \in \mathcal{P}_{X, Y}(\sigma, d)} \mathbb{E} [\|M(D) - \beta\|_2] \gtrsim \sigma \left(\sqrt{\frac{d}{n}} + \frac{d \sqrt{\log(1/\delta)}}{n\varepsilon} \wedge 1 \right).$$

Similar to the other lower bound results, the two terms in this minimax lower bound correspond to the statistical risk and the risk due to privacy constraint respectively.

Differentially private algorithms in the classical setting

In the classical setting of $d \ll n$, the optimal rate of convergence for differentially private linear regression can be directly achieved by perturbing the OLS estimator with suitably chosen noises.

Let $\hat{\beta} = \hat{\beta}(X, \mathbf{y}) := (X^\top X)^{-1} X^\top \mathbf{y}$ denote the OLS estimator, we consider the noisy estimator

$$\hat{\beta}_T := \hat{\beta}(X, f_T(\mathbf{y})) + W,$$

where $f_T(\cdot)$ denotes projection onto the ℓ_∞ ball of radius $T > 0$ in \mathbb{R}^d , and W is an independent draw from $N_d \left(0, \frac{dT^2 \log(1.25/\delta)}{n^2 \varepsilon^2} \mathbf{I}_d \right)$.

Theorem 19. *Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)$ be an i.i.d. sample drawn from $P(\mathbf{x}, y) \in \mathcal{P}_{X, Y}(\sigma, d)$. If we have that*

- Σ_X , the covariance matrix of \mathbf{x} , satisfies $0 < 1/\Lambda < \lambda_{\min}(\Sigma_X) \leq \lambda_{\max}(\Sigma_X) < \Lambda$ for some constant $\Lambda > 0$, and

- $T \geq K\sigma\sqrt{\log n}$ for a sufficiently large constant $K > 0$,

then with high probability, $\hat{\beta}_T$ is (ε, δ) -differentially private, and

$$\mathbb{E}\|\hat{\beta}_T - \beta\|_2 \lesssim \sigma \left(\sqrt{\frac{d}{n}} + \frac{d\sqrt{\log(1/\delta)\log n}}{n\varepsilon} \right).$$

This risk upper bound shows that the lower bound in Theorem 18 is optimal up to a factor of $\sqrt{\log n}$.

5.5. Simulation Studies

The proposed private algorithms can be implemented efficiently. In this section, we perform simulation studies of these algorithms to demonstrate the cost of privacy in different statistical estimation schemes, as well as the merits of the proposed algorithms. More specifically, we study the following four different problems.

Conventional mean estimation The data $\mathbf{x}_1, \dots, \mathbf{x}_n$ is an i.i.d. sample drawn from $N_d(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_d)$, where $\mu_1 = \dots = \mu_d = 1$ and $\sigma = 0.5$.

High-dimensional mean estimation The data $\mathbf{x}_1, \dots, \mathbf{x}_n$ is an i.i.d. sample drawn from $N_d(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_d)$, where $\sigma = 0.5$ and $\boldsymbol{\mu}$ is an s -sparse vector with the first s entries being 1 and the rest being 0.

Conventional linear regression The data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ are generated from the model $y = X\boldsymbol{\beta} + \epsilon$. In the simulation, the entries of design matrix are independently generated from Bernoulli(0.15), and $\epsilon_1, \dots, \epsilon_n$ is an i.i.d sample from $N(0, \sigma^2)$ with $\sigma = 0.5$. The coefficients $\boldsymbol{\beta}$ is set to $\beta_1 = \dots = \beta_d = 1$.

High-dimensional linear regression The data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ are generated from the model $y = X\boldsymbol{\beta} + \epsilon$. In the simulation, we set $\sigma = 0.5$ and the design matrix is generated the same way as in the conventional setting. The coefficients $\boldsymbol{\beta}$ is set to be s -sparse with $\beta_1 = \dots = \beta_s = 1$, and the rest are set to 0.

In all simulations, the privacy parameters (ϵ, δ) take values among $(0.5, 0.5)$, $(0.5, 0.1)$, $(0.5, 0.01)$, $(0.2, 0.01)$, and $(0.2, 0.001)$. In the low-dimensional problems, sample size and dimension (n, d) take values among $(10000, 50)$, $(50000, 50)$, $(10000, 100)$ to $(50000, 100)$; in the high-dimensional problems, sample size, dimension and sparsity (n, d, s) take values among $(2000, 2000, 20)$, $(4000, 2000, 20)$, $(2000, 4000, 20)$, $(4000, 4000, 20)$, $(2000, 2000, 30)$, $(4000, 2000, 30)$, $(2000, 4000, 30)$, and $(4000, 4000, 30)$.

In these simulation studies, we also compare the performance of (ϵ, δ) -differentially private methods with the optimal mechanisms under α -local differential privacy proposed in Duchi et al. (2018), where we set the local privacy parameter to be $\alpha = 10$, corresponding to a weak local privacy constraint. As there is no high-dimensional linear regression algorithms with α -local differential privacy in Duchi et al. (2018), we compare our algorithm with the locally private (low-dimensional) linear regression algorithm proposed in Duchi et al. (2018). The locally private linear regression algorithm is implemented with the knowledge of β 's support, which is not usually available in applications.

Tables 1-4 summarize the estimation errors with respect to ℓ_2 error ($\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2$ in mean estimation problems and $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$ in regression problems) for various methods. In the tables, NP, DP, LDP stand for the non-private algorithms, differentially private algorithms, and locally differentially private algorithms respectively. Each estimation error reported is the average over 100 replications of a given method; the standard error of each case is reported in parentheses. Our proposed differentially private algorithms outperform their locally private counterparts in Duchi et al. (2018), especially in high-dimensional problems. This is expected as Duchi et al. (2018) shows that it is impossible to construct consistent estimators in the high-dimensional problems with α -local differential privacy constraint.

Table 16: Conventional Mean Estimation

| (n, d) | NP | DP | | | | | LDP $\alpha = 10$ |
|-------------|------------|------------|------------|------------|------------|------------|----------------------|
| | | (.5, .5) | (.5, .1) | (.5, .01) | (.2, .01) | (.2, .001) | |
| (10000,50) | .014(.001) | .050(.007) | .084(.011) | .115(.012) | .311(.043) | .367(.058) | 5.435(.507) |
| (50000,50) | .006(.001) | .011(.001) | .018(.002) | .025(.004) | .063(.010) | .074(.010) | 4.260(.213) |
| (10000,100) | .020(.001) | .095(.001) | .168(.020) | .240(.026) | .591(.073) | .741(.088) | 9.523(.645) |
| (50000,100) | .009(.001) | .020(.002) | .035(.003) | .050(.005) | .121(.013) | .150(.015) | 5.411(.201) |

Table 17: High-dimensional Mean Estimation

| (n, d, s) | NP | DP | | | | | LDP $\alpha = 10$ |
|----------------|------------|------------|-------------|------------|-------------|-------------|----------------------|
| | | (.5, .5) | (.5, .1) | (.5, .01) | (.2, .01) | (.2, .001) | |
| (2000,2000,20) | .154(.010) | .179(.024) | .277(.059) | .380(.074) | .882(.162) | 1.173(.271) | 309.999(1.504) |
| (4000,2000,20) | .110(.005) | .121(.015) | .155(.025) | .172(.027) | .452(.100) | .549(.102) | 218.269(1.805) |
| (2000,4000,20) | .162(.009) | .183(.024) | .276(.044) | .361(.085) | .860(.164) | 1.537(.412) | 613.808(4.187) |
| (4000,4000,20) | .117(.006) | .124(.017) | .143(.023) | .191(.035) | .459(.083) | .524(.091) | 432.422(3.070) |
| (2000,2000,30) | .154(.007) | .206(.029) | .347(.049) | .464(.085) | 1.195(.204) | 2.097(.405) | 311.486(2.556) |
| (4000,2000,30) | .110(.006) | .135(.020) | .178(.027) | .238(.034) | .578(.101) | .749(.180) | 220.502(1.761) |
| (2000,4000,30) | .166(.008) | .227(.030) | .361(.074) | .466(.092) | 1.388(.302) | 2.931(.496) | 613.703(4.272) |
| (4000,4000,30) | .117(.007) | .136(.199) | .182(.0259) | .245(.043) | .600(.107) | .731(.122) | 439.823(1.533) |

Table 18: Conventional Linear Regression

| (n, d) | NP | DP | | | | | LDP $\alpha = 10$ |
|-------------|------------|------------|------------|------------|-------------|-------------|----------------------|
| | | (.5, .5) | (.5, .1) | (.5, .01) | (.2, .01) | (.2, .001) | |
| (10000,50) | .061(.014) | .111(.017) | .179(.028) | .233(.042) | .624(.108) | .747(.109) | 85.394(5.441) |
| (50000,50) | .030(.006) | .032(.005) | .044(.006) | .056(.008) | .120(.020) | .148(.018) | 58.412(5.511) |
| (10000,100) | .087(.012) | .206(.022) | .350(.039) | .499(.045) | 1.200(.126) | 1.508(.133) | 297.885(13.325) |
| (50000,100) | .041(.006) | .055(.006) | .079(.008) | .102(.010) | .245(.026) | .294(.028) | 250.424(17.333) |

Table 19: High-dimensional Linear Regression

| (n, d, s) | NP | DP | | | | | LDP $\alpha = 10$ |
|----------------|------------|------------|------------|------------|------------|------------|----------------------|
| | | (.5, .5) | (.5, .1) | (.5, .01) | (.2, .01) | (.2, .001) | |
| (2000,2000,20) | .032(.006) | .033(.004) | .052(.006) | .076(.010) | .176(.023) | .233(.024) | 19.278(1.518) |
| (4000,2000,20) | .022(.002) | .022(.003) | .035(.006) | .043(.007) | .099(.013) | .110(.023) | 18.730(1.741) |
| (2000,4000,20) | .033(.005) | .034(.005) | .052(.005) | .080(.011) | .199(.025) | .206(.190) | 18.899(2.239) |
| (4000,4000,20) | .022(.003) | .023(.004) | .034(.004) | .038(.005) | .096(.011) | .118(.013) | 18.797(2.098) |
| (2000,2000,30) | .037(.006) | .037(.007) | .058(.006) | .104(.012) | .239(.038) | .283(.032) | 38.918(3.224) |
| (4000,2000,30) | .026(.004) | .027(.004) | .035(.004) | .053(.005) | .118(.016) | .130(.021) | 34.552(1.982) |
| (2000,4000,30) | .039(.004) | .040(.005) | .061(.006) | .104(.014) | .238(.038) | .308(.028) | 36.124(3.361) |
| (4000,4000,30) | .027(.003) | .028(.003) | .035(.005) | .055(.005) | .116(.023) | .147(.018) | 32.970(1.947) |

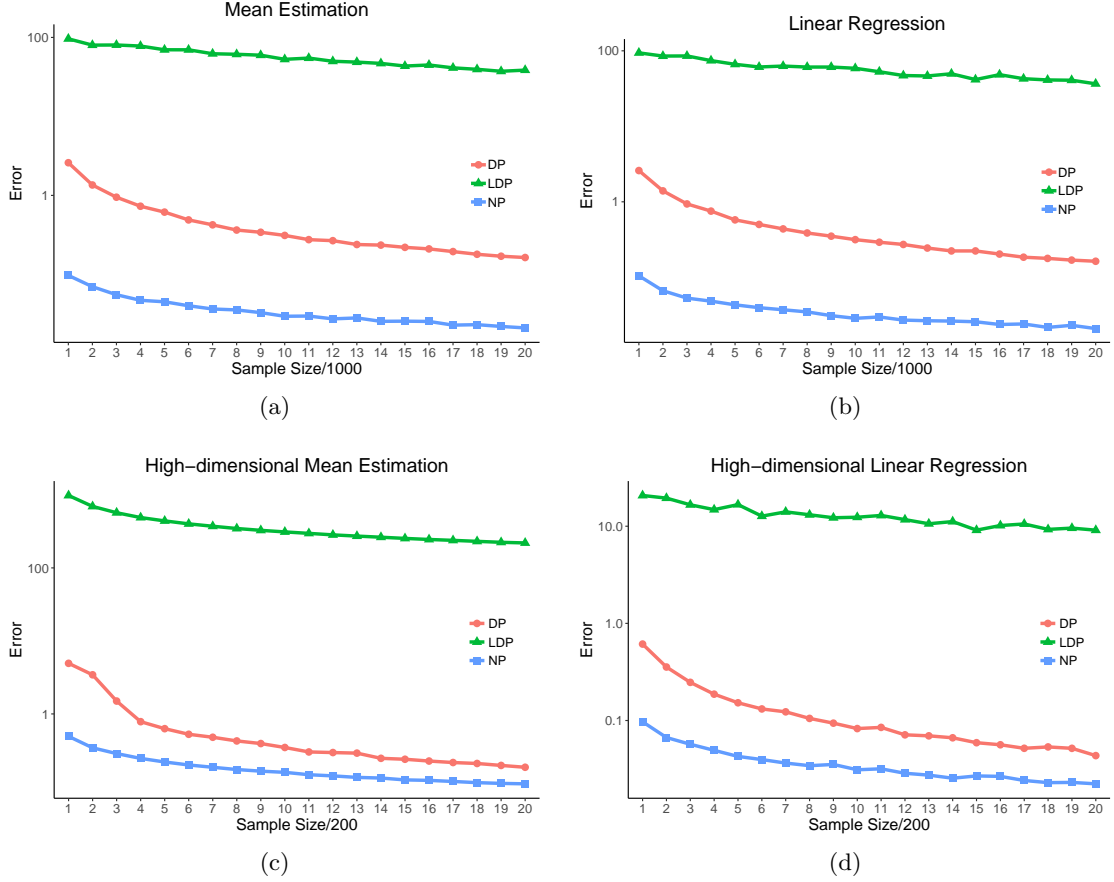


Figure 3: Errors (in \log_{10} -scale) plotted against sample size n , with $(0.5, 10/n^{1.1})$ -differentially privacy guarantee. Top-left: conventional mean estimation with sample size ranging from 1000 to 1000×20 ; top-right: conventional linear regression with sample size ranging from 1000 to 1000×20 ; bottom-left: high-dimensional mean estimation with sample size ranging from 200 to 200×20 ; bottom-right: high-dimensional linear regression with sample size ranging from 200 to 200×20 . The local differentially private algorithms (LDP), differentially private (DP) algorithms and non-private (NP) algorithms are colored in green, red and blue respectively.

As seen in Figure 1 (error in \log_{10} -scale), the gap in estimation errors between the non-private algorithms and differentially private algorithms diminishes as the sample size n increases.

5.6. Data Analysis

5.6.1. SNP array of adults with schizophrenia

We analyze the SNP array data of adults with schizophrenia, collected by Lowther et al. (2017), to illustrate the performance of our high-dimensional sparse mean estimator. In the

dataset, there are 387 adults with schizophrenia, 241 of which are labeled as “average IQ” and 146 of which are labeled as “low IQ”. The SNP array is obtained by genotyping the subjects with the Affymetrix Genome-Wide Human SNP 6.0 platform. For our analysis, we focus on the 2000 SNPs with the highest minor allele frequencies (MAFs); the full dataset is available at <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106818>.

Privacy-perserving data analysis is very much relevant for this dataset and genetic data in general, because as Homer et al. (2008) demonstrates, an adversary can infer the absence/presence of an individual’s genetic data in a large dataset by cross-referencing summary statistics, such as MAFs, from multiple genetic datasets. As MAFs can be easily calculated from the mean of an SNP array, differentially-private estimators of the mean can effectively allow reporting the MAFs without compromising any individual’s privacy.

The data set takes the form of a 387×2000 matrix. The entries of the matrix take values 0, 1 or 2, representing the number of minor allele(s) at each SNP, and therefore the MAF of each SNP location in this sample can be obtained by computing the mean of the rows in this matrix. Sparsity is introduced by considering the difference in MAFs of the two IQ groups: the MAFs of the two groups are likely to differ at a small number of SNP locations among the 2000 SNPs considered.

For m ranging from 10 to 120, we subsample m subjects from each of the two IQ groups, say $\{\mathbf{x}_{11}, \mathbf{x}_{12}, \dots, \mathbf{x}_{1m}\}$ and $\{\mathbf{x}_{21}, \mathbf{x}_{22}, \dots, \mathbf{x}_{2m}\}$, and apply our sparse mean estimator to $\{\mathbf{x}_{11} - \mathbf{x}_{21}, \mathbf{x}_{12} - \mathbf{x}_{22}, \dots, \mathbf{x}_{1m} - \mathbf{x}_{2m}\}$ with $\hat{s} = 20$ and privacy parameters $(\epsilon, \delta) = (0.2, n^{-1.1})$. The error of this estimator is then calculated by comparing with the mean of the entire sample. This procedure is repeated 100 times to obtain Figure 4, which displays the bootstrap estimate of $\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2]/d$ as m increases from 10 to 120. The performance of the sparse mean estimator in Duchi et al. (2018), with privacy parameter $\alpha = 10$, is also plotted for comparison.

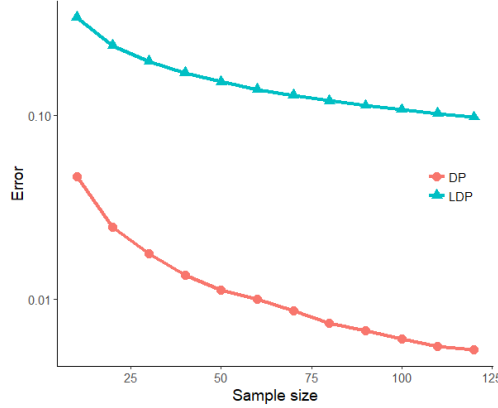


Figure 4: The bootstrap estimate of $\mathbb{E}[\|\hat{\mu} - \mu\|_2]/d$ for the differentially private sparse mean estimator, compared with its locally differentially private counterpart, as sample size increases from 10 to 120.

5.6.2. Housing prices in California

For the linear regression problem, we analyze a housing price dataset with economic and demographic covariates, constructed by Pace and Barry (1997) and available for download at <http://lib.stat.cmu.edu/datasets/houses.zip>. In this dataset, each subject is a block group in California in the 1990 Census; there are 20640 block groups in this dataset. The response variable is the median house value in the block group; the covariates include the median income, median age, total population, number of households, and the total number of rooms of all houses in the block group. In general, summary statistics such as mean or median do not have any differential privacy guarantees, so the absence of information on individual households in the dataset does not preclude an adversary from extracting sensitive individual information from the summary statistics. Privacy-preserving methods are still desirable in this case.

For m ranging from 100 to 20600, we subsample m subjects from the dataset to compute the differentially private OLS estimate, with privacy parameters $(\epsilon, \delta) = (0.2, n^{-1.1})$. The error of this estimator is then calculated by comparing with the non-private OLS estimator computed using the entire sample. This procedure is repeated 100 times to obtain Figure 5, which displays the trend of the bootstrap estimate of $\mathbb{E}[\|\hat{\beta} - \beta\|_2]/d$ as m increases from

100 to 20600. The performance of the linear regression method in Duchi et al. (2018), with privacy parameter $\alpha = 10$, is also plotted for comparison.

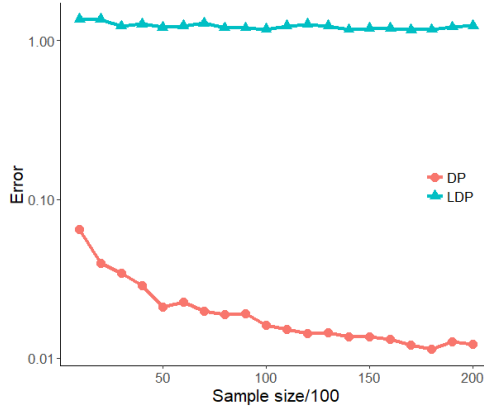


Figure 5: The bootstrap estimate of $\mathbb{E}[\|\hat{\beta} - \beta\|_2]/d$ for the differentially private OLS estimator, compared with its locally differentially private counterpart, as sample size increases from 100 to 20600.

5.7. Discussion

In summary, this paper characterizes the tradeoff between statistical accuracy and privacy guarantees, by providing information-theoretic lower bounds for estimation with differential privacy constraint and differentially private algorithms that has matching upper bounds. For the lower bounds, as standard packing arguments fail to establish sharp results under privacy constraints, we have developed a novel lower bound technique based on tracing adversary. The utility of the new technique is illustrated by establishing minimax lower bounds for differentially private mean estimation and linear regression, in both low-dimensional and high-dimensional settings. We have also proposed computationally efficient algorithms with matching upper bounds up to logarithmic factors.

This line of work can be extended to designing rate-optimal algorithms with (ϵ, δ) -differential privacy guarantee for a greater variety of statistical problems. For instance, the results in the current paper are applicable to estimation of moment-based statistics, such as mean and covariance matrices estimation. It would also be interesting to generalize the results

further to (high-dimensional) empirical risk minimization, (high-dimensional) classification, and nonparametric estimation such as density estimation and nonparametric regression. We are also interested in studying the statistical cost of other notions of privacy, such as concentrated differential privacy (Dwork and Rothblum, 2016) and Renyi differential privacy (Mironov, 2017). These notions of privacy have found many applications such as stochastic gradient Langevin dynamics and stochastic Monte Carlo sampling (Wang et al., 2015).

In addition, as we deepen our understanding of statistical estimation problems with privacy constraints, the next goal should ideally be uncertainty quantification, i.e. statistical inference, with privacy constraints, which is largely unexplored in the statistics literature. We hope to investigate the rate-optimal length of a confidence interval, and the optimal power in hypothesis testing with the constraint of (ε, δ) -differential privacy.

5.8. Proofs

In this section, we prove the main results, Theorem 11, the general lower bound argument, and Theorem 17, the minimax risk upper bound of the private high-dimensional linear regression. For reasons of space, the proofs of other results and technical lemmas are provided in the supplementary material (Cai et al., 2019c).

5.8.1. Proof of Theorem 11

In this section, we prove Theorem 11, the general approach to obtain lower bound with privacy constraint. The applications of Theorem 11 to different models to obtain lower bounds are discussed in the supplementary material (Cai et al., 2019c).

The proof of Theorem 11 consists of three steps, as follows.

Step 1: A preliminary lower bound

For every (ε, δ) -differentially private M , the tracing adversary $\mathbf{a}_P(\cdot, M(X))$ that post-processes M is (ε, δ) -differentially private as well. It follows that for every $i \in [n]$,

$$\mathbb{P}(\mathbf{a}_P(\mathbf{x}_i, M(X)) = \text{IN}) \leq e^\varepsilon \cdot \mathbb{P}(\mathbf{a}_P(\mathbf{x}_i, M(X'_i)) = \text{IN}) + \delta \leq (e^\varepsilon + 1) \delta.$$

Then for $\mathcal{TR}(X, M(X)) := \{i \in [n] : \mathbf{a}_P(\mathbf{x}_i, M(X)) = \text{IN}\}$, the union bound leads to

$$\mathbb{P}(\mathcal{TR}(X, M(X)) \neq \emptyset) \leq \sum_{i \in [n]} \mathbb{P}(\mathbf{a}_P(\mathbf{x}_i, M(X)) = \text{IN}) \leq n(e^\varepsilon + 1) \delta.$$

This inequality and the completeness property together imply that

$$\begin{aligned} & \mathbb{P}(\rho(M(X), \boldsymbol{\theta}(P)) \lesssim \lambda(\mathcal{P}, \delta)) \\ & \leq \mathbb{P}(\{\mathcal{TR}(X, M(X)) = \emptyset\} \cap \{\rho(M(X), \boldsymbol{\theta}(P)) \lesssim \lambda(\mathcal{P}, \delta)\}) \\ & \quad + \mathbb{P}(\mathcal{TR}(X, M(X)) \neq \emptyset) \\ & \leq \delta + n(e^\varepsilon + 1) \delta. \end{aligned}$$

We have $\delta < n^{-(1+\tau)}$ by assumption, so $\mathbb{P}(\rho(M(X), \boldsymbol{\theta}(P)) \lesssim \lambda(\mathcal{P}, \delta)) \leq \delta + n(e^\varepsilon + 1) \delta$ is bounded away from 1. Markov's inequality and the monotonicity of l immediately yield a preliminary lower bound for $n \lesssim \psi(\mathcal{P}, \delta)$:

$$\mathbb{E}[l(\rho(M(X), \boldsymbol{\theta}(P)))] \gtrsim l(\lambda(\mathcal{P}, \delta)).$$

Step 2: An improvement by group privacy

In this step, we show that the preliminary lower bound found in Step 1 is valid for $n \lesssim \psi(\mathcal{P}, \delta) \log(1/\delta)/\varepsilon$.

Consider the following construction: let $k = C \log(\frac{1}{n\delta})/\varepsilon$ and assume that k divides n

without the loss of generality. Since $\frac{1}{n}e^{-3\epsilon n/2} \leq \delta < n^{-(1+\tau)}$, we have $(C\tau/\epsilon) \log n < k \leq n$. The value of C is to be specified later.

We first draw an i.i.d. sample of size n/k , denoted by $\tilde{X} = \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \dots, \tilde{\mathbf{x}}_{n/k}\}$, from P , then sample with replacement from \tilde{X} for n times to obtain $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$. For any $M \in \mathcal{M}_{\epsilon, \delta}$, we define $M_k(\tilde{X}) \equiv M(X)$. Because M is (ϵ, δ) -differentially private, M_k is also differentially private, thanks to the following group privacy lemma:

Lemma 20 (group privacy, Steinke and Ullman (2017)). *For every $m \geq 1$, if M is (ϵ, δ) -differentially private, then for every pair of datasets $X = \{\mathbf{x}_k\}_k$ and $Z = \{\mathbf{z}_k\}_k$ satisfying $\sum_k \mathbb{1}(\mathbf{x}_i \neq \mathbf{z}_i) \leq m$, and every event S ,*

$$\mathbb{P}_{M, X}(S) \leq e^{\epsilon m} \mathbb{P}_{M, Z}(S) + \frac{e^{\epsilon m} - 1}{e - 1} \cdot \delta.$$

The group privacy lemma means that, to characterize the privacy parameters of M_k , it suffices to upper-bound the number of changes in X incurred by replacing one element of \tilde{X} : let m_i denote the number of times that $\tilde{\mathbf{x}}_i$ appears in a sample of size n drawn with replacement from \tilde{X} , then our quantity of interest here is simply $\max_{i \in [n]} m_i$. We shall analyze $\max_{i \in [n]} m_i$ under two separate scenarios: (1). $(1 + \tau) \log n \leq \log(\frac{1}{\delta}) \leq (1 + 2\tau) \log n$ for some $\tau > 0$, and (2). $\log(\frac{1}{\delta}) \gg (1 + \tau) \log n$ for all $\tau > 0$.

(1). $(1 + \tau) \log n \leq \log(\frac{1}{\delta}) \leq (1 + 2\tau) \log n$: under this setting, we have $k = C \log(\frac{1}{n\delta})/\epsilon \asymp (C\tau/\epsilon) \log n$. The analysis makes use of a result from Raab and Steger (1998), stated below:

Lemma 21 (Raab and Steger (1998)). *If (x_1, x_2, \dots, x_d) follows a uniform multinomial(ℓ) distribution, and $\frac{\ell}{d \log d} \asymp c$ for some constant c , then for every $\zeta > 0$,*

$$\mathbb{P} \left(\max_{i \in [d]} x_i > (r_c + \zeta) \log d \right) = o(1),$$

where r_c is the unique root of $1 + x(\log c - \log x + 1) - c = 0$ that is strictly greater than c .

We apply the lemma to obtain that, for any $\zeta > 0$,

$$\mathbb{P}(\max_i m_i > (r + \zeta) \log n) = o(1),$$

where r is the unique root of $1 + x(\log(C\tau/\varepsilon) - \log x + 1) - (C\tau/\varepsilon) = 0$ that is greater than $C\tau/\varepsilon$. To see the existence of such a root, note that $f_{C,\tau,\varepsilon}(x) := 1 + x(\log(C\tau/\varepsilon) - \log x + 1) - (C\tau/\varepsilon)$ is strictly concave and achieves the global maximum value of 1 at $x = C\tau/\varepsilon$.

It follows from Lemma 20 that, with high probability, M_k is an $(\varepsilon(r + \zeta) \log n, \delta e^{\varepsilon(r + \zeta) \log n})$ -differentially private algorithm. Then we repeat the lower bound argument in Step 1, the key ingredient of which is showing that $\mathbb{P}(\mathcal{TR}(\tilde{X}, M_k(\tilde{X})) \neq \emptyset)$ is bounded away from 0. First, for every $i \in [n/k]$,

$$\begin{aligned} & \mathbb{P}(\mathbf{a}_P(\tilde{\mathbf{x}}_i, M_k(\tilde{X})) = \text{IN}) \\ & \leq e^{\varepsilon(r + \zeta) \log n} \mathbb{P}(\mathbf{a}_P(\tilde{\mathbf{x}}_i, M_k(\tilde{X}'_i)) = \text{IN}) + \delta e^{\varepsilon(r + \zeta) \log n} \\ & \leq 2\delta e^{\varepsilon(r + \zeta) \log n} = 2n^{-(1 + \tau) + \varepsilon(r + \zeta)}. \end{aligned}$$

By the union bound,

$$\mathbb{P}(\mathcal{TR}(\tilde{X}, M_k(\tilde{X})) \neq \emptyset) \leq \sum_{i \in [n/k]} \mathbb{P}(A_P(\tilde{\mathbf{x}}_i, M_k(\tilde{X})) = \text{IN}) \leq 2n^{-\tau + \varepsilon(r + \zeta)}.$$

We claim that this probability is always bounded away from 1, because $\varepsilon r < \tau$ with appropriately chosen C : since $f_{C,\tau,\varepsilon}(\tau/\varepsilon) = (\tau/\varepsilon)(1 + \log C - C) + 1$ and $\varepsilon = O(1)$, for every $\tau > 0$ there is a sufficiently small $C > 0$ such that $f_{C,\tau,\varepsilon}(\tau/\varepsilon) < 0$. Since $f_{C,\tau,\varepsilon}(C\tau/\varepsilon) = 1$ is the global maximum, we have $r < \tau/\varepsilon$, or equivalently $\varepsilon r < \tau$, as desired.

(2). $\log(\frac{1}{\delta}) \gg (1 + \tau) \log n$: under this setting, we have $k = C \log(\frac{1}{n\delta})/\varepsilon \gg \log n$.

Each m_i is a sum of n independent Bernoulli(k/n) random variables. We apply Chernoff inequality: since $\mathbb{E}m_i = k$, we have

$$\mathbb{P}(m_i \geq 3k/2) \leq \exp(-k/12).$$

The union bound yields

$$\mathbb{P}(\max m_i > 3k/2) \leq \sum_{i \in [n/k]} \mathbb{P}(m_i \geq 3k/2) \leq n \exp(-k/12) = o(1),$$

since $k \gg \log n$.

By Lemma 20, M_k is a $(3\epsilon k/2, \delta e^{3\epsilon k/2})$ -differentially private algorithm with high probability: for every $i \in [n/k]$,

$$\begin{aligned} \mathbb{P}\left(\mathbf{a}_P(\tilde{\mathbf{x}}_i, M_k(\tilde{X})) = \text{IN}\right) &\leq e^{3\epsilon k/2} \mathbb{P}\left(\mathbf{a}_P(\tilde{\mathbf{x}}_i, M_k(\tilde{X}'_i)) = \text{IN}\right) + \delta e^{3\epsilon k/2} \\ &\leq 2\delta e^{3\epsilon k/2}. \end{aligned}$$

It follows that

$$\mathbb{P}\left(\mathcal{TR}(\tilde{X}, M_k(\tilde{X})) \neq \emptyset\right) \leq \sum_{i \in [n/k]} \mathbb{P}\left(\mathbf{a}_P(\tilde{\mathbf{x}}_i, M_k(\tilde{X})) = \text{IN}\right) \leq 2n\delta e^{3\epsilon k/2}.$$

Choosing $k < \frac{2}{3} \log(\frac{1}{2n\delta})/\epsilon$ guarantees that the probability is bounded away from 1.

To summarize, in both settings of δ , we have $\mathbb{P}\left(\mathcal{TR}(\tilde{X}, M_k(\tilde{X})) \neq \emptyset\right)$ bounded away from 1. A similar argument via Markov's inequality as in Step 1 shows, when $n/k \lesssim \psi(\mathcal{P}, \delta)$, equivalently $n \lesssim k\psi(\mathcal{P}, \delta) \asymp \psi(\mathcal{P}, \delta) \log(1/\delta)/\epsilon$, we have

$$\begin{aligned} \mathbb{P}(\rho(M_k(\tilde{X}), \boldsymbol{\theta}(P)) \lesssim \lambda(\mathcal{P}, \delta)) &\lesssim \lambda(\mathcal{P}, \delta) \\ &\leq \mathbb{P}(\{\mathcal{TR}(\tilde{X}, M_k(\tilde{X})) = \emptyset\} \cap \{\rho(M_k(\tilde{X}), \boldsymbol{\theta}(P)) \lesssim \lambda(\mathcal{P}, \delta)\}) \\ &\quad + \mathbb{P}(\mathcal{TR}(\tilde{X}, M_k(\tilde{X})) \neq \emptyset) < 1. \end{aligned}$$

Since $M_k(\tilde{X}) = M(X)$ by construction, we have extended the range over which the lower bound of $\mathbb{E}[l(\rho(M(X), \boldsymbol{\theta}(P)))] \geq l(\lambda(\mathcal{P}, \delta))$ is valid by an extra factor of $\log(1/\delta)/\varepsilon$.

Step 3: Establishing the lower bound for large n

If $n \gtrsim \psi(\mathcal{P}, \delta) \log(1/\delta)/\varepsilon$, we can choose $0 < \alpha < 1$ such that $n\alpha \asymp \psi(\mathcal{P}, \delta) \log(1/\delta)/\varepsilon$. Consider $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ independently drawn from the mixture distribution $\tilde{P} = \alpha P_{\tilde{\theta}} + (1 - \alpha)P_0 \in \mathcal{P}$, which is assumed to satisfy $\boldsymbol{\theta}(\tilde{P}) = \boldsymbol{\theta}(\alpha P_{\tilde{\theta}} + (1 - \alpha)P_0) = \alpha \tilde{\theta}$, by our regularity conditions on \mathcal{P} and $\boldsymbol{\theta}$.

We then claim that with high probability, $\sum_{i=1}^n \mathbf{1}(\boldsymbol{\theta}(\mathbf{x}_i) = \tilde{\theta}) \asymp n\alpha$: by Chernoff inequality

$$\mathbb{P}\left(\left|\sum_{i=1}^n \mathbf{1}(\boldsymbol{\theta}(\mathbf{x}_i) = \tilde{\theta}) - n\alpha\right| > \frac{n\alpha}{2}\right) \leq 2 \exp(-n\alpha/10) = o(1).$$

Let $A = \{i \in [n] : \boldsymbol{\theta}(\mathbf{x}_i) = \tilde{\theta}\}$; for every $M(X) \equiv M(\mathbf{x}_1, \dots, \mathbf{x}_n)$, we define $\tilde{M}(X_A) = \frac{1}{\alpha} \mathbb{E}_{X_{[n] \setminus A}} M(X)$.

As ρ is induced by a norm, it must be convex in its first argument: for every $\lambda \in [0, 1]$,

$$\begin{aligned} \rho(\lambda x + (1 - \lambda)y, a) &= \|\lambda x + (1 - \lambda)y - a\| \\ &\leq \lambda \|x - a\| + (1 - \lambda) \|y - a\| \\ &= \lambda \rho(x, a) + (1 - \lambda) \rho(y, a). \end{aligned}$$

By convexity, Jensen's inequality implies that

$$\begin{aligned} \mathbb{E}[\rho(M(X), \boldsymbol{\theta}(\tilde{P}))] &\geq \mathbb{E}_{X_A}[\rho(\mathbb{E}_{X_{[n] \setminus A}} M(X), \alpha \tilde{\theta})] = \mathbb{E}_{X_A}[\rho(\alpha \tilde{M}(X_A), \alpha \tilde{\theta})] \\ &\gtrsim \alpha \lambda(\mathcal{P}, \delta). \end{aligned}$$

The last inequality follows from the lower bound developed in the previous steps, since

$\text{card}(A) = \Theta(n\alpha) \lesssim \psi(\mathcal{P}, \delta) \log(1/\delta)/\varepsilon$. Because $n\alpha \asymp \psi(\mathcal{P}, \delta) \log(1/\delta)/\varepsilon$, we have

$$\mathbb{E}[\rho(M(X), \boldsymbol{\theta}(\tilde{P}))] \gtrsim \alpha \lambda(\mathcal{P}, \delta) \asymp \frac{\lambda(\mathcal{P}, \delta) \psi(\mathcal{P}, \delta) \log(1/\delta)}{n\varepsilon}.$$

5.8.2. Proof of Theorem 17

Proof. First, we introduce some useful notation: for a vector $\mathbf{v} \in \mathbb{R}^k$ and a set $\mathcal{S} \subseteq [k]$, let $\text{trunc}(\mathbf{v}, \mathcal{S})$ denote the vector obtained by setting $v_i = 0$ for $i \notin \mathcal{S}$. We also denote $n/N_0 \equiv n_0$ for brevity.

Privacy Guarantee: Because of sample splitting, for $(\mathbf{x}, y) \in (X_{(t)}, \mathbf{y}_{(t)})$ for some $0 \leq t \leq N_0 - 1$, it suffices to prove the privacy guarantee for the t -th iteration of the algorithm: any iteration prior to the t -th does not depend on (\mathbf{x}, y) , while any iteration after the t -th is differentially private by post-processing.

At the t -th iteration, the algorithm first updates the non-sparse estimate of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}^{(t+0.5)} = \hat{\boldsymbol{\beta}}^{(t)} - \eta \cdot \frac{1}{n_0} (X_{(t)}^\top f_T(X_{(t)}, \hat{\boldsymbol{\beta}}^{(t)}) - X_{(t)}^\top f_T(\mathbf{y}_{(t)}))$$

We observe that $\hat{\boldsymbol{\beta}}^{(t)}$ does not depend on $(X_{(t)}, \mathbf{y}_{(t)})$, so the Peeling step applied to $\hat{\boldsymbol{\beta}}^{(t+0.5)}$ would be (ε, δ) -differentially private if it can be shown that: for every $(\tilde{X}_{(t)}, \tilde{\mathbf{y}}_{(t)})$ obtained by replacing one individual in $(X_{(t)}, \mathbf{y}_{(t)})$, we have

$$\left\| (X_{(t)}^\top f_T(X_{(t)}, \hat{\boldsymbol{\beta}}^{(t)}) - X_{(t)}^\top f_T(\mathbf{y}_{(t)})) - (\tilde{X}_{(t)}^\top f_T(\tilde{X}_{(t)}, \hat{\boldsymbol{\beta}}^{(t)}) - \tilde{X}_{(t)}^\top f_T(\tilde{\mathbf{y}}_{(t)})) \right\|_\infty \lesssim T.$$

This fact is straightforward to show thanks to the ℓ_∞ truncations applied to $X_{(t)}, \hat{\boldsymbol{\beta}}^{(t)}$ and $\mathbf{y}_{(t)}$. Without the loss of generality, assume that $(\tilde{X}_{(t)}, \tilde{\mathbf{y}}_{(t)})$ and $(X_{(t)}, \mathbf{y}_{(t)})$ differ by (\mathbf{x}, y) and $(\tilde{\mathbf{x}}, \tilde{y})$, we calculate:

$$\|(f_T(\mathbf{y}) - f_T(\mathbf{x}^\top \boldsymbol{\beta}))\mathbf{x} - (f_T(\tilde{\mathbf{y}}) - f_T(\tilde{\mathbf{x}}^\top \boldsymbol{\beta}))\tilde{\mathbf{x}}\|_\infty \leq 2T(\|\mathbf{x}\|_\infty + \|\tilde{\mathbf{x}}\|_\infty) \lesssim T.$$

Then the privacy guarantee is proved by Lemma 17.

Statistical Accuracy: We define

$$\begin{aligned}\bar{\beta}^{(t+0.5)} &= \hat{\beta}^{(t)} - \eta \cdot \left(\mathbb{E}_{X,y} \left[\frac{1}{n_0} X_{(t)}^\top X_{(t)} \hat{\beta}^{(t)} \right] - \mathbb{E}_{X,y} \left[\frac{1}{n_0} X_{(t)}^\top \mathbf{y}_{(t)} \right] \right), \\ \hat{\beta}^{(t+0.5)} &= \hat{\beta}^{(t)} - \eta \cdot \frac{1}{n_0} \left(X_{(t)}^\top f_T(X_{(t)} \hat{\beta}^{(t)}) - X_{(t)}^\top f_T(\mathbf{y}_{(t)}) \right), \\ \bar{\beta}^{(t+1)} &= \text{trunc}(\bar{\beta}^{(t+0.5)}, \hat{\mathcal{S}}^{(t+0.5)}),\end{aligned}$$

where $\hat{\mathcal{S}}^{(t+0.5)}$ is the index set selected by applying Peeling to $\hat{\beta}^{(t+0.5)}$. Let $\beta^* := \mathbb{E}_{\mathbf{x},y}[\mathbf{x}^\top \mathbf{x}]^{-1} \mathbb{E}_{\mathbf{x},y}[\mathbf{x}^\top \mathbf{y}]$ be the true parameter. Throughout our calculations below, we treat $\hat{\beta}^{(t)}$ as a deterministic quantity, because it does not depend on $(X_{(t)}, \mathbf{y}_{(t)})$ by the design of our algorithm.

We have

$$\|\hat{\beta}^{(t+1)} - \beta^*\|_2 \leq \|\hat{\beta}^{(t+1)} - \bar{\beta}^{(t+1)}\|_2 + \|\bar{\beta}^{(t+1)} - \beta^*\|_2.$$

We shall provide upper bounds for the two terms on the right hand side separately.

For $\|\hat{\beta}^{(t+1)} - \bar{\beta}^{(t+1)}\|_2$, let \mathbf{W} denote the vector of Laplace noises of $|\hat{\mathcal{S}}^{(t+0.5)}| = \hat{s}$ dimensions that is generated when the Peeling algorithm outputs the noisy top \hat{s} coordinates of $\hat{\beta}^{(t+0.5)}$, we have

$$\begin{aligned}\mathbb{E}\|\hat{\beta}^{(t+1)} - \bar{\beta}^{(t+1)}\|_2 &\leq \mathbb{E}\|\text{trunc}(\hat{\beta}^{(t+0.5)}, \hat{\mathcal{S}}^{(t+0.5)}) - \text{trunc}(\bar{\beta}^{(t+0.5)}, \hat{\mathcal{S}}^{(t+0.5)})\|_2 + \|\mathbf{W}\|_2 \\ &\lesssim \sqrt{\hat{s}} \mathbb{E}\|\hat{\beta}^{(t+0.5)} - \bar{\beta}^{(t+0.5)}\|_\infty + \sqrt{\hat{s}} \mathbb{E}\|\mathbf{W}\|_\infty \\ &\lesssim \sqrt{\hat{s}} \mathbb{E}\|\hat{\beta}^{(t+0.5)} - \bar{\beta}^{(t+0.5)}\|_\infty + \sqrt{\hat{s}} \cdot \frac{T}{n_0} \cdot \left(\frac{\varepsilon}{\sqrt{\hat{s} \log(1/\delta)}} \right)^{-1} \cdot \log d \\ &\lesssim \sqrt{\hat{s}} \mathbb{E}\|\hat{\beta}^{(t+0.5)} - \bar{\beta}^{(t+0.5)}\|_\infty + \frac{T s \sqrt{\log(1/\delta)}}{n_0 \varepsilon} \log d.\end{aligned}$$

$\mathbb{E}\|\hat{\beta}^{(t+0.5)} - \bar{\beta}^{(t+0.5)}\|_\infty$ is controlled by the following lemma.

Lemma 22. *Under the same conditions as*

$$\mathbb{E}\|\hat{\boldsymbol{\beta}}^{(t+0.5)} - \bar{\boldsymbol{\beta}}^{(t+0.5)}\|_\infty \lesssim \sigma \sqrt{\frac{\log d}{n_0}}, \quad (5.1)$$

Lemma 22 implies that

$$\mathbb{E}\|\hat{\boldsymbol{\beta}}^{(t+1)} - \bar{\boldsymbol{\beta}}^{(t+1)}\|_2 \lesssim \sigma \sqrt{\frac{s \log d}{n_0}} + \frac{T \sqrt{s \log(1/\delta)}}{n_0 \varepsilon} \log d. \quad (5.2)$$

The next step is bounding $\|\bar{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_2$. We begin with introducing constants $\mu = 2\lambda_{\max}(\Sigma_X)$, $\nu = 2\lambda_{\min}(\Sigma_X)$, such that:

$$\nu/2 \cdot \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2 \leq \mathbb{E}_{X,y} \left[(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^\top \frac{1}{n} X^\top X (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \right] \leq \mu/2 \cdot \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2$$

By invoking standard optimization results for minimizing strongly convex and smooth objective functions, e.g., in Nesterov (2004), for stepsize $\eta = 2/(\nu + \mu)$,

$$\|\bar{\boldsymbol{\beta}}^{(t+0.5)} - \boldsymbol{\beta}^*\|_2 \leq \left(\frac{\mu - \nu}{\mu + \nu} \right) \cdot \|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|_2. \quad (5.3)$$

Then we use the Lemma 5.1 from Wang et al. (2014).

Lemma 23. *Suppose that we have*

$$\|\bar{\boldsymbol{\beta}}^{(t+0.5)} - \boldsymbol{\beta}^*\|_2 \leq \kappa \cdot \|\boldsymbol{\beta}^*\|_2,$$

for some $\kappa \in (0, 1)$. Assuming that we have

$$\hat{s} \geq \frac{4(1 + \kappa)^2}{(1 - \kappa)^2} s, \text{ and } \sqrt{\hat{s}} \|\hat{\boldsymbol{\beta}}^{(t+0.5)} - \bar{\boldsymbol{\beta}}^{(t+0.5)}\|_\infty \leq \frac{(1 - \kappa)^2}{2(1 + \kappa)} \cdot \|\boldsymbol{\beta}^*\|_2,$$

then it holds that

$$\|\bar{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_2 \leq \frac{C\sqrt{s}}{\sqrt{1-\kappa}} \|\hat{\boldsymbol{\beta}}^{(t+0.5)} - \bar{\boldsymbol{\beta}}^{(t+0.5)}\|_\infty + (1 + 4\sqrt{s/\hat{s}})^{1/2} \cdot \|\bar{\boldsymbol{\beta}}^{(t+0.5)} - \boldsymbol{\beta}^*\|_2.$$

The lemma enables us to establish the following result:

Lemma 24. For $t = -1, 0, 1, 2, \dots, n_0 - 1$ and $\rho := \frac{\mu-\nu}{\mu+\nu}$, it holds with high probability that

$$\begin{aligned} \|\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_2 &\leq \kappa \|\boldsymbol{\beta}^*\|_2, \\ \|\bar{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_2 &\leq \rho^{t/2} \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_2 + C \left(\sigma \sqrt{\frac{s \log d}{n_0}} + \frac{Ts \sqrt{\log(1/\delta)}}{n_0 \varepsilon} \log d \right). \end{aligned} \quad (5.4)$$

Finally, with $t \asymp \log n$ (namely $n_0 \asymp n/\log n$), $T \asymp \sigma \sqrt{\log n}$, (5.2) and (5.4) together imply the desired upper bound. □

Proof of Lemma 22

Proof.

$$\begin{aligned} &\mathbb{E} \|\hat{\boldsymbol{\beta}}^{(t+0.5)} - \bar{\boldsymbol{\beta}}^{(t+0.5)}\|_\infty \\ &= \frac{1}{n_0} \mathbb{E} \left\| \left(\mathbb{E}_{X,y} \left[X_{(t)}^\top X_{(t)} \hat{\boldsymbol{\beta}}^{(t)} - X_{(t)}^\top \mathbf{y}_{(t)} \right] \right) - \left(X_{(t)}^\top f_T(X_{(t)} \hat{\boldsymbol{\beta}}^{(t)}) - X_{(t)}^\top f_T(\mathbf{y}_{(t)}) \right) \right\|_\infty \\ &\leq \mathbb{E} \left\| \left(\mathbb{E}_{X,y} \left[\frac{1}{n_0} X_{(t)}^\top X_{(t)} \hat{\boldsymbol{\beta}}^{(t)} - \frac{1}{n_0} X_{(t)}^\top \mathbf{y}_{(t)} \right] \right) - \frac{1}{n_0} \left(X_{(t)}^\top X_{(t)} \hat{\boldsymbol{\beta}}^{(t)} - X_{(t)}^\top \mathbf{y}_{(t)} \right) \right\|_\infty \\ &\quad + \mathbb{E} \left\| \frac{1}{n_0} \left(X_{(t)}^\top X_{(t)} \hat{\boldsymbol{\beta}}^{(t)} - X_{(t)}^\top \mathbf{y}_{(t)} \right) - \frac{1}{n_0} \left(X_{(t)}^\top f_T(X_{(t)} \hat{\boldsymbol{\beta}}^{(t)}) - X_{(t)}^\top f_T(\mathbf{y}_{(t)}) \right) \right\|_\infty. \end{aligned}$$

The first term is at most the order of $\sigma\sqrt{\frac{\log d}{n_0}}$, as follows:

$$\begin{aligned}
& \mathbb{E} \left\| \left(\mathbb{E}_{X,y} \left[\frac{1}{n_0} X_{(t)}^\top X_{(t)} \hat{\boldsymbol{\beta}}^{(t)} - \frac{1}{n_0} X_{(t)}^\top \mathbf{y}_{(t)} \right] \right) - \frac{1}{n_0} \left(X_{(t)}^\top X_{(t)} \hat{\boldsymbol{\beta}}^{(t)} - X_{(t)}^\top \mathbf{y}_{(t)} \right) \right\|_\infty \\
& \leq \mathbb{E} \left\| \frac{1}{n_T} X_{(t)}^\top \mathbf{y}_{(t)} - \mathbb{E}_{X,y} \left[\frac{1}{n_0} X_{(t)}^\top \mathbf{y}_{(t)} \right] \right\|_\infty \\
& \quad + \mathbb{E} \left\| \mathbb{E}_{X,y} \left[\frac{1}{n_0} (X_{(t)}^\top X_{(t)}) \hat{\boldsymbol{\beta}}^{(t)} \right] - \frac{1}{n_0} (X_{(t)}^\top X_{(t)}) \hat{\boldsymbol{\beta}}^{(t)} \right\|_\infty \\
& \leq \mathbb{E} \left\| \frac{1}{n_0} X_{(t)}^\top (X_{(t)} \boldsymbol{\beta}^* + \epsilon) - \mathbb{E}_{X,y} \left[\frac{1}{n_0} X_{(t)}^\top (X_{(t)} \boldsymbol{\beta}^* + \epsilon) \right] \right\|_\infty \\
& \quad + \mathbb{E} \left\| \mathbb{E}_{X,y} \left[\frac{1}{n_0} (X_{(t)}^\top X_{(t)}) \hat{\boldsymbol{\beta}}^{(t)} \right] - \frac{1}{n_0} (X_{(t)}^\top X_{(t)}) \hat{\boldsymbol{\beta}}^{(t)} \right\|_\infty \\
& \leq \mathbb{E} \left\| \mathbb{E}_{X,y} \left[\frac{1}{n_0} (X_{(t)}^\top X_{(t)}) \hat{\boldsymbol{\beta}}^{(t)} \right] - \frac{1}{n_0} (X_{(t)}^\top X_{(t)}) \hat{\boldsymbol{\beta}}^{(t)} \right\|_\infty \\
& \quad + \mathbb{E} \left\| \left(\frac{1}{n_0} X_{(t)}^\top X_{(t)} - \mathbb{E} \left[\frac{1}{n_0} X_{(t)}^\top X_{(t)} \right] \right) \boldsymbol{\beta}^* \right\|_\infty + \mathbb{E} \left\| \frac{1}{n_0} X_{(t)}^\top \epsilon - \mathbb{E}_{X,y} \left[\frac{1}{n_0} X_{(t)}^\top \epsilon \right] \right\|_\infty.
\end{aligned}$$

For the first two terms, we note that for a fixed vector $\mathbf{a} \in \mathbb{R}^d$ and X consisting of sub-Gaussian entries, by the standard Bernstein-type inequality for sub-exponential random variables, we have

$$\mathbb{E} \left\| \Sigma_X \mathbf{a} - \frac{1}{n} (X^\top X) \mathbf{a} \right\|_\infty \leq \mathbb{E} \left(\max_j \frac{1}{n} \sum_{i=1}^n z_{ij} z_i^\top \mathbf{a} \right) \lesssim \sigma \sqrt{\frac{\log d}{n}},$$

where z_{ij} denotes the centered version of x_{ij} . Similarly, by our assumption that $\epsilon := y - \mathbf{x}^\top \boldsymbol{\beta}^*$ is subgaussian, the same bound holds for the third term $\mathbb{E} \left\| \frac{1}{n_0} X_{(t)}^\top \epsilon - \mathbb{E}_{X,y} \left[\frac{1}{n_0} X_{(t)}^\top \epsilon \right] \right\|_\infty$ as well. So we obtain

$$\frac{1}{n_0} \mathbb{E} \left\| \left(\mathbb{E}_{X,y} \left[X_{(t)}^\top X_{(t)} \hat{\boldsymbol{\beta}}^{(t)} - X_{(t)}^\top \mathbf{y}_{(t)} \right] \right) - \left(X_{(t)}^\top X_{(t)} \hat{\boldsymbol{\beta}}^{(t)} - X_{(t)}^\top \mathbf{y}_{(t)} \right) \right\|_\infty \lesssim \sigma \sqrt{\frac{\log d}{n_0}}.$$

Finally, it remains to bound

$$\begin{aligned} & \mathbb{E} \left\| \frac{1}{n_0} \left(X_{(t)}^\top X_{(t)} \hat{\boldsymbol{\beta}}^{(t)} - X_{(t)}^\top \mathbf{y}_{(t)} \right) - \frac{1}{n_0} \left(X_{(t)}^\top f_T(X_{(t)} \hat{\boldsymbol{\beta}}^{(t)}) - X_{(t)}^\top f_T(\mathbf{y}_{(t)}) \right) \right\|_\infty \\ & \leq \frac{1}{n_0} \mathbb{E} \| X_{(t)}^\top (X_{(t)} \hat{\boldsymbol{\beta}}^{(t)} - f_T(X_{(t)} \hat{\boldsymbol{\beta}}^{(t)})) \|_\infty + \frac{1}{n_0} \mathbb{E} \| X_{(t)}^\top (\mathbf{y}_{(t)} - f_T(\mathbf{y}_{(t)})) \|_\infty \end{aligned}$$

In general, for an n_0 -dimensional sub-Gaussian(σ) random vector \mathbf{z} satisfying $\|\mathbb{E}\mathbf{z}\|_\infty = O(1)$, since $|x_{ij}| \leq 1$, we have

$$\mathbb{E} \| X_{(t)}^\top (\mathbf{z} - f_T(\mathbf{z})) \|_\infty \leq \sum_{i=1}^{n_0} \mathbb{E} [|z_i| \mathbb{1}(|z_i| > T)].$$

For each $i \in [n_0]$, since $\|\mathbb{E}\mathbf{z}\|_\infty = O(1)$, for sufficiently large constant K ,

$$\mathbb{P}(|z_i| > t) dt \leq K e^{-t^2/(2\sigma^2)}.$$

Hence,

$$\mathbb{E} [z_i \mathbb{1}(|z_i| > T)] \leq \int_T^\infty \mathbb{P}(|z_i| > t) dt \leq \int_T^\infty K e^{-t^2/\sigma^2} dt.$$

At last, by a standard tail estimate of the Gaussian integral, with $T > \sigma\sqrt{2\log n}$,

$$\int_T^\infty K e^{-t^2/\sigma^2} dt \lesssim e^{-T^2/\sigma^2} \lesssim \frac{1}{n_0}.$$

It then follows that

$$\frac{1}{n_0} \mathbb{E} \| X_{(t)}^\top (\mathbf{z} - f_T(\mathbf{z})) \|_\infty \leq \frac{1}{n_0} \sum_{i=1}^{n_0} \mathbb{E} [|z_i| \mathbb{1}(|z_i| > T)] \lesssim \frac{1}{n_0},$$

which is negligible compared to the $\sigma\sqrt{\frac{\log d}{n_0}}$ terms above. □

Proof of Lemma 24

Proof. In this proof, C refers to a numerical constant that may take different values in different contexts.

We prove the lemma by induction. When $t = -1$, the two inequalities trivially hold. In particular, the first inequality is exactly our condition on the initialization of the algorithm.

For the inductive step, we suppose that with high probability,

$$\begin{aligned} \|\hat{\beta}^{(t)} - \beta^*\|_2 &\leq \kappa \|\beta^*\|_2, \\ \|\bar{\beta}^{(t)} - \beta^*\|_2 &\leq \rho^{(t-1)/2} \|\hat{\beta}^{(0)} - \beta^*\|_2 + C \left(\sigma \sqrt{\frac{s \log d}{n_0}} + \frac{T s \sqrt{\log(1/\delta)}}{n_0 \varepsilon} \log d \right). \end{aligned}$$

The goal is to prove these statements with t replaced by $t + 1$.

To apply Lemma 23, we shall first verify that its conditions are satisfied. We have,

$$\|\bar{\beta}^{(t+0.5)} - \beta^*\|_2 \leq \rho \|\hat{\beta}^{(t)} - \beta^*\|_2 \leq \rho \kappa \cdot \|\beta^*\|_2 \leq \kappa \|\beta^*\|_2.$$

The first inequality is due to (5.3); the second equality is due to our inductive hypothesis. The condition on \hat{s} in Lemma 23 is satisfied by our choice of the tuning parameter, while the third condition

$$\sqrt{\hat{s}} \|\hat{\beta}^{(t+0.5)} - \bar{\beta}^{(t+0.5)}\|_\infty \leq \frac{(1 - \kappa)^2}{2(1 + \kappa)} \cdot \|\beta^*\|_2$$

holds with high probability, thanks to Lemma 22.

Then, we use Lemma 23 and (5.3) to obtain that

$$\begin{aligned}
& \|\bar{\beta}^{(t+1)} - \beta^*\|_2 \\
& \leq \frac{C\sqrt{s}}{\sqrt{1-\kappa}} \|\hat{\beta}^{(t+0.5)} - \bar{\beta}^{(t+0.5)}\|_\infty + (1 + 4\sqrt{s/\hat{s}})^{1/2} \cdot \|\bar{\beta}^{(t+0.5)} - \beta^*\|_2 \\
& \leq C \left(\sigma \sqrt{\frac{s \log d}{n_0}} + \frac{Ts\sqrt{\log(1/\delta)}}{(n_0)\varepsilon} \log d \right) + (1 + 4\sqrt{s/\hat{s}})^{1/2} \cdot \rho \|\hat{\beta}^{(t)} - \beta^*\|_2,
\end{aligned}$$

where the last inequality makes use of (5.1) and (5.3).

When $(1 + 4\sqrt{s/\hat{s}})^{1/2} < \rho^{-1/2}$, that is, $\hat{s} > (\frac{4\rho}{1-\rho})^2 s$, we have

$$\|\bar{\beta}^{(t+1)} - \beta^*\|_2 \leq \rho^{1/2} \|\hat{\beta}^{(t)} - \beta^*\|_2 + C \left(\sigma \sqrt{\frac{s \log d}{n_0}} + \frac{Ts\sqrt{\log(1/\delta)}}{n_0\varepsilon} \log d \right).$$

It follows that

$$\begin{aligned}
& \|\bar{\beta}^{(t+1)} - \beta^*\|_2 \\
& \leq \rho^{1/2} \|\hat{\beta}^{(t)} - \beta^*\|_2 + C \left(\sigma \sqrt{\frac{s \log d}{n_0}} + \frac{Ts\sqrt{\log(1/\delta)}}{n_0\varepsilon} \log d \right) \\
& \leq \rho^{1/2} \left(\|\hat{\beta}^{(t)} - \bar{\beta}^{(t)}\|_2 + \|\bar{\beta}^{(t)} - \beta^*\|_2 \right) + C \left(\sigma \sqrt{\frac{s \log d}{n_0}} + \frac{Ts\sqrt{\log(1/\delta)}}{n_0\varepsilon} \log d \right) \\
& \leq \rho^{t/2} \|\hat{\beta}^{(0)} - \beta^*\|_2 + C \left(\sigma \sqrt{\frac{s \log d}{n_0}} + \frac{Ts\sqrt{\log(1/\delta)}}{n_0\varepsilon} \log d \right).
\end{aligned}$$

The last inequality is a consequence of the inductive hypothesis and (5.2). To complete the induction, the inequality above and (5.2) imply that

$$\begin{aligned}
\|\hat{\beta}^{(t+1)} - \beta^*\|_2 & \leq \mathbb{E} \|\hat{\beta}^{(t+1)} - \bar{\beta}^{(t+1)}\|_2 + \|\bar{\beta}^{(t+1)} - \beta^*\|_2 \\
& \leq \rho^{t/2} \|\hat{\beta}^{(0)} - \beta^*\|_2 + C \left(\sigma \sqrt{\frac{s \log d}{n_0}} + \frac{Ts\sqrt{\log(1/\delta)}}{n_0\varepsilon} \log d \right) \\
& \leq \rho^{t/2} \kappa \|\beta^*\|_2 + C \left(\sigma \sqrt{\frac{s \log d}{n_0}} + \frac{Ts\sqrt{\log(1/\delta)}}{n_0\varepsilon} \log d \right) \\
& \leq \kappa \|\beta^*\|_2.
\end{aligned}$$



APPENDIX

Due to the limit of the space, we don't include the supplements in this thesis. Please refer the supplements of Chapter 2-5 to Cai and Zhang (2018d); Cai et al. (2016a, 2018b, 2019c) respectively.

BIBLIOGRAPHY

- M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.
- J. M. Abowd. The challenge of scientific reproducibility and privacy protection for statistical agencies. In *Census Scientific Advisory Committee*, 2016.
- T. W. Anderson. *An Introduction To Multivariate Statistical Analysis*. Wiley-Interscience, 3rd ed, New York, 2003.
- M. Azizyan, A. Singh, and L. Wasserman. Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems*, pages 2139–2147, 2013.
- M. Azizyan, A. Singh, and L. A. Wasserman. Efficient sparse clustering of high-dimensional non-spherical Gaussian mixtures. In *AISTATS*, 2015.
- M. Bafna and J. Ullman. The price of selection in differential privacy. *arXiv preprint arXiv:1702.02970*, 2017.
- S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Ann. Statist.*, 45(1):77–120, 2017.
- S. Bandyopadhyay, M. Mehta, D. Kuo, M.-K. Sung, R. Chuang, E. J. Jaehnig, B. Bodenmiller, K. Licon, W. Copeland, and M. Shales. Rewiring of genetic networks in response to dna damage. *Science*, 330(6009):1385–1389, 2010.
- R. F. Barber and J. C. Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014.
- R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 464–473. IEEE, 2014.
- M. W. Berry and M. Castellanos. Survey of text mining. *Computing Reviews*, 45(9):548, 2004.
- P. J. Bickel and E. Levina. Some theory for Fisher’s linear discriminant function, ‘naive Bayes’, and some alternatives when there are many more variables than observations. *Bernoulli*, 10:989–1010, 2004.
- P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36(6): 2577–2604, 2008.
- C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

- D. Boneh and J. Shaw. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44(5):1897–1905, 1998.
- C. Bouveyron and C. Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78, 2014.
- P. S. Bradley, U. M. Fayyad, and O. L. Mangasarian. Mathematical programming for data mining: Formulations and challenges. *INFORMS Journal on Computing*, 11(3):217–238, 1999.
- M. Bun, J. Ullman, and S. Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 1–10. ACM, 2014.
- T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.
- T. T. Cai and W. Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577, 2011.
- T. T. Cai and M. Yuan. Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics*, 40(4):2014–2042, 2012.
- T. T. Cai and A. Zhang. Estimation of high-dimensional covariance matrices with incomplete data. *Journal of Multivariate Analysis*, 150:55–74, 2016.
- T. T. Cai and L. Zhang. High-dimensional Gaussian copula regression: Adaptive estimation and statistical inference. *Statistica Sinica*, to appear, 2017.
- T. T. Cai and L. Zhang. High-dimensional linear discriminant analysis: Optimality, adaptive algorithm, and missing data. *arXiv preprint arXiv:1804.03018*, 2018a.
- T. T. Cai and L. Zhang. A convex optimization approach to high-dimensional sparse quadratic discriminant analysis. 2018b.
- T. T. Cai and L. Zhang. High-dimensional linear discriminant analysis: Optimality, adaptive algorithm, and missing data. *arXiv preprint arXiv:1804.03018*, 2018c.
- T. T. Cai and L. Zhang. Supplement to “High-dimensional linear discriminant analysis: Optimality, adaptive algorithm, and missing data”. 2018d.
- T. T. Cai, T. Liang, and H. H. Zhou. Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional gaussian distributions. *Journal of Multivariate Analysis*, 137:161–172, 2015.
- T. T. Cai, J. Ma, and L. Zhang. Supplement to “a convex optimization approach to high-dimensional sparse quadratic discriminant analysis”. 2016a.

- T. T. Cai, Z. Ren, H. H. Zhou, et al. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59, 2016b.
- T. T. Cai, J. Ma, and L. Zhang. CHIME: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, to appear, 2018a.
- T. T. Cai, J. Ma, and L. Zhang. Supplement to “CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality”. 2018b.
- T. T. Cai, J. Ma, L. Zhang, et al. CHIME: Clustering of high-dimensional Gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, 2019a.
- T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*, 2019b.
- T. T. Cai, Y. Wang, and L. Zhang. Supplement to “the cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy”. 2019c.
- E. Candes and J. Romberg. 11-magic: Recovery of sparse signals via convex programming. 2005.
- R. Y. Chen, A. Gittens, and J. A. Tropp. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference: A Journal of the IMA*, 1(1):2–20, 2012.
- G. A. Churchill and D. Iacobucci. *Marketing research: methodological foundations*. Dryden Press New York, 2006.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–38, 1977.
- A. Differential Privacy Team. Privacy at scale. Apple, 2017.
- B. Ding, J. Kulkarni, and S. Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pages 3571–3580, 2017.
- J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 429–438. IEEE, 2013.
- J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.
- R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.

- C. Dwork and V. Feldman. Privacy-preserving prediction. *arXiv preprint arXiv:1803.10266*, 2018.
- C. Dwork and G. N. Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.
- C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.
- C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014a.
- C. Dwork, K. Talwar, A. Thakurta, and L. Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20. ACM, 2014b.
- C. Dwork, A. Smith, T. Steinke, J. Ullman, and S. Vadhan. Robust traceability from trace amounts. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 650–669. IEEE, 2015.
- C. Dwork, A. Smith, T. Steinke, and J. Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4:61–84, 2017.
- C. Dwork, W. J. Su, and L. Zhang. Differentially private false discovery rate control. *arXiv preprint arXiv:1807.04209*, 2018.
- Ú. Erlingsson, V. Pihur, and A. Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.
- B. S. Everitt. *Finite mixture distributions*. Wiley Online Library, 1981.
- J. Fan and Y. Fan. High dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36(6):2605–2637, 2008.
- C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.
- J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.
- R. Ge, Q. Huang, and S. M. Kakade. Learning mixtures of Gaussians in high dimensions. pages 761–770, 2015.
- J. W. Graham. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60:549–576, 2009.

- F. Han and H. Liu. Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. *Bernoulli*, 23(1):23, 2017.
- F. Han, T. Zhao, and H. Liu. Coda: High dimensional copula discriminant analysis. *Journal of Machine Learning Research*, 14(Feb):629–671, 2013.
- M. Hardt and E. Price. Sharp bounds for learning a mixture of two Gaussians. *ArXiv e-prints*, 1404, 2014.
- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning 2nd edition*. New York: Springer, 2009.
- S. M. Hill, L. M. Heiser, T. Cokelaer, M. Unger, N. K. Nesser, D. E. Carlin, Y. Zhang, A. Sokolov, E. O. Paull, and C. K. Wong. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature methods*, 13(4):310–318, 2016.
- N. Homer, S. Szlinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.
- A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.
- B. Jiang, X. Wang, and C. Leng. QUDA: A direct approach for sparse quadratic discriminant analysis. *arXiv preprint arXiv:1510.00084*, 2015.
- C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *NIPS*, pages 4116–4124, 2016a.
- J. Jin, W. Wang, et al. Influential features pca for high dimensional clustering. *Ann. Stat.*, 44(6):2323–2359, 2016b.
- J. Jin, Z. T. Ke, and W. Wang. Phase transitions for high dimensional clustering and related problems. *Ann. Stat.*, to appear, 2017.
- I. M. Johnstone. On minimax estimation of a sparse normal mean vector. *The Annals of Statistics*, 22:271–289, 1994.
- T. Jombart, S. Devillard, and F. Balloux. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics*, 11(1):94, 2010.
- E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2006.

- J. Lei. Differentially private m-estimators. In *Advances in Neural Information Processing Systems*, pages 361–369, 2011.
- E. Levina and R. Vershynin. Partial estimation of covariance matrices. *Probability Theory and Related Fields*, 153(3):405–419, 2012.
- Q. Li and J. Shao. Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica*, pages 457–473, 2015.
- T. Li, X. Yi, C. Carmanis, and P. Ravikumar. Minimax gaussian classification & clustering. In *Artificial Intelligence and Statistics*, pages 1–9, 2017.
- M. W. Libbrecht and W. S. Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015.
- Y. Lin and Y. Jeon. Discriminant analysis through a semiparametric model. *Biometrika*, pages 379–392, 2003.
- B. G. Lindsay. Mixture models: theory, geometry, and applications, volume 5 of nsf-cbms regional conference series in probability and statistics. *Institute for Mathematical Statistics: Hayward, CA*, 1995.
- H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.
- P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *The Annals of Statistics*, 40:1637–1664, 2012.
- K. Lounici. Sparse principal component analysis with missing observations. *High dimensional probability VI*, 66 of Progress in Probability, IMS Collections:327–356, 2013.
- K. Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.
- C. Lowther, D. Merico, G. Costain, J. Wasserman, K. Boyd, A. Noor, M. Speevak, D. J. Stavropoulos, J. Wei, A. C. Lionel, et al. Impact of IQ on the diagnostic yield of chromosomal microarray in a community sample of adults with schizophrenia. *Genome medicine*, 9(1):105, 2017.
- Q. Mai and H. Zou. Sparse semiparametric discriminant analysis. *Journal of Multivariate Analysis*, 135:175–188, 2015.
- Q. Mai, H. Zou, and M. Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42, 2012.
- Q. Mai, Y. Yang, and H. Zou. Multiclass sparse discriminant analysis. *arXiv preprint arXiv:1504.05845*, 2015.

- W. Miao, P. Ding, and Z. Geng. Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111(516): 1673–1683, 2016.
- I. Mironov. Renyi differential privacy. In *Computer Security Foundations Symposium (CSF), 2017 IEEE 30th*, pages 263–275. IEEE, 2017.
- A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of gaussians. In *FOCS, 2010 51st Annual IEEE Symposium on Theory of Computing*, pages 93–102. IEEE, 2010.
- M. Neykov, Y. Ning, J. S. Liu, and H. Liu. A unified theory of confidence regions and testing for high dimensional estimating equations. *arXiv preprint arXiv:1510.08986*, 2015.
- R. K. Pace and R. Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.
- K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.
- M. Raab and A. Steger. Balls into bins – a simple and tight analysis. In *International Workshop on Randomization and Approximation Techniques in Computer Science*, pages 159–170. Springer, 1998.
- M. Rao, T. Javidi, Y. C. Eldar, and A. Goldsmith. Fundamental estimation limits in autoregressive processes with compressive measurements. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 2895–2899. IEEE, 2017.
- G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of convergence for high-dimensional regression under ℓ_q -ball sparsity. In *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, pages 251–257. IEEE, 2009.
- R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.
- D. Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, pages 827–832, 2015.
- J. M. Robins and Y. Ritov. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319, 1997.
- A. Rohde and L. Steinberger. Geometrizing rates of convergence under differential privacy constraints. *arXiv preprint arXiv:1805.01422*, 2018.
- A. Rotnitzky and J. Robins. Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in Medicine*, 16(1):81–102, 1997.

- T. Schneider. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, 14(5):853–871, 2001.
- A. J. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27(2):387–397, 1971.
- J. Shao, Y. Wang, X. Deng, and S. Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, 39(2):1241–1265, 2011.
- A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 813–822. ACM, 2011.
- T. Steinke and J. Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7(2), 2017.
- B. Sun, L. Liu, W. Miao, K. Wirth, J. Robins, and E. T. Tchetgen. Semiparametric estimation with data missing not at random using an instrumental variable. *arXiv preprint arXiv:1607.03197*, 2016.
- K. Talwar, A. G. Thakurta, and L. Zhang. Nearly optimal private lasso. In *Advances in Neural Information Processing Systems*, pages 3025–3033, 2015.
- G. Tardos. Optimal probabilistic fingerprint codes. *Journal of the ACM (JACM)*, 55(2):10, 2008.
- E. J. Tchetgen Tchetgen and K. E. Wirth. A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics*, 73(4):1123–1131, 2017.
- L. Tian and Q. Gu. Communication-efficient distributed sparse linear discriminant analysis. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pages 1178–1187. PMLR, 20–22 Apr 2017.
- R. Tibshirani and G. Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005.
- R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.
- A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.
- J. Ullman. Answering $n^2 + o(1)$ counting queries with differential privacy is hard. *SIAM Journal on Computing*, 45(2):473–496, 2016.

- R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, and J. P. Mesirov. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *pdgfra*, *idh1*, *egfr*, and *nf1*. *Cancer Cell*, 17(1):98–110, 2010.
- Y.-X. Wang, S. Fienberg, and A. Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pages 2493–2502, 2015.
- Z. Wang, Q. Gu, Y. Ning, and H. Liu. High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*, 2014.
- J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.
- L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.
- I. R. White, P. Royston, and A. M. Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011.
- D. M. Witten and R. Tibshirani. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636, 2009.
- D. M. Witten and R. Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.
- J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2009.
- M. C. Wu, L. Zhang, Z. Wang, D. C. Christiani, and X. Lin. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25(9):1145–1151, 2009.
- Y. Xia, T. Cai, and T. T. Cai. Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika*, 102(2):247–266, 2015.
- M. Xu, D. Zhang, and W. B. Wu. l^2 asymptotics for high-dimensional data. *arXiv preprint arXiv:1405.7244*, 2014.
- L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 2012.
- F. Ye and C.-H. Zhang. Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, 11(Dec):3519–3540, 2010.

- X. Yi and C. Caramanis. Regularized em algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2015.
- S. D. Zhao, T. T. Cai, and H. Li. Direct estimation of differential networks. *Biometrika*, 101(2):253–268, 2014.
- H. Zhou, W. Pan, and X. Shen. Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics*, 3:1473–1496, 2009.