ON THE LASSO REGRESSION AND ASYMMETRIC LAPLACE DISTRIBUTION WITH APPLICATIONS

YUE SHI

PhD

The Hong Kong Polytechnic University

2018

The Hong Kong Polytechnic University

Department of Applied Mathematics

On the LASSO regression and Asymmetric Laplace Distribution with applications

Yue Shi

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

March 2018

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

____________________(Signed)

__________Yue SHI________(Name of student)

Dedicated to my mother.

Abstract

In this thesis, we consider four classes of optimization models. One class is LAD Generalized Lasso models. We develop a descent algorithm for LAD-Lasso and a new active zero set descent algorithm for LAD Generalized Lasso under nonsmooth optimality conditions; The second class is constrained LAD Lasso models. We extend the descent algorithm to tackle the constraints as well. Application in Mean Absolute Deviation Lasso portfolio selection is studied. The third class is selection of penalty parameter for compressive sensing. We carry out tests using several criteria for selection of the penalty parameter. The fourth class is optimization under Asymmetric Laplace Distributions, namely robust mixture linear regression model and portfolio selection.

We first consider LAD Generalized Lasso models. Under dynamic nonsmooth optimality conditions, we develop a descent algorithm by selecting fastest descent directions for LAD-Lasso regression. Then we derive a new active zero set descent algorithm for LAD Generalized Lasso regression. The algorithm updates the zero set and basis search directions recursively until optimality conditions are satisfied. It is also shown that the proposed algorithm converges in finitely many steps.

We then consider Constrained LAD Lasso models. We develop a descent algorithm by updating descent directions selected from basis directional set for nonsmooth optimization problems for MAD-Lasso portfolio selection strategy, extensive real data analysis are provided to evaluate the out-of-sample performances.

We next consider selection of penalty parameter. For compressive sensing based signal recovery model, we apply regularized Least Squares for sparse reconstruction since it can reconstruct speech signal from a noisy observation, and proposed a two-level optimization strategy to incorporate the quality design attributes in the sparse solution in compressive speech enhancement by hyper-parameterizing the tuning parameter. The first level involves the compression of the big data and the second level optimizes the tuning parameter by using different optimization criteria (such as Gini index, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)). The set of solutions can then be measured against the desired design attributes to achieve the best trade-off between suppression and distortion.

Finally, we study two models under Asymmetric Laplace Distributions. We first present an efficient two-level latent EM algorithm for parameter estimation of mixture linear regression models, with group label as the first level latent variable and laplace intermediate variable as the second level latent variable. Explicit updating formula of each iteration are derived and computational complexity can thus be reduced significantly. Then we consider robust portfolio selection model, and derived the Expectation-Maximization (EM) algorithm for parameter estimation of Asymmetric Laplace distribution, efficient frontier analysis is provided to evaluate the performance.

Publications Arising from the Thesis

- Y. Shi, Z. Feng, and K.F.C. Yiu. A descent method for least absolute deviation lasso problems, *Optimization Letters*, 1-17, 2017.

- Y. Shi, S.Y. Low, K.F.C. Yiu. Hyper-parameterization of sparse reconstruction for speech enhancement. *Applied Acoustics*, 138, 72-79, 2018.

- Y. Shi, K.F.C. Yiu, Portfolio selection based on Asymmetric Laplace distribution, coherent risk measure, and expectation-maximization estimation, *Quantitative Finance and Economics, accepted*, 2018.

- Y. Shi, C.T. Ng, and K.F.C. Yiu. New active zero set descent algorithm for LAD Generalized Lasso models, *submitted*, 2017.

- Y. Shi, Z. Feng, and K.F.C. Yiu. A descent algorithm for constrained LAD Lasso estimation with applications in portfolio selection, *submitted*, 2017.

- Y. Shi, S.S. Wang, and K.F.C. Yiu. Robust mixture regression model via Asymmetric Laplace distribution, *submitted*, 2018.

Acknowledgements

First and foremost, I would thank my chief supervisor, Prof. Yiu Ka-fai, Cedric deeply for his intensive guidance and encouragement during my PhD study. I learnt how to study and and appreciate the character of Oxford University from him, and affected by his intelligence and integrity greatly. Then I would owe my deepest thanks to Dr. Chun-ling Liu, Catherine for her successive encouragement, love and trust, I enjoy her personal charisma of goodness, diligence and optimism. Additionally, I would owe my sincere gratitude to the committee members: Dr. Xingqiu Zhao, Prof. Shengjie Li and Dr. Siu Pang Yung. Thanks very much for offering me constructive suggestions.

Furthermore, I would thank all my collaborators. I would firstly express my sincere gratitude to Prof. Zhiguo Feng for his successive assistance and guidance in developing optimization algorithms, then I would thank Prof. Chi Tim Ng for providing me careful inspiration of statistical mind and knowledge. Also, I would thank Prof. Siow Yong Low, who has helped me a lot in dealing with signal processing tools. Thanks very much for their great patience and encouragement such that I can finish my thesis smoothly.

I'm very grateful to join Professor Yiu's group, and thanks very much for all my academic brothers and sisters: Dr. Jingzhen Liu, Dr. Mingjie Gao, Dr. Zhibao Li, Ms. Yu Bai, Mr. Tsz Pang Yuen, Mr. Zikai Wei, Ms. Qian Liu, Mr. Qingzheng Wang, for their companion. Meanwhile, I would also thank my former and current officemates, Dr. Lei Yang, Dr. Jin Yang, Dr. Yang Zhou, Dr. Hong Wang, Ms. Qianqian Hou, Dr. Qiujin Peng, Dr. Meiling Hao, Dr. Xiaodong Yan, Dr. Danlin Hou, Dr. Lei Cui, Dr. Zhilong Dong, Dr. Bo Wen, Mr. Yun Shi, Ms. Fei Fang, Dr. Huili Zhang, Ms. Yangchen Ou, Ms. Kaihui Liu, Ms. Jingya Chang, Mr. Dayu Sun, Ms. Qianying Lin, Ms. Peiran Yu, Mr. Chendi Wang, Ms. Chenchen Zu. Thanks very much for their successive companion and help.

Moreover, my sincere acknowledgements goes to Prof. Jiao Jin and Prof. Xingwei Tong in Beijing Normal University. Thanks very much for offering me the opportunity to pursue my PhD degree in The Hong Kong Polytechnic University, I would also owe my gratitude to Dr. Baosheng Liang, Dr. Shanshan Wang for their help and support.

Finally, I would owe my deepest gratitude to my mother for her continuous support and encouragement during my study. Without her help, I can not engaged in work with all my strength and enthusiasm.

Yue SHI
The Hong Kong Polytechnic University
March 2018

Contents

Certificate of Originality **v**

Abstract **ix**

Acknowledgements **xiii**

List of Figures **xvii**

List of Tables **xix**

List of Notations **xxi**

1 Introduction **1**

1.1 LAD Generalized Lasso models 2

1.2 Constrained LAD Lasso models 6

1.3 Selection of penalty parameter 8

1.4 Two models under Asymmetric Laplace Distributions 10

1.5 Contributions of the thesis 14

1.6 Organizations of the thesis 16

2 LAD Generalized Lasso Models **19**

2.1 A descent method for LAD-Lasso model 19

2.1.1 Computational methodology 21

2.1.2 Simulation studies 33

2.1.3 Real data analysis 36

2.2 New active zero set descent algorithm for LAD Generalized Lasso . . 45

2.2.1 Computational method 51

2.2.2 Numerical experiments 60

3 Constrained LAD Lasso models 77

3.1 MAD-Lasso model 77

3.2 Constrained LAD Lasso model 80

3.3 Numerical experiments 88

4 Signal Processing 95

4.1 Signal model background 95

4.2 Proposed two-level optimization process 104

4.3 Numerical experiments 108

5 Two Models under Asymmetric Laplace Distributions 115

5.1 Mixture linear regression under ALDs 115

5.1.1 Methodology 116

5.1.2 Numerical experiments 128

5.2 Portfolio selection under ALD framework 137

5.2.1 Portfolio selection under ALD framework 142

5.2.2 Numerical experiments 151

6 Concluding Remarks 161

6.1 Summary 161

6.2 Further research 163

Bibliography 165

List of Figures

2.1 p=10,50,100,500 n–time plot 34

2.2 Heng Seng AH premium index 38

2.3 Heng Seng AH Index performance 39

2.4 Three cases of descent direction 51

2.5 plot of step length t 57

2.6 Plot of stepwise descent direction of $d^{(k)}$. 58

2.7 Running time tendency of β^* estimation with $p = 50, 100\,, \sigma = 1, 3\,, \rho = 0, 0.1, 0.5$. 69

2.8 Running time tendency of β^{**} estimation with $p = 50, 100\,, \sigma = 1, 3\,, \rho = 0, 0.1, 0.5$. 70

2.9 $p = 50\,, n = 5000$ estimation results of interior point method, proposed method and LS-fuse. 71

2.10 $p = 100\,, n = 5000$ estimation results of interior point method, proposed method and LS-fuse. 72

2.11 Plot of soybean data. 73

2.12 Plot of estimation results of soybean data with LAD-IP, LAD-AZSD and LS-fuse. 74

3.1 Portfolio selection results of Gaussian data A. MAD model; B. MAD-Lasso with interior point method (IP); C. MAD-Lasso with proposed method (NEW). 90

3.2 Portfolio selection results of Asymmetric Laplace data: A. MAD model; B. MAD-Lasso with interior point method (IP); C. MAD-Lasso with proposed method (NEW). 90

3.3 Portfolio selection tendency of FF48d data with increasing tuning parameter. 94

3.4 Portfolio selection tendency of FF48m data with increasing tuning parameter. 94

4.1 *The Gini coefficients for different frequency bins.* 100

4.2 The (a) PESQ and (b) segmental SNR of the different hyperparameter optimization methods as a function of SNRs for babble noise. 110

4.3 The (a) PESQ and (b) segmental SNR of the different hyperparameter optimization methods as a function of SNRs for car noise. 111

4.4 The (a) PESQ and (b) segmental SNR of the different hyperparameter optimization methods as a function of SNRs for subway noise. 112

4.5 The (a) PESQ and (b) segmental SNR of the different hyperparameter optimization methods as a function of SNRs for destroyer noise. . . . 113

5.1 (a) scatter plot of tonedata; (b) histogram of perceived tune ratio. . . 135

5.2 Outlier cases of tone perception data 136

5.3 Univariate densities 138

5.4 Bivariate Asymmetric Laplace data with μ cases: (a1, a2, a3): $\mu = (0,0)$; (b1, b2, b3): $\mu = (1,1)$; (c1, c2, c3): $\mu = (-1,-1)$. Covariance matrix $\mathbf{\Sigma}$ cases. (a1, b1, c1): $\sigma_{11} = \sigma_{22} = 1, \sigma_{12} = \sigma_{21} = 0$; (a2, b2, c2): $\sigma_{11} = \sigma_{22} = 1\,, \sigma_{12} = \sigma_{21} = 0.8$; (a3, b3, c3): $\sigma_{11} = \sigma_{22} = 1\,, \sigma_{12} = \sigma_{21} = -0.8$. 140

5.5 Efficient frontiers of simulated Gaussian data of case (1)-(3) using Gaussian and EM-AL model. 153

5.6 Efficient frontiers of simulated Asymmetric Laplace data of case (1)-(3) using Gaussian and EM-AL model. 154

5.7 ALD efficient frontier of Hang Seng data 157

5.8 Skewness, Kurtosis and Sharpe ratio tendency of Hang Seng data . . 157

5.9 ALD efficient frontier of Hang Seng data 160

5.10 Skewness, Kurtosis and Sharpe ratio tendency of Nasdaq data 160

List of Tables

2.1 Simulation results of $p = 10$. 35

2.2 Simulation results of $p = 50$. 35

2.3 Simulation results of $p = 100$. 35

2.4 Simulation results of $p = 500$. 36

2.5 Time comparison for real datasets 37

2.6 AH Stock Index 40

2.7 The details about AH Stock Index 41

2.8 AH Stock 2012-2013 43

2.9 AH Stock 2012-2013 44

2.10 Estimation results of $\beta^*, \sigma = 1$ 65

2.11 Estimation results of $\beta^*, \sigma = 3$ 66

2.12 Estimation results of $\beta^{**}, \sigma = 1$ 67

2.13 Estimation results of $\beta^{**}, \sigma = 3$ 68

2.14 Estimation results under heavy tailed distributions for $p = 50, 100$. . . 71

2.15 Quantitative comparison of soybean data of 3 methods: LAD-IP, LAD-AZSD and LS-fuse models. 74

2.16 Parameter estimation results of soybean data with LAD-IP, LAD-AZSD, LS-fuse. 75

3.1 Portfolio selection results of simulated datasets. 89

3.2 Portfolio selection results of FF48d Data:$(T, n) = (7573, 48)$; ExpRet: $r_0 = 0.0550$. 92

3.3 Portfolio selection results of FF48m data: $(T, n) = (361, 48)$; ExpRet: $r_0 = 1.2606$. 93

4.1 The Gini coefficients for speech and different types of noise and at different SNRs. 101

5.1 Simulation results of $n = 200, s = 0.2$, SNR=4:1 131

5.2 Simulation results of $n = 200, s = 0.4$, SNR=2:1 132

5.3 Simulation results of $n = 400, s = 0.2$, SNR=4:1 133

5.4 Simulation results of $n = 400, s = 0.4$, SNR=2:1 134

5.5 Model fitting results of Gaussian data and Asymmetric Laplace data using Gauss model and EM-AL model. 153

5.6 Hang Seng data statistics 156

5.7 Efficient frontier results of Hang Seng data 156

5.8 Nasdaq data statistics 158

5.9 Efficient frontier analysis of Nasdaq data 159

List of Notations

$\mathbb{R}$	set of real numbers
$\mathbb{R}^p$	set of p-dimensional real vectors
$\mathbb{R}^{n \times p}$	set of $n \times p$ real matrices
$x^\intercal$	transpose of matrix/vector x
∇_d	directional derivative along direction d
$\lvert x \rvert$	absolute value of real number x
$\lVert x \rVert_1$	l_1 norm of vector x
$\lVert x \rVert$	l_2 norm of vector x
$\operatorname{rank}(X)$	rank of matrix X
$N(0,1)$	standard normal distribution
$t(3)$	student t distribution with 3 degree of freedom
$\text{Cauchy}(0,1)$	standard cauchy distribution
$\Phi^{-1}(\alpha)$	α-th quantile of standard normal distribution
$Y \mid X$	conditional Y given X

Chapter 1

Introduction

In this thesis, we study four related topics about regression and portfolio optimization. The first one is LAD Generalized Lasso models, which arise in a wide range of applications, such as image processing, econometrics, engineering and bioinformatics. We study properties of this kind of models and we develop a descent method for the simple LAD-Lasso problem. Similarly, we develop an active zero set descent algorithm for the LAD Generalized Lasso problem.

The second one is Constrained LAD Lasso models. Under nonsmooth optimality conditions, we derived a descent algorithm for Constrained LAD Lasso problem. Then we investigate the MAD-Lasso strategy by combining MAD portfolio selection model with Lasso penalty, and applied the proposed descent algorithm for finding optimal portfolios. This model can induce sparsity and robustness for portfolio selection, meanwhile the proposed algorithm speed up the calculation process.

The third one is the selection of penalty parameter, and we propose a two-level optimization strategy to incorporate the affective design attributes in the sparse solution in compressive speech enhancement by hyper-parameterizing the penalty parameter. Also, we systematically analyze measures such as GINI, AIC and BIC for finding optimal parameters.

Finally, we investigate two models under Asymmetric Laplace distributions which possess tail-heaviness, skewness and peakedness. Then we derive EM algorithms for robust mixture linear regression models and portfolio selection models, complemented with real data analysis to evaluate the performance of our models.

1.1 LAD Generalized Lasso models

At an era of information explosion, the extraction of useful information from massive datasets becomes an important issue. The process often involves selecting a subset of variables to explain certain observations and phenomena. It can be posed as a regression problem. Since the number of variables are not known in advance, a large dataset is often deployed in the selection process in order not to miss the key variables. In this way, the regression problem becomes a sparse fitting problem.

Motivated by the non-negative garrote procedure of Breiman in [14], Tibshirani added sparsity into regression problems in [109] and constructed the Least Absolute Shrinkage and Selection Operator (Lasso) penalty. By adding a bound to the absolute sum of coefficients, Lasso could shrink some coefficients to zeroes and retain significant variables to maintain model interpretability. As a convex penalty, Lasso is solvable and flexible. Hastie et al. systematically summarized a series of Lasso problems in [46], and displayed that Lasso could be extended to generalized linear models and multivariate analysis. The comprehensive advantages made Lasso popular and active in engineering, finance, marketing, bioinformatics and other related fields.

In practical applications, cases with heavy-tailed errors contain outliers are ubiquitous and would deteriorate estimation accuracy significantly. As an alternative to ordinary least square regression, Least Absolute Deviation (LAD) regression maintains robustness against fat tailed errors or extreme outliers due to its connection with L_1 norm and double exponential distribution. There are several approaches combining LAD regression with certain penalty terms for variable selection problems.

Recently, many researchers concerned about LAD regression with variable selection problem. For example, Zeebari united the LAD regression with ridge penalty, and alleviated the multi-collinearity between variables in [130]. Wang et al. proposed a consistent tuning parameter selection technique for LAD-Lasso, and extensively studied the relative asymptotic properties in [113]. In [39], Gao studied the high dimensional LAD-Lasso problem systematically, and confirmed the corresponding asymptotic properties. In [5], Arslan introduced the weighted LAD-Lasso by adaptively adding up a weighting process to mitigate the influence of outliers against both explanatory variables and response variable. In [120], Xu introduced a two-stage method for tuning parameter selection and obtained the oracle property. Various LAD-Lasso related studies have been conducted and the corresponding theoretical properties are well constructed.

Since LAD-Lasso is more robust and could be easily extended to other situations, efficient solution to this problem become imperative and necessary. Generally, LAD-Lasso could be transformed to classical linear programming problem so that they could be computed easily. As an alternative to simplex method, Koender proposed the interior point method with a preprocessing step in [93]. Watson and Yiu [118] dealt with the error-in-variable l_1 norm regression using Levenberg-Marquardt method, and robust solutions are obtained accordingly. Yiu et al. [127] applied l_1-norm to beamforming design and proposed an algorithm with a set of adaptive grids to speed up the calculation process. However, existing algorithms for solving LAD-Lasso is restrictive and rely heavily on the linear programming solvers. We study and propose a more efficient method by selecting a sequence of fastest descent directions based on dynamic optimality condition.

Robust regression analysis with L_1 norm is ubiquitous in many fields of mathematics and engineering that finds a vast amount of applications in econometrics, genetics, meteorology, engineering, and molecular biology, see for example, [10, 36, 51, 60, 90, 95, 122]. To allow sparse structure in the solution, a variety of penalty functions are adopted in the literature, including Lasso [109], Adaptive Lasso [136], Fused Lasso [110], MCP [131], and SCAD [33]. Lasso-type penalties are popular in practice for their flexibility and simplicity. Moreover, Lasso can be used in the majorization step as approximation to the original penalty function in the majorization-minimization procedure, see [53].

To allow sparsity in the solution, Tibshirani and Taylor [111] imposes structural constraints on the coefficients in a linear regression and studies the following Generalized Lasso problem,

$$\arg\min_{\theta \in \mathbb{R}^p} \|Y - X\theta\|^2 + \lambda\|R\theta\|_1, \tag{1.1}$$

where $X \in \mathbb{R}^{n \times p}$ is the design matrix, $Y \in \mathbb{R}^n$ is the response variable, $R \in \mathbb{R}^{q \times p}$ is a specified penalty matrix, $\theta \in \mathbb{R}^p$ is the coefficient vector we are concerned. Though there are various existing work on Generalized Lasso models using sum of squares objective function. It is widely accepted that Least Absolute Deviation (LAD) regression is robust and resistant to heavy tailed outliers in the response. Combining LAD with Generalized Lasso gives LAD Generalized Lasso problem,

$$\arg\min_{\theta \in \mathbb{R}^p} \|Y - X\theta\|_1 + \lambda\|R\theta\|_1. \tag{1.2}$$

When $R = I_p$, (1.2) reduces to the traditional LAD-Lasso problem, see [39, 113, 116, 101] for details.

LAD Generalized Lasso has wide applications and encompasses LAD Fused Lasso and robust change point detection problems as special cases. Gao and Huang [40] employed LAD Fused Lasso to human genomic DNA copy number data with spatial dependence and sparsity of CNV; Tang [108] investigated LAD Fused Lasso model with censored data. Change point problem has also received considerable amount of attention in the literature, to name a few, Li and Sieling [75] proposed an algorithm using multi-scale segmentation method based on FDR-control; Ng et al. [88] implemented local quadratic approximation strategy with exploitation on the banded structure of Hessian to simplify the computation; Li and Wang [76] investigated LAD change point model based on LAD adaptive Lasso method.

LAD Generalized Lasso can be transformed to a linear programming problem easily. Sparsity of the solution depends on the number of active constraints in the equivalent linear programming problem. Therefore, identification of sparsity is equivalent to the identification of active constraints. It should be noted that the state-of-the-art interior point algorithm can only approach the active constraints approximately by iterative procedure. If only finitely-many iterations are done, closeness to the active constraints must be determined by some user-chosen threshold value that is very arbitrary. Interior point method is employed by Koenker [63] with a preprocessing step for quantile regression. There are a number of alternatives to the interior point method. Wang et al. [115] established an efficient algorithm for LAD-Lasso problem that can solve the entire regularization path in one pass. Wang et al. [116] posed augmented Lagrangian method for fused lasso under general convex loss. Shi et al. [101] constructed a descent method by iteratively selecting fastest descent directions for LAD-Lasso problems under nonsmooth optimality conditions. However, none of these methods guarantees the convergence of the algorithm in finitely-many steps.

The main contribution is to propose a new active zero set descent algorithm that can stop in finitely-many steps, where the stopping conditions do not involve any user-chosen threshold value or tolerance level. The proposed algorithm check certain dynamic nonsmooth optimality conditions in each iteration and updates the active zero set and basis search directions. This makes our approach different from many other numerical approximation methods such as interior point method that requires a user-chosen threshold value to determine if an absolute value in the objective function is zero. On the contrary, the interior point method cannot be terminated in finitely-many steps. This is because it approximates the original problem using nonlinear "logarithmic barrier". As a result, Newton-like iteration is required. Moreover, the size logarithmic barrier needs to be decrease gradually in another iteration. This entails a nested iteration that do Newton update in the inner loop and decrease logarithmic barrier in the outer loop.

1.2 Constrained LAD Lasso models

The mean-variance framework of Markowitz [87] is the cornerstone for modern portfolio selection theory. Under this framework, in order to balance the risk and return, the portfolio variance is minimized at a given level expected return. This entails the estimation of the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. However, as shown in [15, 27, 37, 58], if the sample mean and sample covariance are taken as the estimation of $\boldsymbol{\mu}\,,\boldsymbol{\Sigma}\,,$ the out-of-sample performance of the asset allocation is not satisfactory in practice. In the context of regression analysis, it is well known that least absolute deviance (LAD) is more robust and resistant to outliers in the response compared to the usual least square (LS) regression, see [39, 113, 114]. The statistical properties of the constrained Lasso estimates are studied in [38, 56]. As an analogy, it is natural to believe that in the portfolio selection problem, the out-of-sample performance of a portfolio can be improved if the portfolio variance is replaced by the mean absolute value. Indeed, Konno [67] propose a mean absolute deviation (MAD) based robust portfolio selection method without involving mean vector and covariance matrix explicitly.

Sparsity is also desirable in portfolio selection because it reduces the management cost. However, this cannot be achieved by applying the method of [67] directly.

Though the Lasso penalty of Tibshirani [109] is introduced in the context of variable selection, it finds extensive applications in portfolio selection. For example, Brodie [16] develops a sparse and stable portfolio selection strategy by incorporating the idea of Lasso regularization. It is shown that the out-of-sample performance of the Lasso regularized method is consistently better than naive equal-weight portfolio in terms of Sharpe ratio. Further studies of regularized Markowitz's theory include, to name a few, [20, 34, 35, 123, 124]. However, all these methods are developed under the traditional mean-variance framework. The purpose of this paper is to incorporate Lasso penalty into MAD based portfolio selection method.

In this section, we illustrate that the proposed MAD-Lasso method can be reformulated as a constrained LAD problem with linearly equality constraints. In the absence of constraints, Shi [101] develop a steepest descent algorithm for the LAD-Lasso problem. In the present paper, we further generalize the ideas of "nonsmooth optimality conditions" and "basis directional set" to allow equality constraints. Interior point method is a competitor of the proposed algorithm. Notice that the constrained LAD problem can be transformed into a linear programming problem and therefore can be solved by the interior point method provided in the Matlab interface. However, interior point requires nested iteration that increase the tuning parameter in the outer-loop and do optimization to an approximated problem in the inner-loop. Since the solution is never exact if only finitely-many iterations are done, one needs to specify a thresholding value to determine if a component in the approximated solution equals zero. The choice of such thresholding value can be very arbitrary. On the contrary, thresholding value is not required by the proposed algorithm.

1.3 Selection of penalty parameter

The ever growing demand for mobile electronic devices, e.g., smart phones, has made voice interfaces ubiquitous. Given the mobility of these electronic devices, the input speech signal will suffer from the various environmental noise. Clearly, delivering a clean speech signal in the communication system is an important aspect of the product requirement. The objective of speech enhancement is to estimate the desired speech signal from the noisy observation, which consists of both speech and noise signals. The two key performance measures for speech enhancement are usually measured in terms of noise suppression and speech distortion [126, 82]. Interestingly, these two measures can be viewed as engineering design and quality design requirements, respectively [57, 73, 24]. In terms of engineering design, the enhancement must yield the highest signal to noise ratio (SNR) possible, which translates to noise suppression capability. In order to satisfy its quality design, the enhancement process must also maintain the perceptual features, i.e., minimizes speech quality degradation. Indeed, it is a challenge to optimize the overall noisy speech as the engineering and quality requirements [57] are at times conflicting as maximizing SNR tend to result in speech degradation [83], resulting in a natural trade-off.

Given its volume, speech signal is considered to be a big data. Additionally, speech is highly non-stationarity across the time and frequency domains. The varying nature of speech adds to the challenge as the data is not just 'big' but also changing as a function of time and frequency. There is a wealth of literature examining the characteristics of speech to reveal its patterns and trends, which are useful in applications such as speech recognition, speech enhancement and computational auditory scene analysis. Of late, one important characteristics of speech is its sparsity. Speech sparsity has gained popularity as it may hold the key to making the 'big' speech data, 'small'. Whilst speech is fairly compact and dense in the time domain, speech signals are in fact sparse in the time-frequency representations [91, 41]. This is because speech is highly non-stationary and there will be lapses of time-frequency periods where the speech power is negligible compared to the average power. On average, a speech signal consists of approximately ten to fifteen phonemes per second and each of these phonemes has a varying spectral rate [43].

The notion of sparsity has led to sparse reconstruction methods such as compressed sensing (CS) [29, 19]. CS theory states that sparse signals with a small set of linear measurements can be reconstructed with an overwhelming probability [17, 18]. Potentially, CS has the capability to compress big data such as speech signal. In speech enhancement, CS exploits the sparsity of speech and non-sparse nature of environmental noise in its reconstruction. Low et al. [84] demonstrated the use of CS as a speech enhancer by relying upon the strength of CS to maintain only the sparse components (speech) and its weakness in preserving the non-sparse components (noise). Various CS based methods with favorable results have been reported [84, 103, 119], demonstrating its efficacy for speech enhancement applications. A very popular technique for sparse signal reconstruction is the regularized ℓ_1-norm least squares [61]. This is because ℓ_1 regularized least squares yields a sparser solution since the solution tends to have fewer nonzero coefficients compared to the ℓ_2 based Tikhonov regularization [61]. One important parameter in solving the regularized sparse solution is the tuning parameter or the penalty constant, λ. The regularization parameter, λ holds significance as a heavier weighting would penalize the Tikhonov regularization. In other words, the tuning parameter holds the key in determining how sparse a solution is reconstructed.

Whilst a sparse solution indicates the existence of a sparse component such as speech, there is no measure incorporated in the CS reconstruction to optimize on the overall speech quality. The idea is to establish the relationships between sparsity and quality. Since the tuning parameter has influence over the sparsity of the solution, then a quality measure should be factored in to link the two. This is akin to factoring aspects of consumers perceive quality by using a quality measure in the overall product design [81]. We proposes to formulate the solution in compressive speech enhancement by hyper-parameterizing the tuning parameter. The tuning parameter is then optimized by using different optimization criteria (such as Gini index, the Akaike information criterion (AIC) and Bayesian information criterion (BIC)) to achieve the sparsest set of solutions. The set of solutions is then evaluated against the perceptual evaluation speech quality (PESQ) as a quality measure [96], which can be used in a wide range of operating conditions depending on the requirements. The development of such a process can then be used to describe a systematic approach to the analysis of consumer reactions to candidate designs, which ultimately provides a definition of better products and increases the product appeal [9].

1.4 Two models under Asymmetric Laplace Distributions

In this section, we consider two models under Asymmetric Laplace Distribution (ALD). The first model is mixture linear regression model with Asymmetric Laplace error, the second model is portfolio selection model under Asymmetric Laplace Distribution framework.

We first consider mixture linear regression models with error term follows mixture Asymmetric Laplace distributions. Let X be a $n \times p$ design matrix and Y be a response variable. The relationship between Y and X is often modelled via linear regression. With the framework of mixture linear regression, we assume that with probability $\pi_k, k = 1, \cdots, K$, (X_i, Y_i) comes from the k-th component if latent variable $W = k$:

$$Y = X^{\intercal}\beta_k + \sigma_k\epsilon_k, \quad k = 1, \cdots, K, \quad K \geqslant 2, \tag{1.3}$$

where the mixing proportion $\pi = (\pi_1, \cdots, \pi_K)'$ satisfies $\pi_k > 0$, and $\sum_{k=1}^K \pi_k = 1$, β_k is the unknown $p \times 1$ vector of the $100\tau\%$ regression quantiles for the k-th component with $0 < \tau < 1$, and $\sigma_k > 0$ is the corresponding unknown scalars. The random error terms ϵ_ks are assumed to be independent of X, and it is commonly assumed that the $100\tau\%$ quantile of ϵ_k is zero with variances one. When $\tau = 1/2$, it reduces to the Least Absolute Derivation (LAD) regression.

For a given random sample $\{Y_i, X_i\}_{i=1}^n$ from model (1.3), when $K = 1$, the τ-th QR is defined as any vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ minimizing the target function $Q(\boldsymbol{\beta}) = \sum_{i=1}^n \rho_\tau(Y_i - X_i^\intercal \boldsymbol{\beta})$, where $\rho_\tau(t) = t(\tau - I(t < 0))$ is the so-called check function, and $I(\cdot)$ is the usual indicator function [62, 63]. Many algorithms have been developed in the literature to tackle the minimization problem $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta})$, such as the interior point algorithm [65], the MM algorithm [52], and references therein. It is easy to show that minimizing $Q(\boldsymbol{\beta})$ is equivalent to maximizing the likelihood function of a linear regression model with random errors following the Asymmetric Laplace Distribution, see [42, 64, 128], among many others. Since ALD can be represented as a normal-variance-mean mixture with an exponential mixing distribution, which makes it easy to implement the EM algorithm for unknown parameter estimation, see [133]. Wang and Xiang [117] proposed a two-layer EM algorithm for ALD mixture regression models for composite quantile regression, which provide another form of likelihood function for composite quantile regression. The objective function for quantile regression of model (1.3) is generally

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \rho_\tau(Y_i - X_i^\intercal \boldsymbol{\beta}), \tag{1.4}$$

the robustness property of the QR procedure, and the natural connection between QR estimation and maximum likelihood estimation for the regression coefficients given the asymmetric laplace distributed random error when $K = 1$ as in [133], motivate us to consider the possible extension of the algorithm to the mixture model setup as in (1.3). For model (1.3), Yao et al. [121] proposed a robust estimation procedure for mixture linear regression models based on t distribution by extending [89]'s work, while [102] investigated a robust estimation procedure for mixture linear regression models based on Laplace distribution. The research deals with the same questions as in [121] or [102], but with the QR technique, or the ALD, instead of the less commonly used t-distribution or the special case of the standard Laplace distribution, used for achieving robustness. That is, we propose a new robust mixture regression model via ALD for (1.3), and investigate its estimates based on EM algorithm. The natural connection between the QR procedure and the MLE based on ALD error made the proposed procedures more appealing.

The MLE works well under Gaussian error case. However, MLE is sensitive to outliers or heavy tailed errors. Yao et al. [121] conducted mixture t regression models to overcome the heavy tail error cases; Song et al. [102] noticed the special connection of Laplace distribution and quantile regression, a Laplace error based linear regression is proposed to solve this problem. We consider mixture Laplace errors for quantile regression with different skewness level, a robust EM procedure is conducted to verify the robustness of our algorithm.

Then we focus on portfolio selection models under Asymmetric Laplace Distribution (ALD) framework. Portfolio selection aims at either maximizing the return or minimizing the risk. In 1952, Markowitz [87] suggests to select the portfolio by minimizing the standard deviation at a given expected return under the assumption that asset returns are normally distributed. This means that standard deviation is chosen as the risk measure. Markowitz's work laid down the cornerstone for modern portfolio selection theory framework.

Risk measures and probability distributions are two important constituents of the portfolio selection theory. Traditional Markowitz's model [87] is established based on normality assumption and standard deviation is chosen as the risk measure.

One disadvantage of taking standard deviation (StD) as a risk measure is that the loss in the extreme cases tends to be underestimated. To overcome such a difficulty, the idea of Value at Risk (VaR) is also widely used in practice. Artzner et al. [6] suggests that desirable risk measure should be "coherent". However, VaR does not fulfill the subadditivity condition as required by the definition of "coherence". Yiu [125] proposed an optimal portfolio selection under Value-at-Risk. On the other hand, Expected Shortfall (ES) is coherent as a popular risk measure for portfolio selection that aims at averaging the tail uncertainties.

It is well-known that financial data cannot be described satisfactorily by normal distribution. The normality assumption is restrictive and is generally violated due to financial market uncertainties and managers' risk aversion. As Behr and Ptter [11] pointed out, alternatives for multivariate normal distribution are necessary for portfolio selection. A desirable alternative model should be able to explain tail heaviness, skewness, and excess kurtosis. Various heavy tailed distributions have been applied to portfolio selection problems. Among these, Mandelbrot [3] concluded that the daily rate of return of stock price data exhibit heavy tailed distributions; Hu and Kercheval [49] apply multivariate skewed t and student t distribution for efficient frontier analysis; Generalized hyperbolic distribution is extensively studied in [11, 30, 47, 48, 106, 107], with special cases including hyperbolic distribution [13, 31], Variance Gamma distribution [100], Normal Inverse Gaussian distribution [8], etc.

Recently, Asymmetric Laplace distribution has received various attention in the literature, to name a few, [7, 66, 70, 71, 94]. Compared to Normal distribution, the Asymmetric Laplace distribution describes asymmetry, steep peak, and tail heaviness better. Portfolio selection models are extensively studied under Asymmetric Laplace framework. Zhu [134], Kozubowski and Podgrski [71] apply Asymmetric Laplace distribution to financial data. By assuming that the asset data is generated from autoregressive moving average (ARMA) time series models with Asymmetric Laplace noise, Zhu [134] establish the asymptotic inference theory under very mild conditions and present methods of computing conditional Value at Risk (CVaR). Zhao et al. [132] further propose a so-called mean-CVaR-skewness portfolio selection strategy under Asymmetric Laplace distribution, this model can be further transformed to quadratic programming problem with explicit solutions.

In this subsection, we extended Hu [49]'s work to Asymmetric Laplace framework. We first derived the equivalence of mean-VaR/ES/Std-skewness-kurtosis models, and show that these models can be reduced to quadratic programming problem. S-ince Zhao [132] utilized moment estimation for parameter estimation of Asymmetric Laplace distribution which is less efficient compare to maximum likelihood estimation. Taken into consideration of the normal mean-variance mixture of Asymmetric Laplace distribution, followed by Expectation-Maximization algorithm for multivariate Laplace distribution in Arslan [4], we derived the EM algorithm for Asymmetric Laplace Distributions that outperforms moment estimation in [132]. The advantage of the proposed EM algorithm is to alleviate the complicated calculation of Bessel function. This improves many existing methods of estimating Asymmetric Laplace distributions, for example, Hrlimann [55], Kollo and Srivastava [66], Visk [112]. Extensive simulation studies and efficient frontier analysis are complemented to confirm that our algorithm performs better than moment estimation for parameter estimation.

1.5 Contributions of the thesis

The contributions of this thesis can be divided into four parts:

1. **LAD Generalized Lasso regression**.

   In Chapter 2, we first focus on LAD-Lasso regression, we derived the optimality condition for optimal solutions, and developed a descent algorithm such that the nonsmooth optimization problem can be optimized directly. Then we construct the active zero set descent algorithm for LAD Generalized Lasso. Under dynamic nonsmooth optimality conditions, based on zero set and basis directional set, we update the descent directions and optimal step length recursively without user-chosen threshold value. Simulation studies and real data analysis are provided to confirm that our algorithms perform well.

2. **Constrained LAD Lasso for portfolio optimization**.

   In Chapter 3, we established the MAD-Lasso portfolio selection strategy, reformulated as Constrained LAD Lasso with linearly equality constraints. We develop a descent algorithm by updating descent directions from basis directional set and optimal step length iteratively for solutions of MAD-Lasso model.

3. **Penalty parameter selection for compressive sensing**.

   In Chapter 4, we first propose a two-level optimization strategy to incorporate the affective design attributes in the sparse solution in compressive speech enhancement by hyper-parameterizing the tuning parameter, and provide selection criteria for tuning parameter selection.

4. **Two models under Asymmetric Laplace Distributions**.

   In Chapter 5, we first propose a two-level latent EM algorithm for parameter estimation of mixture linear regression models by assuming that the error term follows mixture laplace distribution. Then we consider portfolio selection under Asymmetric Laplace Distribution (ALD) framework, and derived the EM algorithm for parameter estimation, we also prove that minimize VaR, ES and StD under ALD framework can be simplified to quadratic programming with explicit solutions. Extensive simulation studies and real data analysis confirmed that our proposed methodology works well.

1.6 Organizations of the thesis

The thesis is structured as follows.

- In Chapter 1, we introduce existing background knowledge of four topics: LAD Generalized Lasso models, Constrained LAD Lasso models, penalty parameter selection for compressive sensing, two models under Asymmetric Laplace Distribution. Then we summarized the main contributions of the thesis.

- In Chapter 2, we focus on LAD Generalized Lasso models. We first develop a descent method by choosing the fastest decent direction for LAD-Lasso model, then we present a new active zero set descent algorithm for LAD Generalized Lasso by updating the descent directions and optimal step length recursively based on zero set and basis directional set without user-chosen threshold value, convergence analysis are conducted.

- In Chapter 3, we consider Constrained LAD Lasso models, and conduct a descent algorithm by iteratively updating descent directions and optimal step length, then we apply the algorithm to MAD-Lasso portfolio selection strategy.

- In Chapter 4, we derive a Two-Level tuning parameter selection strategy for compressive sensing based signal processing model.

- In Chapter 5, we study two models under Asymmetric Laplace Distributions. We first conduct mixture linear regression model and derived a two-level Expectation-Maximization algorithm for model fitting, then we investigate robust portfolio selection models under Asymmetric Laplace framework.

- In Chapter 6, we summarize our main results in this thesis and provide several further possible research directions.

Chapter 2

LAD Generalized Lasso Models

In this chapter, we focus on LAD Generalized Lasso models. Based on nonsmooth optimality conditions and directional derivatives, we derive a descent method for LAD-Lasso model and a new active zero set descent algorithm for LAD Generalized Lasso model. Compared to interior point method, we verify that our algorithms are much more time efficient than state-of-the-art linear programming solver: interior point method.

2.1 A descent method for LAD-Lasso model

Consider linear regression problem

$$Y = X\beta + \varepsilon, \tag{2.1}$$

where X is the $n \times p$ design matrix with row vectors $X_i \in \mathbb{R}^p, i = 1, \cdots, n$, and $Y = (y_1, \cdots, y_n)^\intercal$ is the response vector, $\beta = (\beta_1, \cdots, \beta_p)^\intercal$ is the parameter vector we are concerned.

Generally, the LAD-Lasso regression is to minimize the l_1 norm loss function

$$\min_{\beta} \sum_{i=1}^{n} |y_i - X_i\beta|$$

subject to the constraint

$$\sum_{i=1}^{p} |\beta_i| < c,$$

where c is a positive constant.

This problem can be transformed into the following optimization problem:

$$\min_{\beta} \sum_{i=1}^{n} |y_i - X_i\beta| + \gamma \sum_{j=1}^{p} |\beta_j|,$$

or the matrix representation

$$\min_{\beta} \|Y - X\beta\|_1 + \gamma\|\beta\|_1. \tag{2.2}$$

Note that the terms in (2.2) are nonsmooth. A typical way to tackle this problem is to transform it into a linear programming problem. Denote

$$\|Y - X\beta\|_1 = u_1 + v_1 \, , \|\beta\|_1 = u_2 + v_2, \tag{2.3}$$

where $u_1, v_1, u_2, v_2 \geqslant 0$ and $u_1, v_1 \in \mathbb{R}^n$, $u_2, v_2 \in \mathbb{R}^p$ are defined as

$$\begin{aligned} u_1 &= \max\big(Y - X\beta, 0\big), \\ v_1 &= \max\big(-(Y - X\beta\big), 0), \\ u_2 &= \max(\beta, 0), \\ v_2 &= \max(-\beta, 0). \end{aligned}$$

Hence

$$Y - X\beta = u_1 - v_1 \, , \beta = u_2 - v_2,$$

and (2.2) is equivalent to the following minimization problem:

$$\begin{aligned} \min \quad & 0 \cdot \beta + u_1 + v_1 + \gamma(u_2 + v_2) \\ \text{s.t.} \quad & X\beta + u_1 - v_1 = Y, \\ & \beta - u_2 + v_2 = 0_p, \\ & u_1, v_1 \geqslant 0_n, u_2, v_2 \geqslant 0_p. \end{aligned}$$

Denote

$$A = \begin{pmatrix} X & I & -I & 0_n & 0_n \\ I_p & 0_{p\times n} & 0_{p\times n} & -I_p & I_p \end{pmatrix}, b = \begin{pmatrix} Y \\ 0_p \end{pmatrix},$$

the optimization problem becomes

$$\begin{aligned} \min \quad & c^\intercal x \\ \text{s.t.} \quad & Ax = b, \\ & u_1, v_1 \geqslant 0_n, u_2, v_2 \geqslant 0_p, \end{aligned} \tag{2.4}$$

where $x = (\beta^\intercal, u_1^\intercal, v_1^\intercal, u_2^\intercal, v_2^\intercal)^\intercal, c = (0_p, I, I, \gamma I, \gamma I)$.

Thus, (2.4) is a canonical linear programming problem and interior point method can be applied to solve it. This is currently the state-of-art technique for tackling the LAD-Lasso problem. However, when n and p become large, the computational time still grows significantly and becomes very expensive. Problem (2.2) can be written as a canonical form by introducing the symbols as follows:

$$Y^* = \begin{pmatrix} Y \\ 0_p \end{pmatrix}, X^* = \begin{pmatrix} X \\ \gamma \cdot I_p \end{pmatrix},$$

where 0_p is $p \times 1$ vector, I_p is p-dimensional identity matrix. Then, Problem (2.2) becomes

$$\min_\beta \|Y^* - X^*\beta\|_1. \tag{2.5}$$

For simplicity of notation, we omit the superscript $*$ and consider the canonical form

$$\min_\beta \|Y - X\beta\|_1. \tag{2.6}$$

2.1.1 Computational methodology

Introducing the objective function $f(\beta)$, the optimization problem (2.6) is standardized as

$$\min_{\beta\in\mathbb{R}^p} f(\beta) = \sum_{i=1}^n |X_i\beta - y_i| = \sum_{i=1}^n f_i(\beta), \tag{2.7}$$

where

$$f_i(\beta) = |X_i\beta - y_i| = \begin{cases} X_i\beta - y_i, & \text{if } X_i\beta - y_i > 0, \\ -X_i\beta + y_i, & \text{if } X_i\beta - y_i < 0, \\ 0, & \text{if } X_i\beta - y_i = 0. \end{cases}$$

To develop an efficient method for solving Problem (2.6), the optimality conditions are needed. The derivative of f_i with respect to β is given by

$$\frac{\partial f_i}{\partial \beta} = \begin{cases} X_i, & \text{if } X_i\beta - y_i > 0, \\ -X_i, & \text{if } X_i\beta - y_i < 0. \end{cases}$$

At the point when $X_i\beta - y_i = 0$, it's not differentiable. However, its directional derivative exists. For a direction $d \in \mathbb{R}^n$, the directional derivative of f_i along d is defined as

$$\nabla_{d^+} f_i = \lim_{\lambda \to 0^+} \frac{|X_i(\beta + \lambda d) - y_i| - |X_i\beta - y_i|}{\lambda \|d\|} = \frac{|X_i d|}{\|d\|}.$$

Similarly, for the direction $-d$, directional derivative of f_i along $-d$ is defined as

$$\nabla_{d^-} f_i = \lim_{\lambda \to 0+} \frac{|X_i(\beta - \lambda d) - y_i| - |X_i\beta - y_i|}{\lambda \|d\|} = \frac{|X_i d|}{\|d\|}.$$

Hence, for the absolute linear function, we have

$$\nabla_{d^-} f_i = \nabla_{d^+} f_i.$$

Furthermore, if $X_i\beta - y_i \neq 0$, then f_i is smooth and we have

$$\nabla_{d^-} f_i = -\nabla_{d^+} f_i.$$

Denote $X_i\beta - y_i = u_i$, we rewrite the objective function as

$$f(\beta) = A(\beta) + C(\beta),$$

where $A(\beta)$ relate to the smooth part of $\nabla_d f(\beta)$,

$$A(\beta) = \sum_{i=1}^{n} \chi(u_i > 0)(X_i\beta - y_i) + \sum_{i=1}^{n} \chi(u_i < 0)(-X_i\beta + y_i) \triangleq a^\intercal \beta + b,$$

in which

$$\chi(\nu) = \begin{cases} 1, & \text{if } \nu \text{ is true,} \\ 0, & \text{otherwise,} \end{cases}$$

$$a^{\mathsf{T}} = \sum_{i=1}^{n} \chi(u_i > 0) X_i - \sum_{i=1}^{n} \chi(u_i < 0) X_i,$$

$$b = -\sum_{i=1}^{n} \chi(u_i > 0) y_i + \sum_{i=1}^{n} \chi(u_i < 0) y_i,$$

and $C(\beta)$ relate to the nonsmooth part of $\nabla_d f(\beta)$.

Denote the zero set in each iteration by $\Omega_k = \{k_1, \cdots, k_m\}$, which is the set of all the indices i such that $u_i = 0$. Then

$$C(\beta) = \sum_{i=1}^{n} \chi(u_i = 0)|X_i\beta - y_i| = \sum_{i=1}^{m} |X_{k_i}\beta - y_{k_i}| = \sum_{i \in \Omega_k} |X_i\beta - y_i|.$$

Since $f(\beta)$ is the sum of n convex functions, it is convex and its local minimizer is also the global minimizer. The optimality condition of the minimizer is that any directional derivatives are greater than or equal to zero. That is, β^* is the optimal solution of (2.7) if and only if

$$\nabla_d f(\beta^*) = \nabla_d A(\beta^*) + \nabla_d C(\beta^*) \geqslant 0\,, \forall d \in \mathbb{R}^p. \tag{2.8}$$

However, it is not easy to verify this condition during computation since d is arbitrary. We should derive an equivalent condition such that it can be verified easily. Consider the function $C(\beta)$ such that

$$X_{k_i}\beta = y_{k_i}\,, i = 1, \cdots, m.$$

Denote

$$X_a = \begin{pmatrix} X_{k_1} \\ \vdots \\ X_{k_m} \end{pmatrix},$$

and suppose that the rank of X_a is m, we can find its generalized inverse matrix as V_a such that $X_a V_a = I_m$, where I_m is the $m \times m$ identity matrix and $V_a = (V_1, \cdots, V_m)$.

Consider the null space $\{V \in \mathbb{R}^p | X_a V = 0\}$. There exist $p - m$ linear independent vectors $V_j, j = m+1, \cdots, p$, which are the basis of the null space. Hence, we have

$$X_a V_j = 0, \forall j = m+1, \cdots, p.$$

Therefore, $\{V_i : i = 1, \cdots, p\}$ form a basis of $\mathbb{R}^p$ and the following orthonormality holds:

$$X_{k_i} V_j = \begin{cases} 1, & \text{when } i = j; \\ 0, & \text{when } i \neq j, \end{cases} \qquad i = 1, \cdots, m\,, j = 1, \cdots, p, \tag{2.9}$$

Then we can obtain the directional derivatives of f along the vectors $\{V_j : j = 1, \cdots, p\}$. If $i \in \{1, \cdots, m\}$, we have

$$\begin{aligned} \nabla_{V_i^+} C(\beta) &= \frac{\sum_{j=1}^m |X_{k_j} V_i|}{\|V_i\|} = \frac{1}{\|V_i\|}\,, i = 1, \cdots, m, \\ \nabla_{V_i^-} C(\beta) &= \frac{\sum_{j=1}^m |X_{k_j}(-V_i)|}{\|-V_i\|} = \frac{1}{\|V_i\|}\,, i = 1, \cdots, m. \end{aligned}$$

If $i \in \{m+1, \cdots, p\}$, we have

$$\nabla_{V_i} C(\beta) = \frac{\sum_{j=1}^m |X_{k_j} V_i|}{\|V_i\|} = 0, i = m+1, \cdots, p.$$

Consequently, we have

$$\begin{aligned} &\nabla_{V_i^+} f(\beta) = \nabla_{V_i^+} A(\beta) + \frac{1}{\|V_i\|} = (a^\intercal V_i + 1)/\|V_i\|\,, \quad i = 1, \cdots, m. \\ &\nabla_{V_i^-} f(\beta) = \nabla_{V_i^-} A(\beta) + \frac{1}{\|V_i\|} = (-a^\intercal V_i + 1)/\|V_i\|\,, \quad i = 1, \cdots, m. \qquad (2.10) \\ &\nabla_{V_i} f(\beta) = \nabla_{V_i} A(\beta) = a^\intercal V_i/\|V_i\|\,, \quad i = m+1, \cdots, p. \end{aligned}$$

An equivalent optimal condition of (2.8) is given by the following theorem.

Theorem 2.1. *β^* is the optimal solution if and only if the directional derivatives satisfy*

$$\begin{aligned}
&\nabla_{V_i^+} f(\beta^*) \geqslant 0\,, \quad i = 1, \cdots, m. \\
&\nabla_{V_i^-} f(\beta^*) \geqslant 0\,, \quad i = 1, \cdots, m. \\
&\nabla_{V_i} f(\beta^*) = 0\,, \quad i = m+1, \cdots, p.
\end{aligned} \tag{2.11}$$

Proof. Note that (2.11) is a special case of (2.8), the necessary condition is obvious. Therefore, we only prove the sufficient condition, that is, we prove that if (2.11) are satisfied, then (2.8) holds.

For any direction d, since $\{V_i : i = 1, \cdots, p\}$ is a basis of $\mathbb{R}^p$, there exists a vector λ, such that

$$d = \sum_{i=1}^{p} \lambda_i V_i. \tag{2.12}$$

Without loss of generality, we can set $\lambda_i \geqslant 0, \forall i = 1, \cdots, p$, because if $\lambda_i < 0$, we have $\lambda_i V_i = (-\lambda_i) \cdot V_i^-$. Then V_i^+ is replaced by V_i^-, and λ_i is replaced by $-\lambda_i > 0$.

Hence, by adjusting the order adequately, (2.12) can be reorganized as

$$d = \sum_{i=1}^{m_1} \lambda_i V_i^+ + \sum_{i=m_1+1}^{m} \lambda_i V_i^- + \sum_{i=m+1}^{p} \lambda_i V_i.$$

where $\lambda_i \geqslant 0, \forall i = 1, \cdots, p$. It follows from (2.10) that

$$\begin{aligned}
\nabla_d C(\beta^*) &= \frac{\sum_{i=1}^{m} |X_{k_i} d|}{\|d\|} \\
&= \frac{\sum_{i=1}^{m} \left| X_{k_i} \big(\sum_{j=1}^{m_1} \lambda_j V_j^+ + \sum_{j=m_1+1}^{m} \lambda_j V_j^- + \sum_{j=m+1}^{p} \lambda_j V_j \big) \right|}{\|d\|} \\
&= \frac{\sum_{i=1}^{m_1} \left| \lambda_i X_{k_i} V_i^+ \right| + \sum_{i=m_1+1}^{m} \left| \lambda_i X_{k_i} V_i^- \right|}{\|d\|} \\
&= \frac{\sum_{i=1}^{m} \lambda_i}{\|d\|}.
\end{aligned}$$

Hence, by (2.10), we have

$$
\begin{aligned}
\nabla_d f(\beta^*) &= \left(\sum_{i=1}^{m_1} \lambda_i a^\intercal V_i^+ + \sum_{i=m_1+1}^{m} \lambda_i a^\intercal V_i^- + \sum_{i=m+1}^{p} \lambda_i a^\intercal V_i + \sum_{i=1}^{m} \lambda_i\right) \Big/ \|d\| \\
&= \left(\sum_{i=1}^{m_1} \lambda_i (a^\intercal V_i^+ + 1) + \sum_{i=m_1+1}^{m} \lambda_i (a^\intercal V_i^- + 1) + \sum_{i=m+1}^{p} \lambda_i a^\intercal V_i\right) \Big/ \|d\| \\
&= \left(\sum_{i=1}^{m_1} \nabla_{V_i^+} f(\beta^*) \cdot \|V_i\| + \sum_{i=m_1+1}^{m} \nabla_{V_i^-} f(\beta^*) \cdot \|V_i\|\right) \Big/ \|d\| \\
&\geqslant 0.
\end{aligned}
$$

Thus for any direction d, the directional derivative is greater than or equals to zero. Hence, (2.8) holds and β^* is the optimal solution. □

Remark 2.1. *If the rank of X_a is l, and $l < m$, we can find l rows such that they are rank l. Then, the generalized inverse matrix $V_a = (V_1, \cdots, V_l)$ can be computed. Since Theorem 2.1 does not hold by replacing m by l, extend Theorem 2.1 to multi-collinear cases need further exploration.*

If the condition (2.11) is not satisfied, then there exists a direction d such that the cost function value decreases along with this direction. If the i-th condition is not satisfied, that is,

$$\nabla_{V_i^+} f(\beta) \geqslant 0 \text{ and } \nabla_{V_i^-} f(\beta) \geqslant 0$$

can not be satisfied at the same time, then V_i^+ or V_i^- is the descent direction. For an iterative point $\beta^{(k)}$, denote the zero set by Ω_k. The function can be rewritten as

$$f(\beta) = a^{(k)\intercal} \beta + \sum_{i \in \Omega_k} |X_i \beta - y_i| + b^{(k)}. \tag{2.13}$$

We need to find a descent direction such that (2.13) decreases along it whenever the condition (2.11) is not satisfied.

Lemma 2.1. *Suppose that $d_1, \cdots, d_m$ are the descent directions and are also linear independent, then for any $w_i \geqslant 0$, and at least one i such that $w_i > 0$, $\sum_{i=1}^m w_i d_i$ is also the descent direction.*

Proof. By the definition of directional derivative, the directional derivative of d_i and $w_i d_i \, (w_i > 0)$ are the same. Hence, $w_i d_i$ is also the descent direction. Hence, we can assume that $\sum_{i=1}^m w_i = 1, w_i \geqslant 0$.

Suppose that

$$\nabla_{d_i} f(\beta) = \lim_{t \to 0^+} \frac{f(\beta + td_i) - f(\beta)}{\|td_i\|} = \alpha_i < 0.$$

Since f is linear along with d_i when $t > 0$ is small, there exists $\varepsilon_i > 0$ such that

$$f(\beta + td_i) \approx f(\beta) + t\alpha_i \|d_i\|, \quad t \in [0, \varepsilon_i].$$

Let $\varepsilon = \min\{\varepsilon_1, \cdots, \varepsilon_m\}$. When $t \in (0, \varepsilon]$, we have

$$\begin{aligned} & \frac{f\big(\beta + t\sum_{i=1}^m w_i d_i\big) - f(\beta)}{\|t\sum_{i=1}^m w_i d_i\|} \\ = & \frac{f\big(\sum_{i=1}^m w_i \beta + t\sum_{i=1}^m w_i d_i\big) - \sum_{i=1}^m w_i f(\beta)}{\|t\sum_{i=1}^m w_i d_i\|} \\ = & \frac{f\big(\sum_{i=1}^m w_i (\beta + td_i)\big) - \sum_{i=1}^m w_i f(\beta)}{\|t\sum_{i=1}^m w_i d_i\|} \\ \leqslant & \frac{\sum_{i=1}^m w_i \big(f(\beta + td_i) - f(\beta)\big)}{\|t\sum_{i=1}^m w_i d_i\|} \\ = & \frac{t\sum_{i=1}^m w_i \alpha_i \|d_i\|}{\|t\sum_{i=1}^m w_i d_i\|}. \end{aligned}$$

Hence,

$$\nabla_{\sum_{i=1}^m w_i d_i} f(\beta) = \lim_{t \to 0^+} \frac{f(\beta + t\sum_{i=1}^m w_i d_i) - f(\beta)}{\|t\sum_{i=1}^m w_i d_i\|} \leqslant \lim_{t \to 0^+} \frac{t\sum_{i=1}^m \alpha_i w_i \|d_i\|}{t\|\sum_{i=1}^m w_i d_i\|} = \frac{\sum_{i=1}^m \alpha_i w_i \|d_i\|}{\|\sum_{i=1}^m w_i d_i\|}.$$

Since $d_1, \cdots, d_m$ are linear independent and $\{w_i | i = 1, \cdots, m\}$ are not all zero, $\| \sum_{i=1}^m w_i d_i \|$ is not equal to zero. Note that $\alpha_i < 0$, $w_i > 0$ and $\|d_i\| > 0$, we have $\frac{\sum_{i=1}^m \alpha_i w_i \|d_i\|}{\| \sum_{i=1}^m w_i d_i \|} < 0$. Hence, $\sum_{i=1}^m w_i d_i$ is a descent direction. □

Lemma 2.1 indicate that linear combination of descent directions is still a descent direction, thus the following descent direction search is feasible.

Since there exists at least one $i \in \{1, \cdots, m\}$ such that condition (2.11) is not satisfied. Denote the set of all such indices k_i by Ω'_k, where (2.11) is not satisfied for V_i^+ or V_i^-. Then, we can choose the descent direction d in the space spanned by

$$\{V_i : k_i \in \Omega'_k\} \cup \{V_i : i = m+1, \cdots, p\}.$$

To speed up the search, we check the descent directional derivatives $\nabla_{V_i^+} f$ or $\nabla_{V_i^-} f$, and choose the indices where they descent most. That is, we choose a subset Λ_1 of Ω'_k, which is a proportional α of the indices in Ω'_k such that the corresponding descent directional derivatives $\nabla_{V_i^+} f$ or $\nabla_{V_i^-} f$ is less than the other $1 - \alpha$ of the directional derivatives. Denote

$$\Omega_{0k} = \Omega_k \backslash \Lambda_1,$$

we choose the descent direction d in the space spanned by

$$\{V_i : k_i \in \Lambda_1\} \cup \{V_i : i = m+1, \cdots, p\}$$

such that

$$d = \sum_{k_i \in \Lambda_1} \lambda_i V_i + \sum_{i=m+1}^{p} \lambda_i V_i.$$

It can be verified that

$$X_i d = 0\,, \forall i \in \Omega_{0k}.$$

Hence, the descent direction should keep the set Ω_{0k} unchanged, we set the descent direction $d^{(k)}$ as the optimal solution of

$$\begin{aligned} \max_{h \in \mathbb{R}^p} \quad & -a^{(k)}h \\ s.t. \quad & X_i h = 0, \forall i \in \Omega_{0k}, \\ & \|h\| = 1. \end{aligned} \tag{2.14}$$

It means that the solution h is chosen as the vector nearest to the deepest descent direction $-a^{(k)}$, and still keep the set Ω_{0k} unchanged at the same time. The optimal solution of Problem (2.14) is

$$\tilde{d} = -a^{(k)} - X_{0k}^{\intercal}(X_{0k}X_{0k}^{\intercal})^{-1}X_{0k} \cdot (-a^{(k)}), \tag{2.15}$$

where $X_{0k}^{\intercal}(X_{0k}X_{0k}^{\intercal})^{-1}X_{0k}(-a^{(k)})$ is the projected direction of $-a^{(k)}$ in the subspace $\{h : X_i h = 0, i \in \Omega_{0k}\}$, and

$$X_{0k} = \begin{pmatrix} X_{k_1} \\ \vdots \\ X_{k_l} \end{pmatrix}, k_1, \cdots, k_l \in \Omega_{0k}.$$

The sparsity indicate that $m \ll p$ and $m \ll n$, thus computational complexity of $\tilde{d}$ is very low during iterations. Hence, the descent direction $d^{(k)}$ can be chosen as the normalized vector of $\tilde{d}$

$$d^{(k)} = \tilde{d}/\|\tilde{d}\|, \tag{2.16}$$

and the zero set is updated as $\Omega_k = \Omega_{0k}$.

The cost function value will decrease along the descent direction $d^{(k)}$, when the step length is small. The next iteration point will be generated by

$$\beta^{(k+1)} = \beta^{(k)} + \lambda_k d^{(k)}, \lambda_k > 0,$$

where λ_k is the step length, which should be maximized such that the cost function value is reduced in largest magnitude. For this, we define a new problem as

$$\min_{\lambda \geqslant 0} g(\lambda)$$

where

$$g(\lambda) = f(\beta^{(k+1)}) = f(\beta^{(k)} + \lambda d^{(k)}), \quad \lambda \geqslant 0.$$

Since f is convex, $g(\lambda)$ is also convex, we can choose λ_k as the optimal solution of the problem $\min\limits_{\lambda} g(\lambda)$. This problem is equivalent to the problem as follows:

$$\begin{aligned} \max_{\lambda \geqslant 0} \quad & \lambda \\ \text{s.t.} \quad & \nabla_{d^{(k)}} f(\beta^{(k)} + \lambda d^{(k)}) \geqslant 0 \\ & \nabla_{d^{(k)}-} f(\beta^{(k)} + \lambda d^{(k)}) \geqslant 0. \end{aligned} \tag{2.17}$$

For this problem, we have the following observation.

Theorem 2.2. *There exists an optimal solution $\lambda^{(k)} > 0$ and at least one i in $\{1, \cdots, n\}$ such that $X_i(\beta^{(k)} + \lambda^{(k)} d^{(k)}) = y_i$, that is, i is in the zero set at the point $\beta^{(k)} + \lambda^{(k)} d^{(k)}$.*

Proof. If $\lambda = 0$, $d^{(k)}$ is a descent direction at $\beta^{(k)}$, that is,

$$\nabla_{d^{(k)}} f(\beta^{(k)}) < 0.$$

Since $g(\lambda)$ is convex, $\frac{\partial g(\lambda)}{\partial \lambda}$ is monotonically increasing.

Note that

$$\begin{aligned} \frac{\partial g(\lambda)}{\partial \lambda} &= \lim_{\Delta\lambda \to 0} \frac{g(\beta^{(k)} + (\lambda + \Delta\lambda) d^{(k)}) - g(\beta^{(k)} + \lambda d^{(k)})}{\Delta\lambda} \\ &= \|d^{(k)}\| \nabla_{d^{(k)}} f(\beta^{(k)} + \lambda d^{(k)}), \end{aligned}$$

then, the directional derivative

$$\nabla_{d^{(k)}} f(\beta^{(k)} + \lambda d^{(k)})$$

is monotonically increasing with respect to λ.

Note that each term is absolute linear function, $f(\beta^{(k)} + \lambda d^{(k)})$ is piecewise linear and $\nabla_{d^{(k)}} f(\beta^{(k)} + \lambda d^{(k)})$ is piecewise constant. For each point where $\nabla_{d^{(k)}} f(\beta^{(k)} + \lambda d^{(k)})$ increases, there exists at least one index i such that u_i changes from negative to positive or from positive to negative. All these indices i is in $\{1, \cdots, n\}$, which is finite. Suppose that

$$\lim_{\lambda \to +\infty} \nabla_{d^{(k)}} f(\beta^{(k)} + \lambda d^{(k)}) < 0,$$

we have

$$\lim_{\lambda \to +\infty} \nabla_{d^{(k)}} f(\beta^{(k)} + \lambda d^{(k)}) = \nabla_{d^{(k)}} f(\beta^{(k)} + \lambda' d^{(k)}) < 0,$$

where λ' is a sufficiently large value. Therefore,

$$f(\beta^{(k)} + \lambda' d^{(k)}) \leqslant f(\beta^{(k)}) + \nabla_{d^{(k)}} f(\beta^{(k)} + \lambda' d^{(k)}) \to -\infty.$$

This contracts to the fact that $f \geqslant 0$, which is impossible. Thus we must have

$$\lim_{\lambda \to +\infty} \nabla_{d^{(k)}} f(\beta^{(k)} + \lambda d^{(k)}) \geqslant 0.$$

Since $\nabla_{d^{(k)}} f(\beta^{(k)} + \lambda d^{(k)})$ is piecewise linear, we can find a point λ' such that $\nabla_{d^{(k)}} f(\beta^{(k)} + \lambda' d^{(k)})$ becomes positive or zero in the first time. That is,

$$\begin{aligned} \nabla_{d^{(k)}} f(\beta^{(k)} + \lambda d^{(k)}) < 0\,, \lambda < \lambda', \\ \nabla_{d^{(k)}} f(\beta^{(k)} + \lambda d^{(k)}) \geqslant 0\,, \lambda \geqslant \lambda'. \end{aligned}$$

Hence, $\beta^{(k)} + \lambda' d^{(k)}$ is the minimum point of $f(\beta^{(k)} + \lambda d^{(k)})$.

Note that $\nabla_{d^{(k)}} f(\beta^{(k)} + \lambda d^{(k)})$ is discontinuous at λ', there exists at least one index i in $\{1, \cdots, n\}$ such that

$$X_i(\beta^{(k)} + \lambda' d^{(k)}) = 0.$$

□

Hence, we can find the optimum step length in each iteration.

We denote λ_k as the optimum step length along the direction $d^{(k)}$. By using the step length λ_k, the cost function becomes

$$f(\beta^{(k+1)}) = (a^{(k+1)})^{\intercal}\beta^{(k+1)} + \sum_{i \in \Omega_{k+1}} |X_i \beta^{(k+1)}| + b^{(k+1)},$$

and the k-th iteration terminated and moved to the $(k+1)$-th iteration. For this update, the indices in Λ_1 have been removed from the zero set Ω_k. It follows from Theorem 2.2 that some indices move to the zero set. We denote all these indices by Λ_2, then a new zero set at $(k+1)$-th iteration is generated as

$$\Omega_{k+1} = \Omega_k \cup \Lambda_2.$$

Hence, we find a new iterate as $\beta^{(k+1)} = \beta^{(k)} + \lambda_k d^{(k)}$, and the zero set is updated as Ω_{k+1}. We continue the iteration until the optimal conditions (2.11) are satisfied. In summary, the algorithm is as follows:

Algorithm 1

Initialization: Choose an initial point $\beta^{(0)}$, compute the corresponding set Ω_0, and compute the cost function $f(\beta^{(0)})$. Set $k = 0$.

Step 1: (Terminate)
Generate the matrix V for the zero set Ω_k. If conditions (2.11) is satisfied, then stop and return the optimal solution and value. Otherwise, go to Step 2.

Step 2: (Descent Direction)
Find the α fastest descent directions as Λ_1, where α denotes the percentage of selected descent directions that decrease faster than the other $1 - \alpha$ directions. Set $\Omega_{0k} = \Omega_k \backslash \Lambda_1$, and compute the descent direction $d^{(k)}$ using (2.16).

Step 3: (Optimal Step Length) Find the best step length λ_k by (2.31).

Step 4: (Iteration) Update $\beta^{(k+1)} = \beta^{(k)} + \lambda_k d^{(k)}$. Find Λ_2 and update the zero set as $\Omega_{k+1} = \Omega_{0k} \cup \Lambda_2$. Then we compute the cost function $f(\beta^{(k+1)})$, let $k = k + 1$ and go to Step 1.

2.1.2 Simulation studies

In this section, Algorithm 1 is implemented to solve the LAD-Lasso problem, where parameter α controls the percentage of directions selected from the descent direction set. Experiments show that too small or too large α values may result in unsteadiness or time inefficiency, e.g., $\alpha = 0.01, 0.20$. Here α is set as 0.05 to reach a balance between stability and time consumption.

We compare our proposed method with Interior Point method and Gurobi based on Matlab platform, where the default solver of function `linprog` is Interior Point method.

To solve LAD-Lasso problem, a key consideration is the tuning parameter selection. In [114], Wang focused on the high dimensional penalized least absolute deviation problem, and a tuning parameter selection procedure is given. Denote $\boldsymbol{x}_i$ as the i-th column vector of design matrix X, we first scaled the dataset such that $\|\boldsymbol{x}_i\|^2 = n, i = 1, \cdots, p$, and choose $\lambda = \sqrt{2n \log p}$, which is rate consistent. Similar to Gao and Huang [39], we consider four simulation examples with data generated by

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, 1),$$

where design matrix X follows multivariate Gaussian distribution with zero mean vector and covariance matrix Σ, the elements of Σ is given by $(\Sigma)_{ij} = 0.5^{|i-j|}$ such that the correlation between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is $0.5^{|i-j|}$. For simplicity, the true coefficient β is given by

$$\beta = (2, 2, 2, 2, 2, 0, \cdots, 0),$$

where the first five elements equal to 2 and the remaining $p - 5$ elements are zeroes, thus there are 5 nonzero components.

We consider four cases of p as $10\,, 50\,, 100\,, 500$, respectively. For each p, the value of n increases from 500 (100 for $p = 10$ case) to 10000 gradually. For each p and

n, the data X and Y are simulated 100 times. Interior point method, the proposed method and Gurobi are applied to these problems for comparison. The running time of these methods are depicted in Figure 2.1. It can be seen that the proposed method is more efficient than the other methods, especially when n increases. That is, the larger n/p is, the more efficient the proposed method becomes. We choose $\beta = 0_p$ as the initial value, since there are 5 nonzero entries for each p, experiments show that $m < 10$ for all the iterations, thus the computational complexity is relatively small.

Several representative simulation results are listed in Table 2.1–2.4, where Running Time denotes the average time taken; MSE evaluates the average prediction error; Degree of Freedom (Zou [137]) refers to the number of nonzero components of the estimator; Correctly Fitted Ratio indicates accurate estimation of nonzero component locations relative to the total simulation. Results show that MSE, Degree of Freedom and Correctly Fitted Ratio are same for all methods, which indicate that they have converged to the same optimal solution. Thus, our proposed method achieves both time efficiency and estimation accuracy.

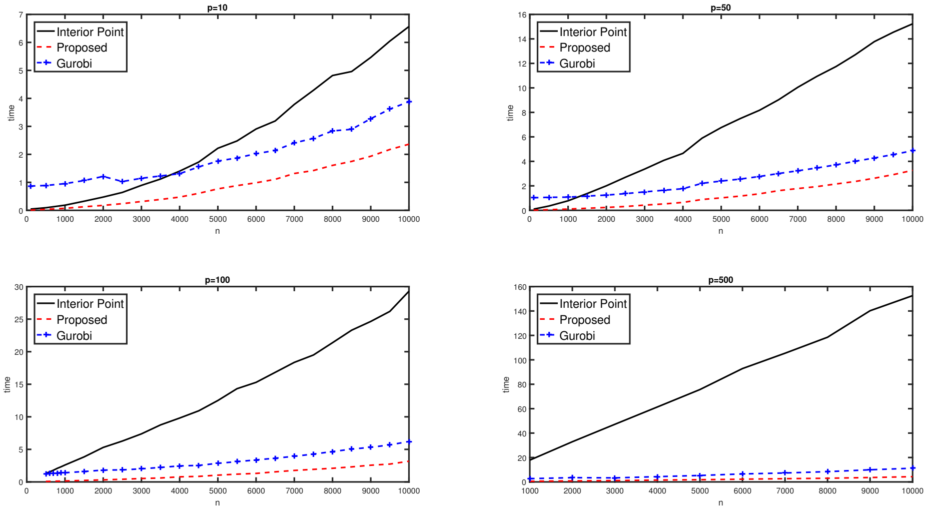

Figure 2.1: p=10,50,100,500 n–time plot

		TIME			MSE		
n	p	Interior Point	Proposed	Gurobi	Interior Point	Proposed	Gurobi
100	10	0.0456	0.0103	0.8684	0.0680	0.0680	0.0680
500	10	0.0928	0.0355	0.8877	0.0110	0.0110	0.0110
1000	10	0.1845	0.0722	0.9540	0.0055	0.0055	0.0055
2000	10	0.4740	0.1771	1.2126	0.0027	0.0027	0.0027
5000	10	2.2176	0.7675	1.7595	0.0010	0.0010	0.0010
10000	10	6.5670	2.3630	3.8824	0.0005	0.0005	0.0005

		Degree of Freedom			Correctly Fitted Ratio		
n	p	Interior Point	Proposed	Gurobi	Interior Point	Proposed	Gurobi
100	10	5.19	5.19	5.19	0.83	0.83	0.83
500	10	5.29	5.29	5.29	0.77	0.77	0.77
1000	10	5.31	5.31	5.31	0.74	0.74	0.74
2000	10	5.23	5.23	5.23	0.79	0.79	0.79
5000	10	5.25	5.25	5.25	0.76	0.76	0.76
10000	10	5.18	5.18	5.18	0.83	0.83	0.83

Table 2.1: Simulation results of $p = 10$.

		TIME			MSE		
n	p	Interior Point	Proposed	Gurobi	Interior Point	Proposed	Gurobi
100	50	0.0990	0.0182	1.0301	0.1044	0.1044	0.1044
1000	50	0.7803	0.1039	1.0755	0.0065	0.0065	0.0065
2000	50	2.0035	0.2363	1.2457	0.0033	0.0033	0.0033
5000	50	6.7602	1.0177	2.3907	0.0014	0.0014	0.0014
10000	50	15.2296	3.2660	4.8684	0.0007	0.0007	0.0007

		Degree of Freedom			Correctly Fitted Ratio		
n	p	Interior Point	Proposed	Gurobi	Interior Point	Proposed	Gurobi
100	50	5.21	5.21	5.21	0.80	0.80	0.80
1000	50	5.23	5.23	5.23	0.80	0.80	0.80
2000	50	5.35	5.35	5.35	0.69	0.69	0.69
5000	50	5.20	5.20	5.20	0.81	0.81	0.81
10000	50	5.25	5.25	5.25	0.81	0.81	0.81

Table 2.2: Simulation results of $p = 50$.

		TIME			MSE		
n	p	Interior Point	Proposed	Gurobi	Interior Point	Proposed	Gurobi
500	100	1.3169	0.0770	1.2476	0.0161	0.0161	0.0161
1000	100	2.5999	0.1487	1.4152	0.0081	0.0081	0.0081
2000	100	5.2977	0.3179	1.8193	0.0036	0.0036	0.0036
5000	100	12.5036	1.0418	2.8586	0.0015	0.0015	0.0015
10000	100	29.2864	3.1812	6.2155	0.0008	0.0008	0.0008

		Degree of Freedom			Correctly Fitted Ratio		
n	p	Interior Point	Proposed	Gurobi	Interior Point	Proposed	Gurobi
500	100	5.33	5.33	5.33	0.75	0.75	0.75
1000	100	5.17	5.17	5.17	0.85	0.85	0.85
2000	100	5.29	5.29	5.29	0.74	0.74	0.74
5000	100	5.25	5.25	5.25	0.77	0.77	0.77
10000	100	5.39	5.39	5.39	0.67	0.67	0.67

Table 2.3: Simulation results of $p = 100$.

		TIME			MSE		
n	p	Interior Point	Proposed	Gurobi	Interior Point	Proposed	Gurobi
1000	500	17.8460	0.5508	2.6924	0.0089	0.0089	0.0089
2000	500	32.9152	0.8235	3.4965	0.0047	0.0047	0.0047
5000	500	75.6648	1.8116	5.2965	0.0017	0.0017	0.0017
8000	500	118.5086	3.0154	8.4263	0.0011	0.0011	0.0011
10000	500	152.5622	4.3109	11.2252	0.0009	0.0009	0.0009

		Degree of Freedom			Correctly Fitted Ratio		
n	p	Interior Point	Proposed	Gurobi	Interior Point	Proposed	Gurobi
1000	500	5.26	5.26	5.26	0.77	0.77	0.77
2000	500	5.22	5.22	5.22	0.82	0.82	0.82
5000	500	5.23	5.23	5.23	0.77	0.77	0.77
8000	500	5.24	5.24	5.24	0.78	0.78	0.78
10000	500	5.31	5.31	5.31	0.72	0.72	0.72

Table 2.4: Simulation results of $p = 500$.

2.1.3 Real data analysis

Example 1: In the first example, we have selected 5 different real datasets for numerical experiment. Again, we compare our method with the interior point method and the Gurobi method. The datasets are as follows:

1. Prostate Cancer Data, which is studied by Stamey et al. [105] dealing with the correlation of 9 predictors and prostate specific antigen (lpsa).

2. Boston Housing Data, which is derived from Harrison and Rubinfeld [45] focussing on the 14 predictors that affect medv (median value of owner-occupied homes in \$1000s).

3. Bardet Data, which is the simplified gene expression data presented by Scheetz et al. [97], where design matrix X is a 120×100 matrix expanded from the expression levels of 20 filtered genes. The objective is to discover the correlation between 100 predictors and the expression level of gene TRIM32 that causes Bardet-Biedl syndrome.

4. Diabetes Data, which is studied by Efron [32] containing 442 patients with 10 clinical measures: age, sex, body mass index (bmi), average blood pressure

(map), and six blood serum measurements. The aim is to find the correlation between response y and the above 10 predictors.

5. China Stock Data, which considered by Wang [113] exploring the relationship of Return on Equity (ROE_{t+1}) and other 9 predictors.

Since all three methods found the same result, we focus on the execution time. Table 2.5 shows the running results for the 5 datasets:

Name	n	p	Interior Point	Proposed	Gurobi
Prostate Cancer	97	8	0.0243	0.0067	0.6855
Boston Housing	506	13	0.0878	0.0382	0.7414
Bardet	120	100	0.1304	0.0514	0.6988
Diabetes	442	10	0.5127	0.0113	0.6989
China Stock	1946	9	0.2632	0.1163	1.3954

Table 2.5: Time comparison for real datasets

For the 5 datasets, time comparison of Interior Point (IP) method, our proposed method and Gurobi are summarized in Table 2.5, again our proposed method is faster than other methods.

Example 2: In the second example, we apply the proposed methodology to China stock market data of AH premium as follows. China stock market participants are mostly retail investors with mental gambler, most of them have little experience, this leads to misvaluation and detachment from intrinsic value of companies. Hence it is attractive to look at stock prices of Shanghai Exchange (A share) and Hong Kong Exchange (H share). AH Premium is a great demonstration of this irrationality and education level of investors.

Chinese stock market is relatively young and come into the market since 1990, with two main stock markets: the Shanghai Stock Exchange and Shenzhen Stock Exchange. The Chinese stock market has boomed during the last several decades.

From Hang Seng index, the China stock market (A Share) and Hong Kong stock market (H share) listed 60 AH premium stock market in Table 2.2. Both connections allow China investors get access to Hong Kong and Hong Kong investors buy stocks listed in China.

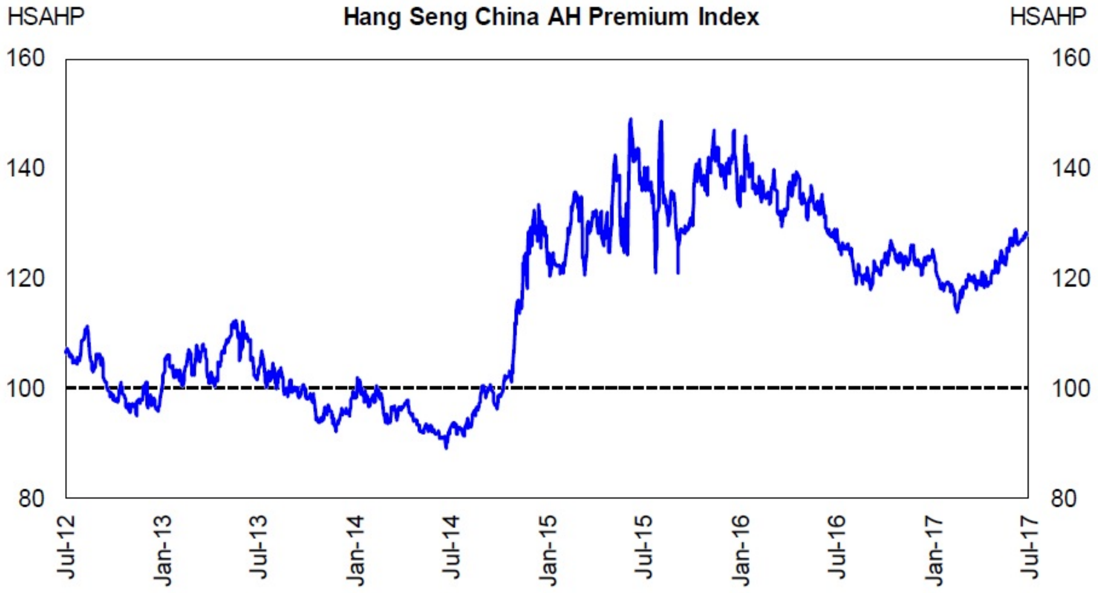

Figure 2.2: Heng Seng AH premium index

The difference in close price tendency are illustrated in Figure 2.2. The index takes value 100 yields equivalence between A share and H share, HSAHP above 100 indicate that A share is more expensive than H share and vice versa. There may be several reasons for trading AH premium is specifically different. International investors are deviate from the A share stocks, and made traditional convergence trades impossible; Mainland investors and international investors may get different information about China stock generally, and therefore they look at China stock market with different viewpoints. Many mainland investors treat H share as a type of western form of market since they can be traded freely with less control, whereas the China stock market is strictly controlled by government; H share surged more heavily than A share trading in Hong Kong with modern investmental strategies and techniques. On the other side, international investor doubted about Chinese firms and the government control management.

The weight proportions of different industries are illustrated in Figure 2.3, it is evident that the financial industries take up most percentage of the AH stock market, around 68.01%. Other medium level significant industries including Energy, Materials, Industrials, Consumer Goods, Consumer Services and Properties & Construction, the summation occupy nearly 28.1% of the stock market. The rest low level industries contain Telecommunications, Utilities, Information Technology and Conglomerates, totally sum up to 2.09%. Hence there are 11 main industry types in total.

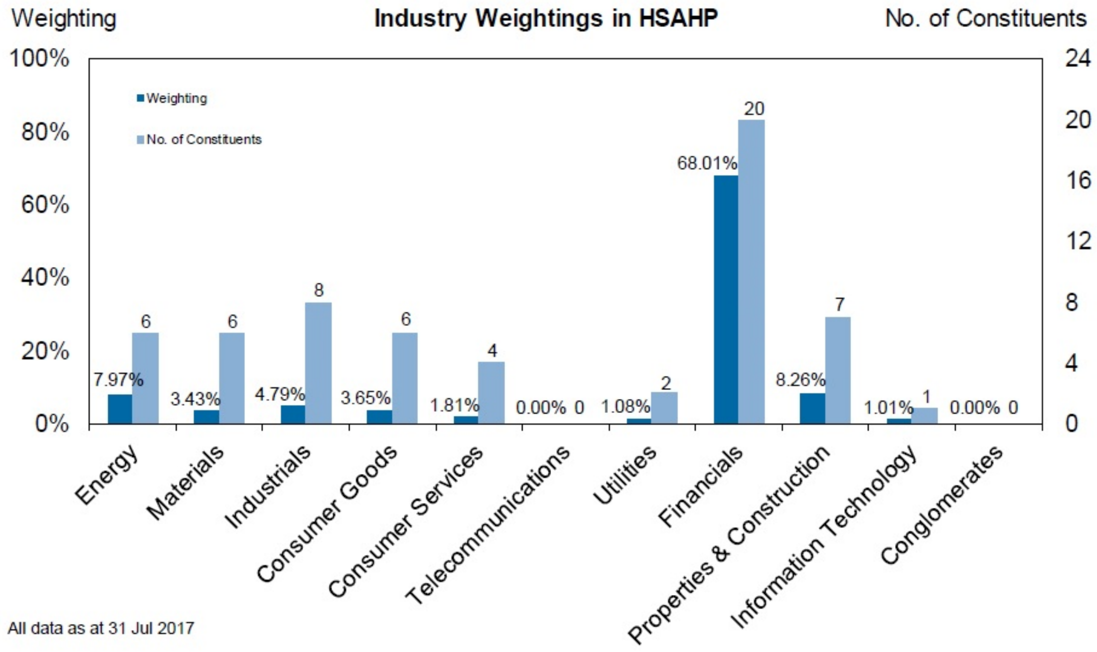

Figure 2.3: Heng Seng AH Index performance

Figure 2.3 introduced the main industry weightings in HSAHP market in detail, with the corresponding industry weighting proportions. Financial industries contain the most proportion of HSAHP, Energy, Material, Industrials, Consumer Goods, Properties & Construction are prominent industries in AH Premium, which shown in Figure 2.3 for detailed description. Table 2.7 displayed the complete 2017 AH premium index list from Hang Seng Index. Table 2.7 contain all the AH Premium stock index until 2017.

Similar to Wang [113], we consider Hang Seng 2017 AH premium Index for A Share and H Share respectively. Forecasting and prediction performances are based on the following response and predictor variables.

- Response: Return on Equity (ROE_{t+1}) of year $t + 1$ as the response variable;
- Predictors: Return on Equity (ROE_t), Asset Turnover Ratio (ATO), Profit Margin (PM), Debt to Asset Ratio or leverage (LEV), Sales Growth rate (GROWTH), Price-to-Book ratio (PB), Accounts Receivable/Revenues (AR-R), Inventories (INV) and Logarithm of Total Assets (ASSET) of year t.

We collect data from Bloomberg during 1/1/2012–12/31/2015, with the following A stock CH Equity and H stock HK Equity indexes in Table 2.6.

CH Equity Index									
600871	601727	601866	600685	600188	603993	601898	601857	601800	600115
601992	600688	601919	600362	601766	600027	600029	600011	601899	002202
600332	601633	601186	601111	601333	601600	601607	000157	000063	600028
600196	000002	002594	601088	000898	000338	600660	600585		

HK Equity Index									
1033	2727	2866	1171	3993	1898	1800	2009	1919	1766
1071	1055	2899	2208	2333	1186	2600	2607	1157	2196
1211	1088	2338							

Table 2.6: AH Stock Index

No.	Company Name	A Stock	H Stock	Industry	A/H Price Ratio
1	Sinopec SSC	600871	1033	Energy	308.90
2	SH Electric	601727	2727	Industrials	246.87
3	COSCO Ship Dev	601866	2866	Industrials	244.71
4	DFZQ	600958	3958	Financials	233.30
5	COMEC	600685	0317	Industrials	224.14
6	MCC Properties	601618	1618	Construction	221.80
7	Yanzhou Coal	600188	1171	Energy	215.25
8	CMOC	603993	3993	Materials	197.47
9	China Coal	601898	1898	Energy	190.36
10	PetroChina	601857	0857	Energy	187.62
11	China Comm Cons	601800	1800	Properties & Construction	183.81
12	China East Air	600115	0670	Consumer Services	182.33
13	BBMG Properties	601992	2009	Construction	179.72
14	Shanghai Pechem	600688	0338	Materials	178.56
15	COSCO SHIP Hold	601919	1919	Industrials	174.00
16	Jiangxi Copper	600362	0358	Materials	168.75
17	CMSC	600999	6099	Financials	168.47
18	CRRC	601766	1766	Industrials	167.63
19	EB Securities	601788	6178	Financials	167.44
20	China Railway	601390	0390	Properties & Construction	165.35
21	Huadian Power	600027	1071	Utilities	161.10
22	China South Air	600029	1055	Consumer Services	156.78
23	Huaneng Power	600011	0902	Utilities	154.63
24	Zijin Mining	601899	2899	Materials	152.99
25	Goldwind	002202	2208	Industrials	152.97
26	CITIC Bank	601998	0998	Financials	152.25
27	Baiyunshan Ph	600332	0874	Consumer Goods	152.03
28	GreatWall Motor	601633	2333	Consumer Goods	150.60
29	HTSC	601688	6886	Financials	147.13
30	China Rail Cons	601186	1186	Properties & Construction	147.01
31	Air China	601111	0753	Consumer Services	145.35
32	Guangshen Rail	601333	0525	Consumer Services	143.51
33	CHALCO	601600	2600	Materials	141.66
34	Sh Pharma	601607	2607	Consumer Goods	139.97
35	Zoomlion	000157	1157	Industrials	139.70
36	Haitong Sec	600837	6837	Financials	138.53
37	ZTE	000063	0763	Information Technology	132.54
38	NCI	601336	1336	Financials	132.21
39	GF Sec	000776	1776	Financials	130.74
40	CEB Bank	601818	6818	Financials	129.55
41	China Life	601628	2628	Financials	129.45
42	Bankcomm	601328	3328	Financials	128.31
43	Minsheng Bank	600016	1988	Financials	127.01
44	CITIC Sec	600030	6030	Financials	126.98
45	CCB	601939	0939	Financials	120.36
46	Sinopec Corp	600028	0386	Energy	119.60
47	Fosun Pharma	600196	2196	Consumer Goods	118.31
48	Bank of China	601988	3988	Financials	118.19
49	ABC	601288	1288	Financials	117.99
50	China Vanke	000002	2202	Properties & Construction	117.69
51	ICBC	601398	1398	Financials	116.93
52	BYD Company	002594	1211	Consumer Goods	116.79
53	CM Bank	600036	3968	Financials	115.49
54	China Shenhua	601088	1088	Energy	115.24
55	Angang Steel	000898	0347	Materials	114.98
56	CPIC	601601	2601	Financials	114.97
57	Weichai Power	000338	2338	Industrials	108.06
58	Ping An	601318	2318	Financials	104.20
59	Fuyao Glass	600660	3606	Consumer Goods	101.27
60	Conch Cement	600585	0914	Properties & Construction	97.48

Table 2.7: The details about AH Stock Index

Following Wang [114], we consider regression models with tuning parameters $\lambda_1 = \sqrt{2n\log p}, \lambda_2 = \sqrt{n\log p}, \lambda_3 = \frac{\sqrt{n\log p}}{2}, \lambda_4 = \frac{\sqrt{n\log p}}{4}$, and consider the following regression models:

- LS (Least Square): $\arg\min_\beta \|Y - X\beta\|^2$ with explicit solution $\hat{\beta} = (X^\intercal X)^{-1} X^\intercal Y$.
- LAD (Least Absolute Deviation): $\arg\min_\beta \|Y - X\beta\|_1$.
- LS-Lasso: $\hat{\beta} = \arg\min_\beta \|Y - X\beta\|^2 + \lambda\|\beta\|_1$.
- LAD-Lasso: $\hat{\beta} = \arg\min_\beta \|Y - X\beta\|_1 + \lambda\|\beta\|_1$.

The LAD-Lasso model can be reformulated as linear programming, existing solvers include Interior Point method (IP), Dual Simplex (DS), and our proposed algorithm (Section 2.1) with detailed description in Shi et al. [101].

We consider two datasets: (1) Training data as A-Stock close price of Year 2012; Test Data as A-Stock close price of Year 2013; (2) Training data as H-Stock close price of Year 2012; Test Data as H-Stock close price of Year 2013, then we compare time consumption of Interior Point (IP) and the proposed descent algorithm, we evaluate the prediction performance through the following measures:

- R^2: $\mathrm{R}^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$;
- Degree of Freedom (DF): DF = number of nonzero elements of β; (see Zou [137]);
- Mean Absolute Percentage Error (MAPE): MAPE $= \sum_{t=1}^{T} \left|\frac{y_t - \hat{y}_t}{y_t}\right|$.

Consider dataset of A stock and H stock during 1/1/2012–12/31/2013. Model fitting results are displayed in Table 2.8.

	Linear Regression		LS-Lasso		LAD-Lasso			
	LS	LAD	LS-aic	LS-bic	λ_1	λ_2	λ_3	λ_4
Train: A2012; Test: A2013								
ROE	0.8167	0.7663	0.8167	0.8167	0.7423	0.7455	0.7861	0.7184
ATO	4.2406	6.8013	4.2406	4.2406	0.0000	0.0000	1.1567	2.1259
PM	-0.0026	0.0896	-0.0026	-0.0026	-0.0000	-0.0000	-0.0000	0.0794
LEV	0.0229	-0.0400	0.0229	0.0229	0.0000	-0.0000	-0.0067	-0.0242
GRO	-0.0695	-0.0638	-0.0695	-0.0695	0.0278	0.0270	-0.0147	-0.0120
PB	1.9227	0.8930	1.9227	1.9227	0.0000	0.0000	0.6817	1.0352
ARR	-0.0001	-0.0001	-0.0001	-0.0001	-0.0000	-0.0000	-0.0000	-0.0001
INV	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ASS	-0.4858	-0.2763	-0.4858	-0.4858	0.0916	0.0962	0.0000	-0.0782
R^2	0.8086	0.7825	0.8086	0.8086	0.7217	0.7229	0.7718	0.7820
TIME	0.0130	0.0417	0.0741	0.0258	0.0452	0.0432	0.0435	0.0433
DF	7	8	7	7	3	3	5	7
MAPE	2.2927	1.9268	2.2927	2.2927	1.8433	1.8616	2.0953	2.0126
Train: H2012; Test: H2013								
ROE	0.7418	0.6732	0.8024	0.8024	0.8648	0.8879	0.8039	0.8034
ATO	4.1722	3.0442	0.0000	0.0000	0.0000	0.0000	0.0000	0.2281
PM	-0.0467	-0.1664	-0.1649	-0.1649	-0.0161	-0.1087	-0.2603	-0.2529
LEV	-0.1676	-0.2501	-0.2352	-0.2352	-0.0564	-0.0875	-0.2795	-0.2712
GRO	-0.0910	-0.0408	-0.0685	-0.0685	-0.0538	-0.0401	-0.0629	-0.0634
PB	2.0473	2.8224	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ARR	-0.0003	-0.0005	-0.0004	-0.0004	-0.0002	-0.0002	-0.0005	-0.0005
INV	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
ASS	0.1180	0.3766	0.8336	0.8336	0.3585	0.4903	1.1061	1.0665
R^2	0.9061	0.8922	0.8913	0.8913	0.8434	0.8492	0.8858	0.8862
TIME	0.0128	0.0365	0.0742	0.0278	0.0440	0.0426	0.0423	0.0420
DF	9	9	7	7	6	6	7	8
MAPE	1.8443	2.1205	2.9188	2.9188	2.7584	2.7312	2.9039	2.8491

Table 2.8: AH Stock 2012-2013

Table 2.8 indicate that with $\lambda = \lambda_1 = \sqrt{2n \log p}$ for Year 2012-2013 data, we can make a balance between MAPE and R^2. For A stock, ROE, GROWTH and ASSET are significant variables; For H stock, ROE, PM, LEV, GROWTH, ARR and ASSET are significant variables. Of all these estimation methods, our proposed method can achieve more accurate estimation results. Table 2.9 show that the simulation results are generally better than other methods.

	Linear Regression		LS-Lasso		LAD-Lasso			
	LS	LAD	LS-aic	LS-bic	λ_1	λ_2	λ_3	λ_4
Train: A2014; Test: A2015								
ROE	0.6950	0.8795	0.6950	0.6677	0.9077	0.9187	0.8644	0.8702
ATO	4.1606	-0.6577	4.1606	0.0000	0.0000	0.0000	0.0000	-0.0000
PM	0.0141	-0.2379	0.0141	0.0000	-0.1703	-0.1704	-0.2064	-0.2211
LEV	0.0774	0.0782	0.0774	0.0484	0.0314	0.0278	0.0611	0.0869
GRO	-0.0443	-0.0629	-0.0443	-0.0337	-0.0465	-0.0457	-0.0595	-0.0615
PB	1.1454	1.6788	1.1454	0.0000	0.0000	0.0000	1.2704	1.4356
ARR	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
INV	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ASS	-0.5575	-0.4004	-0.5575	0.0000	0.0000	-0.0000	-0.3107	-0.4160
R^2	0.6629	0.5575	0.6629	0.5715	0.5405	0.5366	0.5817	0.5765
TIME	0.0137	0.0357	0.0706	0.0272	0.0455	0.0425	0.0426	0.0413
DF	7	7	7	3	4	4	6	6
MAPE	1.3236	1.5640	1.3236	1.8097	2.0931	2.0703	1.6854	1.5699
Train: H2014; Test: H2015								
ROE	0.4640	0.4496	0.4737	0.4737	0.4106	0.4463	0.4632	0.4301
ATO	1.4082	3.0138	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PM	-0.0237	-0.0270	-0.0527	-0.0527	-0.0000	-0.0625	-0.0657	-0.0503
LEV	0.0022	-0.0367	-0.0077	-0.0077	-0.0144	-0.0659	-0.0158	-0.0613
GRO	0.0428	0.0184	0.0472	0.0472	0.0347	0.0256	0.0278	0.0195
PB	3.0447	3.0404	3.0210	3.0210	0.0000	1.5327	1.9088	3.8030
ARR	-0.0001	-0.0001	-0.0001	-0.0001	-0.0000	-0.0001	-0.0000	-0.0000
INV	0.0000	0.0000	0.0001	0.0001	0.0000	0.0001	0.0000	0.0000
ASS	-0.2658	-0.2096	-0.1514	-0.1514	0.1397	0.2482	0.0000	-0.0534
R^2	0.8007	0.7845	0.7984	0.7984	0.6618	0.7617	0.7690	0.7741
TIME	0.0128	0.0398	0.0714	0.0296	0.0527	0.0492	0.0461	0.0422
DF	7	7	6	6	4	6	5	6
MAPE	0.8307	0.8128	0.8258	0.8258	0.7402	0.6967	0.7544	0.8306

Table 2.9: AH Stock 2012-2013

Table 2.9 report results of Year 2014-2015. For A stock, ROE, PM, LEV, GROWTH are significant variables for LAD-Lasso; For H stock, ROE, PM, LEV, GROWTH, PB and ASSET are significant variables for prediction of ROE_{t+1}.

Table 2.8, 2.9 show that with proper tuning parameter, LAD-Lasso can achieve a tradeoff between R^2 and MAPE, and suggest that ROE, PM, LEV, GRO, PB, ASSET are significant variables that affect prediction and forecasting.

2.2 New active zero set descent algorithm for LAD Generalized Lasso

Consider LAD Generalized Lasso problem,

$$\arg\min_{\theta\in\mathbb{R}^p} \|Y - X\theta\|_1 + \lambda\|R\theta\|_1, \tag{2.18}$$

where $X \in \mathbb{R}^{n\times p}$ is the design matrix with row vectors $X_i, i = 1, 2, \cdots, n$, $Y = (y_1, \cdots, y_n)' \in \mathbb{R}^n$ is the response vector, $R \in \mathbb{R}^{q\times p}$ is the constraint matrix, $\theta \in \mathbb{R}^p$ is the coefficient vector, and $\lambda > 0$ is the tuning parameter.

Denote

$$Y^* = \begin{pmatrix} Y \\ 0 \end{pmatrix}, X^* = \begin{pmatrix} X \\ \lambda R \end{pmatrix}, n^* = n + q. \tag{2.19}$$

Problem (2.18) becomes

$$\arg\min_{\theta\in\mathbb{R}^p} \|Y^* - X^*\theta\|_1.$$

For convenience, the superscript $*$ in n, X, Y in (2.42) are dropped. Hopefully, there is no confusion. Then, LAD Generalized Lasso problem becomes

$$\arg\min_{\theta\in\mathbb{R}^p} \|Y - X\theta\|_1 = \arg\min_{\theta\in\mathbb{R}^p} \sum_{i=1}^{n} |X_i\theta - y_i|.$$

Denote $u_i = X_i\theta - y_i$.

Optimality conditions and directional derivatives

Consider optimization problem

$$\min_{\theta\in\mathbb{R}^p} f(\theta) = \min_{\theta\in\mathbb{R}^p} \sum_{i=1}^{n} f_i(\theta) = \arg\min_{\theta\in\mathbb{R}^p} \sum_{i=1}^{n} |X_i\theta - y_i|. \tag{2.20}$$

Note that the absolute value function is convex, so does $f(\theta)$. Hence, the local minimizer of $f(\theta)$ is the global minimizer, too. The optimality condition is that all directional derivatives are greater than or equal to zero. For a function $g(\theta)$, the directional derivative of g along a direction d is defined as

$$D_d g(\theta) = \lim_{t \to 0^+} \frac{g(\theta + td) - g(\theta)}{t\|d\|} \,. \tag{2.21}$$

Below, we highlight the main difference between the directional derivatives of smooth functions and non-smooth absolute value function. It is well-known that when g is differentiable,

$$D_d g(\theta) = \frac{\partial g}{\partial \theta} \cdot d \,, D_{d^-} g(\theta) = \frac{\partial g}{\partial \theta} \cdot (-d).$$

Hence, we have

$$D_{d^-} g(\theta) = -D_d g(\theta). \tag{2.22}$$

However, this is not true when g is non-differentiable absolute value function. Consider the special case of f with $n = 1$, that is, the absolute value function

$$f = |b^{\intercal}\theta + c| \,.$$

Given a direction d and a point θ fulfilling $b^{\intercal}\theta + c = 0$, we have

$$\begin{aligned} D_d f(\theta) &= \lim_{t \to 0^+} \frac{|b^{\intercal}(\theta + td) + c| - |b^{\intercal}\theta + c|}{t\|d\|} \\ &= \lim_{t \to 0^+} \frac{|tb^{\intercal}d|}{t\|d\|} = \frac{|b^{\intercal}d|}{\|d\|}, \\ D_{d^-} f(\theta) &= \lim_{t \to 0^+} \frac{|b^{\intercal}(\theta - td) + c| - |b^{\intercal}\theta + c|}{t\|d\|} \\ &= \lim_{t \to 0^+} \frac{|-tb^{\intercal}d|}{t\|d\|} = \frac{|b^{\intercal}d|}{\|d\|}. \end{aligned}$$

Hence,

$$D_{d^-} f(\theta) = D_d f(\theta). \tag{2.23}$$

In the general cases $n > 1$, the directional derivative of the cost function f can be decomposed into two parts

$$D_d f(\theta) = A(\theta, d) + C(\theta, d),$$

where $A(\theta, d)$ and $C(\theta, d)$ are the smooth and non-smooth part of the directional derivative $D_d f(\theta)$ respectively, that is,

$$\begin{aligned} A(\theta, d) &= \sum_{i=1}^{n} \delta(u_i > 0) \cdot \frac{X_i d}{\|d\|} + \sum_{i=1}^{n} \delta(u_i < 0) \cdot \frac{(-X_i) d}{\|d\|}, \\ C(\theta, d) &= \sum_{i=1}^{n} \delta(u_i = 0) \cdot \frac{X_i d}{\|d\|}, \end{aligned}$$

where $\delta(\cdot)$ is the indicator function

$$\delta(\nu) = \begin{cases} 1, & \text{if } \nu \text{ is true,} \\ 0, & \text{otherwise.} \end{cases}$$

The vector θ^* is the optimal solution to (2.20) if and only if

$$D_d f(\theta^*) = A(\theta^*, d) + C(\theta^*, d) \geqslant 0 \,, \forall d \in \mathbb{R}^p. \tag{2.24}$$

Basis direction set and zero set

The optimality condition (2.24) involves infinitely many arbitrary directions d. To overcome the difficulties related to the infinite dimension, an equivalent finite representation of the optimality condition is derived based on the concepts of basis direction set and zero set introduced in what follows. For a given point θ, define the zero set $\Omega = \{i | u_i = 0, i = 1, \cdots, n\} = \{\omega_1, \omega_2, \cdots, \omega_m\}$, where m is the cardinality of Ω and ω_i are the indexes in $\{1, \cdots, n\}$. Thus,

$$X_{\omega_i} \theta = y_{\omega_i} \,, i = 1, \cdots, m. \tag{2.25}$$

Consider the set of directions along which the zero set remains unchanged in the proximity of θ. Basis direction set refers to the basis of such a direction set.

The basis directions $v_1, v_2, \ldots, v_p$ are constructed as follows. Let

$$X_a = \begin{pmatrix} X_{\omega_1} \\ \vdots \\ X_{\omega_m} \end{pmatrix}.$$

Without loss of generality, we assume that rank $(X_a) = m$. The generalized inverse of X_a is V_a with columns $\{v_j, j = 1, \cdots, m\}$ such that $X_a V_a = I_m$, where I_m is the $m \times m$ identity matrix. The remaining $p-m$ linear independent vectors $v_j, j = m+1, \cdots, p$ are defined as the basis of the null space $\{v \in \mathbb{R}^p | X_a v = 0\}$. Then, we have

$$X_a v_j = 0\,, \forall j = m+1, \cdots, p$$

and

$$C(\theta^*, v_j) = C(\theta^*, v_j^-) = 0\,, j = m+1, \cdots, p.$$

Therefore, $f(\theta^*)$ is smooth along with these directions $\{v_j | j = m+1, \cdots, p\}$. Moreover, orthogonality holds,

$$X_{\omega_i} v_j = \begin{cases} 1, & \text{when } i = j, \\ 0, & \text{when } i \neq j, \end{cases} \quad i = 1, \cdots, m\,; j = 1, \cdots, p. \tag{2.26}$$

The following theorem gives an equivalent finite-dimensional representation of the optimality condition (2.24).

Theorem 2.3. *Suppose that rank* $(X_a) = m$. *Then,* θ^* *is the optimal solution if and only if the directional derivatives satisfy*

$$D_{v_i} f(\theta^*) = A(\theta^*, v_i) + C(\theta^*, v_i) \geqslant 0\,, i = 1, \cdots, m, \tag{2.27}$$

$$D_{v_i^-} f(\theta^*) = A(\theta^*, v_i^-) + C(\theta^*, v_i^-) \geqslant 0\,, i = 1, \cdots, m, \tag{2.28}$$

$$D_{v_i} f(\theta^*) = A(\theta^*, v_i) = 0\,, i = m+1, \cdots, p. \tag{2.29}$$

Remark 2.2. *If rank* $(X_a) < m$, *it is natural to consider the maximal subset of linear independent vectors and redefine the cost function by removing some redundant terms.*

However, in both simulation studies and real data analysis later on in sections 3 and 4, we never come across with such situations. Therefore, the non-full-rank cases are not discussed in this paper. It can be an interesting future research topic to study the non-full-rank cases.

Proof. Note that (2.27) and (2.28) are special cases of (2.24). From (2.22), if $D_{v_i}f(\theta^*) \geqslant 0$, and $D_{v_i^-}f(\theta^*) \geqslant 0$, we have $D_{v_i}f(\theta^*) = 0$. Hence, (2.29) is a special case of (2.24). Thus, the necessary condition is obvious.

Next, we prove the sufficient condition. That means (2.24) holds if (2.27)-(2.29) are satisfied. Take $d = v_i\,, i = 1, \cdots, p$. By (2.26), we have

$$\begin{aligned} D_{v_i}f(\theta^*) &= A(\theta^*, v_i) = 0\,, i = m+1, \cdots, p, \\ C(\theta^*, v_i) &= \frac{\sum_{j=1}^m |X_{k_j} v_i|}{\|v_i\|} = \frac{1}{\|v_i\|}\,, i = 1, \cdots, m, \\ C(\theta^*, v_i^-) &= \frac{\sum_{j=1}^m |X_{k_j}(-v_i)|}{\|-v_i\|} = \frac{1}{\|v_i\|}\,, i = 1, \cdots, m. \end{aligned}$$

Then, (2.27)-(2.29) can be simplified as

$$\begin{aligned} D_{v_i}f(\theta^*) &= A(\theta^*, v_i) = \frac{av_i}{\|v_i\|} = 0\,, i = m+1, \cdots, p, \\ D_{v_i}f(\theta^*) &= A(\theta^*, v_i) + C(\theta^*, v_i) = \frac{av_i}{\|v_i\|} + \frac{1}{\|v_i\|} \geqslant 0\,, i = 1, \cdots, m, \qquad (2.30) \\ D_{v_i^-}f(\theta^*) &= A(\theta^*, v_i^-) + C(\theta^*, v_i^-) = \frac{-av_i}{\|v_i\|} + \frac{1}{\|v_i\|} \geqslant 0\,, i = 1, \cdots, m. \end{aligned}$$

Here,

$$a = \sum_{i=1}^n \delta(u_i > 0)X_i - \sum_{i=1}^n \delta(u_i < 0)X_i\,.$$

For any direction d, since $\{v_i | i = 1, \cdots, p\}$ is a basis of $\mathbb{R}^p$, there exists a vector

$\mu = (\mu_1, \cdots, \mu_p)$ such that

$$d = \sum_{i=1}^{p} \mu_i v_i. \tag{2.31}$$

Without loss of generality, we can set $\mu_i \geqslant 0\,, \forall i = 1, \cdots, p$, because if $\mu_i < 0$, we have $\mu_i v_i = (-\mu_i) \cdot v_i^-$. Then, v_i can be replaced by v_i^- and μ_i can be replaced by $-\mu_i > 0$. Hence, by adjusting the order adequately, (2.31) can be rewritten as

$$d = \sum_{j=1}^{m_1} \mu_j v_j + \sum_{j=m_1+1}^{m} \mu_j v_j^- + \sum_{j=m+1}^{p} \mu_j v_j,$$

where $\mu_j \geqslant 0\,, \forall j = 1, \cdots, p$. It follows from (2.30) that

$$\begin{aligned} C(\theta^*, d) &= \frac{\sum_{i=1}^{m} \left| X_{\omega_{ki}} d \right|}{\|d\|} \\ &= \frac{\sum_{i=1}^{m} \left| X_{\omega_{ki}} \big(\sum_{j=1}^{m_1} \mu_j v_j + \sum_{j=m_1+1}^{m} \mu_j v_j^- + \sum_{j=m+1}^{p} \mu_j v_j \big) \right|}{\|d\|} \\ &= \frac{\sum_{i=1}^{m_1} \left| \mu_i X_{\omega_{ki}} v_i \right| + \sum_{i=m_1+1}^{m} \left| \mu_i X_{\omega_{ki}} v_i^- \right|}{\|d\|} \\ &= \frac{\sum_{i=1}^{m} \mu_i}{\|d\|}. \end{aligned}$$

Hence, by (2.30), we have

$$\begin{aligned} D_d f(\theta^*) &= A(\theta^*, d) + C(\theta^*, d) \\ &= \left(\sum_{i=1}^{m_1} \mu_i a v_i + \sum_{i=m_1+1}^{m} \mu_i a v_i^- + \sum_{i=m+1}^{p} \mu_i a v_i + \sum_{i=1}^{m} \mu_i \right) \Big/ \|d\| \\ &= \left(\sum_{i=1}^{m_1} \mu_i (a v_i + 1) + \sum_{i=m_1+1}^{m} \mu_i (a v_i^- + 1) + \sum_{i=m+1}^{p} \mu_i a v_i \right) \Big/ \|d\| \\ &= \left(\sum_{i=1}^{m_1} D_{v_i} f(\theta^*) \cdot \|v_i\| + \sum_{i=m_1+1}^{m} D_{v_i^-} f(\theta^*) \cdot \|v_i\| \right) \Big/ \|d\| \\ &\geqslant 0\,. \end{aligned}$$

Thus, for any direction d, the directional derivative is greater than or equals to zero. Hence, (2.24) holds and θ^* is the optimal solution. □

2.2.1 Computational method

Proposition 2.2.1. *If conditions* (2.27)-(2.29) *are not all satisfied, there exists a direction d such that the cost function value decreases along the direction d.*

Proof. Clearly, if the i-th condition of (2.29) is violated, v_i^- is a descend direction. Suppose that the i-th condition of (2.27) or (2.28) is not satisfied. Then,

$$D_{v_i} f(\theta) \geqslant 0, \text{ and } D_{v_i^-} f(\theta) \geqslant 0$$

can not be satisfied at the same time. Consider the following three cases.

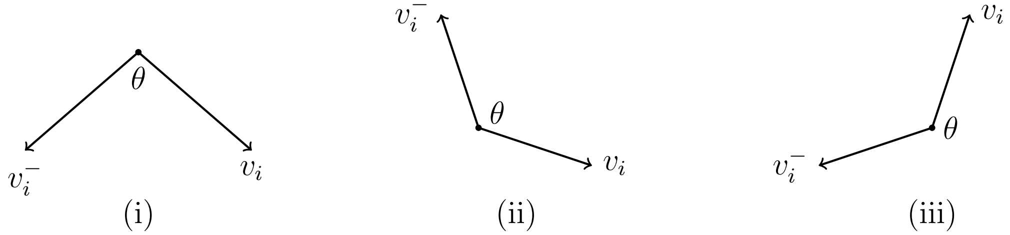

Figure 2.4: Three cases of descent direction

(i) $D_{v_i} f(\theta) < 0$ and $D_{v_i^-} f(\theta) < 0$.

If this case is true, then f is concave at the point θ along with the direction v_i, which contradicts the convexity of f.

(ii) $D_{v_i} f(\theta) < 0$ and $D_{v_i^-} f(\theta) \geqslant 0$.

It can be seen that

$$D_{v_i^-} f(\theta) \geqslant -D_{v_i} f(\theta).$$

Otherwise f becomes concave, which is a contradiction. Then, v_i is a descent direction.

(iii) $D_{v_i} f(\theta) \geqslant 0$ and $D_{v_i^-} f(\theta) < 0$.

It can be seen that

$$D_{v_i} f(\theta) \geqslant -D_{v_i^-} f(\theta).$$

Otherwise f becomes concave, which is a contradiction. Then, v_i^- is a descent direction.

Hence, the cost function value decreases along the directions either v_i or v_i^-. □

Descent direction search

To find the optimal point, we propose an algorithm so that the iterative points $\{\theta^{(k)}, k = 1, 2, 3, \cdots\}$ converge to the optimal solution and the corresponding cost function values are monotonic decreasing.

The algorithm is designed based on Theorem 2.3 and Proposition 2.2.1. If (2.27) and (2.28) hold for all $i = 1, 2, \ldots, m$ and (2.29) holds for all $i = m + 1, \cdots, p$, the optimal solution is found and the algorithm terminates. Otherwise, there are two mutually exclusive situations as follows.

First phase (steepest phase) refers to the situation where $m < p$ and Condition (2.29) is violated for some $i \in \{m + 1, \cdots, p\}$. In the steepest phase, choose a descent direction so that all zeros in Ω_k are kept while the smooth part is updated. That means the descent direction is selected from the space spanned by $\{v_i | i = m + 1, \cdots, p\}$. In a neighborhood of $\theta^{(k)}$, the cost function is

$$f(\theta^{(k)}) = a^{(k)}\theta^{(k)} + b^{(k)} + \sum_{\omega_{ki} \in \Omega_k} |X_{\omega_{ki}}\theta^{(k)} - y_{\omega_{ki}}|,$$

where

$$a^{(k)} = \sum_{i=1}^{n} \delta(u_i > 0)X_i - \sum_{i=1}^{n} \delta(u_i < 0)X_i\,, \quad b^{(k)} = \sum_{i=1}^{n} \delta(u_i > 0)y_i - \sum_{i=1}^{n} \delta(u_i < 0)y_i,$$

and Ω_k is the zero set index set of $\theta^{(k)}$ that defined as $\Omega_k = \{\omega_{ki} : u_{\omega_{ki}} = X_{\omega_{ki}}\theta^{(k)} - y_{\omega_{ki}} = 0\}$. The descent direction is chosen as the projection of $-a^{(k)}$ in $\{v_i | i = m+1, \cdots, p\}$.

Second phase (decreasement phase) refers to the situation where (a) $m = p$, or (b) $m < p\,, \forall i \in \{m+1, \cdots, p\}$, and Condition (2.29) is satisfied. Then, Conditions (2.27) and (2.28) are violated for some $i \in \{1, \cdots, m\}$. Otherwise, the algorithm has to be terminated. In the decreasement phase, some zeros in Ω_k are set free to avoid deadlock of the algorithm. Denote by Λ_1 the zeros to be set free. Λ_1 is chosen to contain the indexes ω_{ki} corresponding to the fastest descending directions v_i or v_i^-. Similar to the steepest phase, the steepest descent direction is chosen as the projection of $-a^{(k)}$ in $\{v_i | i = m+1, \cdots, p\}$.

In the steepest phase, set $\Omega_{0k} = \Omega_k$ and in the decreasement phase, set $\Omega_{0k} = \Omega_k \backslash \Lambda_1$. Then, Ω_{0k} contains zeros in Ω_k that we intend to keep. The projection h can be obtained by solving

$$\begin{aligned} \max_{h \in \mathbb{R}^p} \quad & -a^{(k)} h \\ \text{s.t.} \quad & X_i h = 0, \forall i \in \Omega_{0k}, \\ & \|h\| = 1. \end{aligned} \tag{2.32}$$

It means that the solution h is chosen as the direction nearest to the deepest descent direction $-a^{(k)}$ and along which the set Ω_{0k} remains unchanged in a neighborhood of $\theta^{(k)}$. The optimal solution to Problem (2.32) can be obtained by normalizing

$$\tilde{d} = -a^{(k)} - X_{0k}^{\intercal}(X_{0k}X_{0k}^{\intercal})^{-1}X_{0k} \cdot (-a^{(k)}), \tag{2.33}$$

where $X_{0k}^{\intercal}(X_{0k}X_{0k}^{\intercal})^{-1}X_{0k}(-a^{(k)})$ is the projected direction of $-a^{(k)}$ in the subspace $\{h | X_i h = 0, i \in \Omega_{0k}\}$ and

$$X_{0k} = \begin{pmatrix} X_{\omega_{k01}} \\ \vdots \\ X_{\omega_{k0m_0}} \end{pmatrix}, \omega_{k01}, \cdots, \omega_{k0m_0} \in \Omega_{0k}.$$

Then, the descent direction $d^{(k)}$ can be chosen as the normalized vector of $\tilde{d}$, that is,

$$d^{(k)} = \tilde{d}/\|\tilde{d}\| \,. \tag{2.34}$$

Proposition 2.2.2. *For any integers $k = 1, 2, \ldots,$ if the k-th iteration is in the first (steepest) phase, there exists $\ell > k$ such that the ℓ-th iteration is in the second (decreasement) phase.*

Proof. For any k-th iteration, $m < p$ and $A(\theta^{(k)}, v_i) \neq 0$ for at least one i in $\{m+1, \cdots, p\}$. Then, the number of zero set increases, because the indices in Ω_k will not be removed in first phase. Hence, by repeating the first phase, we obtain $m = p$, or $A(\theta^{(k)}, v_i) = 0 \,, \forall i = \{m+1, \cdots, p\}$, which will go to the second phase. □

Optimal step length

This section describes the procedure of determining the optimal step length for the descent direction in the previous section. It is shown that after moving, there is always new zero in the new point. Throughout this paper, Λ_2 refers to the set of indexes corresponding to the new zeros and Ω_{k+1} refers to the updated zero set. Without loss of generality, consider the following assumption.

Assumption 2.2.1. *The design matrix X is of full rank, i.e., $rank\,(X) = p$.*

Lemma 2.2. *For any θ, $d \in \mathbb{R}^p \,, d \neq 0 \,, \lim_{t\to\infty} f(\theta + td) = \infty \,.$*

Proof. According to Assumption 2.2.1, for any $d \in \mathbb{R}^p$, there exists at least one $i \in \{1\,, \cdots, n\}$ such that

$$X_i d \neq 0 \,.$$

Otherwise, there exists one $d \in \mathbb{R}^p$ such that $X_i d = 0 \,, \forall i \in \{1, \cdots, n\}$. Then the direction d is redundant for the space $\mathbb{R}^p$. For this, at least one dimension related to d can be reduced from the space $\mathbb{R}^p$.

Then, for any $\theta \in \mathbb{R}^p$,

$$\begin{aligned}\lim_{t\to\infty} f(\theta + td) &\geqslant \lim_{t\to\infty} |X_i(\theta + td) - y_i| \\ &\geqslant \lim_{t\to\infty} |tX_i d| - |X_i\theta - y_i| \\ &= \infty.\end{aligned}$$

□

Lemma 2.3. *If* $m = p\,, \theta^{(k)}$ *is unique.*

Proof. Since $m = p$, we have $X_a\theta^{(k)} = Y_a$, where

$$X_a = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}, Y_a = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix}.$$

By definition, X_a is the maximal subset of independent vector. Therefore, X_a is invertible and $\theta^{(k)} = X_a^{-1}Y_a$. Hence, $\theta^{(k)}$ is unique. □

Lemma 2.4. *If* $m < p$ *and* $A\,(\theta^{(k)}, v_i) = 0\,, \forall i = m+1, \cdots, p$, *then,* $f(\theta^{(k)})$ *is unique.*

Proof. Denote

$$\theta^{(k)} = \sum_{i=1}^{p} \mu_i v_i = V_a\mu_a + V_b\mu_b\,,$$

where

$$V_a = (v_1, \cdots, v_m)\,, V_b = (v_{m+1}, \cdots, v_p)\,, \mu_a \in \mathbb{R}^m\,, \mu_b \in \mathbb{R}^{p-m}\,.$$

If $\mu_b = 0$, then $\mu_a = X_aV_a\mu_a = X_a\theta^{(k)} = Y_a$ and $\theta^{(k)} = V_a\mu_a = V_aY_a = X_a^{\intercal}(X_aX_a^{\intercal})^{-1}Y_a$. Note that $A(\theta^{(k)}, v_i) = 0\,, \forall i \in \{m+1, \cdots, p\}$, we have $f(\theta^{(k)}+t^*v_i) = f(\theta^{(k)})\,, \forall t, \forall i = m+1, \cdots, p$. The cost function value will not changed along these directions. Hence, $f(\theta^{(k)})$ is unique and $\theta^{(k)}$ is in a subspace which contains the point $X_a^{\intercal}(X_aX_a^{\intercal})^{-1}Y_a$.

□

Theorem 2.4. *There exists an optimal solution for the line search problem of the descent direction. Let $t^* > 0$ be the optimal step size. Then, there is at least one i in $\{1, \cdots, n\}$ such that $X_i(\theta^{(k)} + t^* d^{(k)}) = y_i$. That means i is a new zero after moving to the point $\theta^{(k)} + t^* d^{(k)}$.*

Proof. For $(k+1)$-th iteration, the updating formula for θ is

$$\theta^{(k+1)} = \theta^{(k)} + td^{(k)}.$$

If $t = 0\,, d^{(k)}$ is a descent direction at $\theta^{(k)}$, that is,

$$D_{d^{(k)}} f(\theta^{(k)}) < 0\,.$$

Denote $g(t) = f(\theta^{(k)} + td^{(k)})$, since $g(t)$ is convex, $\frac{\partial g(t)}{\partial t}$ is monotonically increasing. Note that

$$\frac{\partial g(t)}{\partial t} = \lim_{\Delta t \to 0} \frac{g(\theta^{(k)} + (t + \Delta t)d^{(k)}) - g(\theta^{(k)} + td^{(k)})}{\Delta t} = \|d^{(k)}\| D_{d^{(k)}} f(\theta^{(k)} + td^{(k)}),$$

we have that

$$D_{d^{(k)}} f(\theta^{(k)} + td^{(k)})$$

is monotonically increasing.

From Lemma 2.2, if $t \to \infty\,$, then

$$\theta^{(k)} + td^{(k)} \to \infty, \text{ and } f(\theta^{(k)} + td^{(k)}) \to \infty.$$

Hence, we must have

$$D_{d^{(k)}} f(\theta^{(k)} + td^{(k)}) > 0\,.$$

If t is sufficiently large, then there exists one t such that

$$D_{d^{(k)}} f(\theta^{(k)} + td^{(k)}) \geqslant 0\,, D_{d^{(k)}-} f(\theta^{(k)} + td^{(k)}) \geqslant 0,$$

If

$$D_{d^{(k)}} f(\theta^{(k)} + td^{(k)}) \neq 0, \text{ or } D_{d^{(k)}-} f(\theta^{(k)} + td^{(k)}) \neq 0,$$

then the gradient does not exist at this point and there exists at least one i in $\{1, \cdots, n\}$ such that i is in the zero set, that is,

$$X_i(\theta^{(k)} + td^{(k)}) = 0\,.$$

If

$$D_{d^{(k)}} f(\theta^{(k)} + td^{(k)}) = D_{d^{(k)}-} f(\theta^{(k)} + td^{(k)}),$$

it can be seen that

$$D_{d^{(k)}} f(\theta^{(k)} + td^{(k)}) = 0\,.$$

Then, we move the point along with $d^{(k)}$ until we find one i in $\{1, \cdots, n\}$ such that

$$X_i(\theta^{(k)} + td^{(k)}) - y_i = 0$$

and $t^{(k)}$ can be chosen as t_1 or t_2 in Figure 2.5.

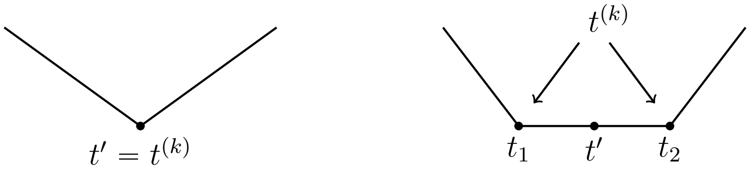

Figure 2.5: plot of step length t

Since the gradient is zero, the cost function value does not change, and the corresponding $t^{(k)}$ is still optimal. □

The line search problem in Theorem 2.4 can be solved as follows. Consider

$$\min_{t \geqslant 0} g(t),$$

where

$$g(t) = f(\theta^{(k+1)}) = f(\theta^{(k)} + td^{(k)})\,, t \geqslant 0\,.$$

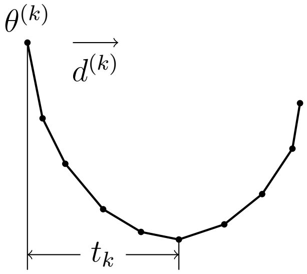

Figure 2.6: Plot of stepwise descent direction of $d^{(k)}$.

Since f is convex, $g(t)$ is convex, too, as depicted in Figure 2.6. Therefore, the line search problem is equivalent to

$$\begin{aligned} \max_{t \geqslant 0} \quad & t \\ \text{s.t.} \quad & D_{d^{(k)}} f(\theta^{(k)} + t d^{(k)}) \geqslant 0, \\ & D_{d^{(k)-}} f(\theta^{(k)} - t d^{(k)}) \geqslant 0. \end{aligned} \tag{2.35}$$

It suffices to consider the points $t_i = -(X_i \theta^{(k)} - y_i)/X_i d^{(k)}$ for $i = 1, 2, \ldots, n$. Note that $X_i(\theta^{(k)} + t_i d^{(k)}) - y_i = 0$. Sort the values t_i fulfilling $t_i > 0$ but $t_i \neq \infty$ in the ascending order and denote the sorted series by $0 < t_{k1} \leqslant t_{k2} \leqslant \ldots$ and define $\tau_{k1}, \tau_{k2}, \ldots$ as the indexes corresponding to $0 < t_{k1} \leqslant t_{k2} \leqslant \ldots$. The optimal step length is t_{ks} that fulfills the following optimality conditions,

$$D_{d^{(k)}} f(\theta^{(k)} + t_{ks} d^{(k)}) \geqslant 0\,, D_{d^{(k)}} f(\theta^{(k)} + t_{k,s-1} d^{(k)}) < 0\,, \tag{2.36}$$

where the direction derivatives can be obtained recursively via

$$\begin{aligned} D_{d^{(k)}} f(\theta^{(k)} + t_{ks} d^{(k)}) &= D_{d^{(k)}} f(\theta^{(k)} + t_{k,s-1} d^{(k)}) + 2|X_{\tau_{ks}} d^{(k)}|\,, s = 2, 3, \cdots, \\ D_{d^{(k)}} f(\theta^{(k)} + t_{k1} d^{(k)}) &= D_{d^{(k)}} f(\theta^{(k)}) + 2|X_{\tau_{k1}} d^{(k)}|\,. \end{aligned} \tag{2.37}$$

Denote the set of all such indexes τ_{ks} by Λ_2. Then, a updated zero set in the $(k+1)$-th iteration is defined as

$$\Omega_{k+1} = \Omega_{0k} \cup \Lambda_2\,.$$

By moving with step length t_{ks}, the cost function becomes

$$f(\theta^{(k+1)}) = (a^{(k+1)})\theta^{(k+1)} + \sum_{\omega_{k+1,i} \in \Omega_{k+1}} |X_{\omega_{k+1,i}} \theta^{(k+1)}| + b^{(k+1)}\,.$$

Theorem 2.5. *Algorithm 1 below terminates in a finite number of steps.*

Proof. For each step, the direction is a descent direction. Then, if the optimal condition is not satisfied, we can find a descent direction $d^{(k)}$ and the cost function value is reduced. Furthermore, the algorithm will stop in a point $\theta^{(k)}$ in each step, where there exists at least one index i such that $X_i\theta^{(k)} - y_i = 0$, that is, i is in the zero set. Hence, except the initial point, the zero set at any point stop in each step is not empty.

Note that there are totally n terms, which is finite, hence the number of the zero set in each step is finite. And the number of the zero set in Decreasement phase is also finite. For each step in second phase, $m = p$, or $m < p\,, A(\theta^{(k)}, v_i) = 0\,, \forall i = m+1, \cdots, p$. By Lemma 2.3 and Lemma 2.4, $\theta^{(k)}$ is unique corresponding to each zero set. Note that $f(\theta^{(k)})$ is monotonically decreasing which will not be repeated. The zero set series will not be repeated. Since the zero set is finite, the algorithm will stop in finite steps until the optimal conditions are satisfied. $\square$

Algorithm 2 Active Zero Set Descent Algorithm

Initialization: Choose an initial point $\theta^{(0)}$. Compute the corresponding set Ω_0 and the cost function value $f(\theta^{(0)})$. Set $k = 0$.

Step 1: (Terminate)
Generate the matrix V for the zero set Ω_k as described in subsection 2.2. If condition (2.27)–(2.29) are satisfied, then stop and return the optimal solution and value. Otherwise, if $m < p$ and there exists one i, $A(\theta^{(k)}, v_i) \neq 0$, go to Step 2, else go to Step 3.

Step 2: (Steepest phase) Set $\Omega_{0k} = \Omega_k$. Set the descent direction as $d^{(k)} = \text{Proj}\{-a^{(k)}|v_i, i = m+1, \cdots, p\}$ described in subsection 2.2.1. Go to Step 4.

Step 3: (Decreasement phase)
Find Λ_1, set $\Omega_{0k} = \Omega_k \backslash \Lambda_1$, and compute the descent direction $d^{(k)}$ by using (2.33) , (2.34). Go to Step 4.

Step 4: Find the best step length t^* by (2.35) and update $\theta^{(k+1)} = \theta^{(k)} + t^* d^{(k)}$. Find Λ_2 using the method described in Subsection 2.2.1 and update the zero set as $\Omega_{k+1} = \Omega_{0k} \cup \Lambda_2$. Then, compute the cost function $f(\theta^{(k+1)})$ and go to Step 1 for the $(k+1)$-th iteration.

2.2.2 Numerical experiments

In this part, we conduct extensive simulation studies and real data analysis to evaluate the estimation performance of our new active zero set descent algorithm compared to interior point method for LAD Fused Lasso model.

Simulation studies

Consider LAD Fused Lasso problem

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{\bar{n}} |\bar{y}_i - \bar{X}_i \beta| + \lambda \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|. \tag{2.38}$$

where $\bar{X} \in \mathbb{R}^{\bar{n} \times p}$ is the design matrix with row vectors $\bar{X}_i$, $\bar{y} = (\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_{\bar{n}})' \in \mathbb{R}^{\bar{n}}$ is the response vector, $\lambda > 0$ serves as the tuning parameter, and $\beta \in \mathbb{R}^p$ is the coefficient vector.

Problem (2.38) is a special case of the LAD Generalized Lasso problem (2.18) with penalty matrix

$$R = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}.$$

Interior point method for LAD Fused Lasso

Consider LAD Fused Lasso problem (2.38) and reparameterize β as

$$\theta_1 = \beta_1, \theta_2 = \beta_2 - \beta_1, \cdots, \theta_p = \beta_p - \beta_{p-1},$$

where $\theta = (\theta_1, \theta_2, \cdots, \theta_p)' \in \mathbb{R}^p$. The links between θ and β are given as

$$\theta = W\beta\,, \beta = M\theta,$$

where

$$W = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}, M = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}. \tag{2.39}$$

Then, Problem (2.38) becomes

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{\bar{n}} |\bar{y}_i - \bar{X}_i M\theta| + \lambda \sum_{j=2}^{p} |\theta_j|. \tag{2.40}$$

This can be viewed as a LAD-Lasso problem. Denote $H = \bar{X}M$, with j-th column as H_j. Following Wang et al. [116], we normalize the design matrix H such that $\|H_j\|_2^2 = n$, $j = 1, 2, \cdots, n$, and choose $\lambda = \sqrt{2n \log p}$. Let ϵ^+ , ϵ^- , θ^+ , and θ^- be the positive parts and negative parts of $\bar{y} - H\theta$ and θ respectively. Then,

$$\|\bar{y} - H\theta\|_1 = \varepsilon^+ + \varepsilon^-\,, \|\theta\|_1 = \theta^+ + \theta^-$$

and

$$\bar{y} - H\theta = \varepsilon^+ - \varepsilon^- \,, \theta = \theta^+ - \theta^-.$$

Problem (2.40) can therefore be rewritten as

$$\begin{aligned} \arg\min_{\theta} \quad & \sum_{i=1}^{\bar{n}} \varepsilon_i^+ + \sum_{i=1}^{\bar{n}} \varepsilon_i^- + \lambda(\sum_{j=2}^{p} \theta_j^+ + \sum_{j=2}^{p} \theta_j^-), \\ & H\theta + \varepsilon^+ - \varepsilon^- = \bar{y}, \\ & \theta - \theta^+ + \theta^- = 0_p, \\ & \varepsilon^+ \,, \varepsilon^- \geqslant 0_{\bar{n}} \,, \theta^+ \,, \theta^- \geqslant 0_p. \end{aligned} \tag{2.41}$$

The above problem can be solved using the state-of-the-art linear programming solver, interior point method that is available in Matlab function `linprog`.

Alternatively, denote

$$Y = \begin{pmatrix} \bar{y} \\ 0_{p-1} \end{pmatrix} \triangleq \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} & \bar{X} \\ 0 & \lambda I_{p-1} \end{pmatrix} \triangleq \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}, n \triangleq \bar{n} + p - 1. \tag{2.42}$$

Problem (2.40) becomes

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \|Y - X\theta\|_1, \tag{2.43}$$

which is a LAD regression problem. The solution θ can be obtained using active zero set descent algorithm (see Section 2.2.1). Then, β can be estimated by the transformation $\hat{\beta} = M\hat{\theta}$ with M defined in (2.39). Equivalently, we can get $\hat{\beta}$ back as

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_1 + \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_1 + \hat{\theta}_2 + \cdots + \hat{\theta}_p \end{pmatrix}.$$

Then we examine the performance of active zero set descent algorithm (LAD-AZSD, see Section 2.2) and interior point method (LAD-IP) for LAD Fused Lasso problem under different regression models. The experiments are performed on an Intel(R) Core(TM) i7-4790 CPU 3.60GHz processor and the algorithms are implemented in Matlab. For each dataset, we choose $\lambda = \sqrt{2n \log p}$ as suggested in [114, 116, 25].

Experiment 1. In this example, we study the effects of five factors, namely (i) the correlation in the covariates X, (ii) time-varying pattern in the coefficient β, (iii) variance of the error distribution, (iv) sample size n, and (v) the number of covariates p. Consider a $3 \times 2 \times 2 \times 5 \times 2$ experiment design. The covariates are generated from Gaussian distribution $\mathcal{N}_p(\mu, \Sigma)$ with correlation matrix $\Sigma = \rho^{|i-j|}\,, i, j = 1, 2, \cdots, p$. Three levels of ρ are considered, $\rho = 0\,, 0.1\,, 0.5$. The response vector is generated by

$$Y = X\beta + \sigma\varepsilon, \tag{2.44}$$

where $\{\varepsilon_i\}_{1 \leqslant i \leqslant \bar{n}}$ are independent standard Normal random variables. Two levels of σ are considered, $\sigma = 1\,, 3$. Consider two time-varying patterns of β, namely

$$\texttt{Case 1} \qquad \beta^* = (\underbrace{5, \cdots, 5}_{1-5}, \underbrace{0, \cdots, 0}_{6-10}, \underbrace{2, \cdots, 2}_{11-15}, \underbrace{0, \cdots, 0}_{16-p}) \in \mathbb{R}^p,$$

$$\texttt{Case 2} \qquad \beta^{**} = (\underbrace{5, 4.5, 4, 3.5, 3, 2.5, 2, 1.5, 1, 0.5}_{1-10}, \underbrace{0, \cdots, 0}_{11-p}) \in \mathbb{R}^p,$$

Case 1 is sparse and blocky while Case 2 is sparse and smooth. Choose p and n, with $p = 50\,, 100$, and n varying from 1000 to 5000. The above simulation settings are similar to those in [109, 3]. For each combination of levels, the experiment is repeated for 100 datasets.

The performances are measured in terms of the averaged run-time (`TIME`), averaged number of nonzero entries (`DF`: see [137]), and averaged mean absolute deviation

(`MAD`). The simulation results are summarized in Table 2.10-2.13. We have the following observations:

(i) For different $n\,,p$ values, AZSD perform considerably faster than interior point method for all cases. Under strong correlation settings, time superiority of AZSD is slightly weakened. Figure 2.7,2.8 illustrate that both algorithms have computation time approximately linear in n.

(ii) The results of `MAD` and `DF` suggest that the new algorithm does not lose accuracy comparing to the interior point algorithm. Both algorithms gives estimated β that are close to their true values. Also, estimation accuracy increased as n increases, suggesting estimation consistency.

	`TIME`		`MAD(sd)`		`DF`	
n	LAD-IP	LAD-AZSD	LAD-IP	LAD-AZSD	LAD-IP	LAD-AZSD
$\mathbf{p = 50}$						
Case 1: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.$						
1000	0.2308	0.1266	0.0223(0.0024)	0.0223(0.0025)	10.2	10.2
2000	0.5205	0.2949	0.0151(0.0017)	0.0151(0.0017)	10.3	10.2
3000	0.8463	0.5170	0.0120(0.0013)	0.0121(0.0013)	10.1	10.1
4000	1.2325	0.7968	0.0104(0.0012)	0.0104(0.0012)	10.1	10.1
5000	1.7181	1.0690	0.0094(0.0012)	0.0094(0.0012)	10.2	10.2
Case 2: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.1.$						
1000	0.2260	0.1289	0.0193(0.0026)	0.0193(0.0026)	10.2	10.2
2000	0.5074	0.2942	0.0127(0.0014)	0.0127(0.0015)	10.2	10.2
3000	0.8633	0.5257	0.0106(0.0014)	0.0106(0.0014)	10.2	10.2
4000	1.3677	0.8603	0.0089(0.0010)	0.0089(0.0010)	10.1	10.1
5000	1.9654	1.2667	0.0083(0.0010)	0.0083(0.0010)	10.3	10.3
Case 3: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.5.$						
1000	0.2552	0.1912	0.0137(0.0025)	0.0137(0.0025)	10.4	10.4
2000	0.6253	0.4143	0.0101(0.0018)	0.0101(0.0018)	10.4	10.5
3000	0.9820	0.7119	0.0079(0.0014)	0.0079(0.0014)	10.4	10.4
4000	1.4284	1.0524	0.0067(0.0011)	0.0067(0.0011)	10.2	10.2
5000	1.9240	1.4807	0.0061(0.0010)	0.0061(0.0010)	10.3	10.3
$\mathbf{p = 100}$						
Case 1: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.1.$						
1000	0.4914	0.1616	0.0119(0.0015)	0.0119(0.0015)	10.2	10.2
2000	1.3266	0.5097	0.0083(0.0009)	0.0083(0.0009)	10.1	10.2
3000	2.1535	0.8775	0.0067(0.0008)	0.0067(0.0008)	10.2	10.9
4000	3.0055	1.2941	0.0057(0.0006)	0.0057(0.0006)	10.1	10.2
5000	3.8056	1.7229	0.0050(0.0006)	0.0050(0.0006)	10.2	10.2
Case 2: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.1.$						
1000	0.5561	0.2103	0.0103(0.0014)	0.0103(0.0014)	10.3	10.3
2000	1.2715	0.4997	0.0071(0.0009)	0.0071(0.0009)	10.2	10.2
3000	2.0313	0.8822	0.0057(0.0006)	0.0057(0.0006)	10.2	10.2
4000	2.8906	1.2687	0.0049(0.0005)	0.0049(0.0005)	10.2	10.1
5000	3.8144	1.8367	0.0043(0.0005)	0.0044(0.0005)	10.2	10.4
Case 3: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.5.$						
1000	0.5383	0.2377	0.0073(0.0014)	0.0073(0.0014)	10.4	10.4
2000	1.3066	0.5901	0.0051(0.0008)	0.0052(0.0009)	10.3	10.3
3000	1.9853	1.0170	0.0042(0.0007)	0.0042(0.0007)	10.3	10.3
4000	2.9280	1.5267	0.0036(0.0006)	0.0036(0.0006)	10.3	10.3
5000	3.9526	1.9660	0.0033(0.0005)	0.0033(0.0005)	10.4	10.4

Table 2.10: Estimation results of $\beta^*, \sigma = 1$

	TIME		MAD(sd)		DF	
n	LAD-IP	LAD-AZSD	LAD-IP	LAD-AZSD	LAD-IP	LAD-AZSD
$\mathbf{p = 50}$						
Case 1: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.$						
1000	0.2596	0.1362	0.0675(0.0097)	0.0675(0.0097)	10.2	10.3
2000	0.5836	0.2912	0.0451(0.0052)	0.0450(0.0052)	10.3	10.3
3000	1.0197	0.5289	0.0366(0.0045)	0.0366(0.0045)	10.2	10.2
4000	1.5034	0.7327	0.0318(0.0038)	0.0318(0.0038)	10.2	10.1
5000	2.0822	1.0256	0.0285(0.0031)	0.0285(0.0031)	10.2	10.2
Case 2: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.1.$						
1000	0.2533	0.1356	0.0581(0.0075)	0.0581(0.0076)	10.2	10.2
2000	0.5829	0.3051	0.0398(0.0047)	0.0398(0.0048)	10.2	10.2
3000	1.0102	0.5445	0.0320(0.0038)	0.0320(0.0038)	10.2	10.2
4000	1.3586	0.6926	0.0273(0.0036)	0.0273(0.0036)	10.2	10.2
5000	1.7779	0.9191	0.0242(0.0033)	0.0242(0.0033)	10.1	10.1
Case 3: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.5.$						
1000	0.2260	0.1247	0.0416(0.0064)	0.0417(0.0065)	10.4	10.4
2000	0.5194	0.2905	0.0290(0.0047)	0.0291(0.0047)	10.3	10.3
3000	0.8290	0.5156	0.0245(0.0045)	0.0245(0.0045)	10.4	10.4
4000	1.2174	0.7667	0.0202(0.0033)	0.0202(0.0033)	10.3	10.3
5000	1.6814	1.0587	0.0184(0.0033)	0.0184(0.0034)	10.4	10.4
$\mathbf{p = 100}$						
Case 1: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.$						
1000	0.6197	0.1790	0.0360(0.0045)	0.0360(0.0045)	10.2	10.2
2000	1.3610	0.4161	0.0247(0.0028)	0.0247(0.0028)	10.1	10.1
3000	2.1097	0.7272	0.0198(0.0020)	0.0198(0.0020)	10.2	10.2
4000	3.0466	1.0314	0.0168(0.0022)	0.0168(0.0022)	10.2	10.2
5000	3.9469	1.3925	0.0151(0.0017)	0.0151(0.0017)	10.2	10.2
Case 2: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.1.$						
1000	0.5697	0.1655	0.0317(0.0038)	0.0317(0.0039)	10.2	10.2
2000	1.2843	0.4101	0.0215(0.0026)	0.0215(0.0026)	10.2	10.2
3000	2.0810	0.7351	0.0170(0.0019)	0.0170(0.0019)	10.2	10.2
4000	2.9373	1.0792	0.0148(0.0019)	0.0148(0.0019)	10.2	10.2
5000	3.8968	1.3963	0.0132(0.0015)	0.0132(0.0015)	10.2	10.2
Case 3: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.5.$						
1000	0.5525	0.2105	0.0229(0.0035)	0.0228(0.0035)	10.4	10.3
2000	1.2764	0.4977	0.0149(0.0024)	0.0148(0.0024)	10.3	10.4
3000	1.9854	0.8489	0.0124(0.0020)	0.0125(0.0020)	10.4	10.4
4000	2.9019	1.3003	0.0108(0.0018)	0.0108(0.0018)	10.2	10.2
5000	4.0582	1.6930	0.0095(0.0015)	0.0095(0.0015)	10.3	10.3

Table 2.11: Estimation results of $\beta^*, \sigma = 3$

	TIME		MAD(sd)		DF	
n	LAD-IP	LAD-AZSD	LAD-IP	LAD-AZSD	LAD-IP	LAD-AZSD
$\mathbf{p = 50}$						
Case 1: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.$						
1000	0.2464	0.1387	0.0215(0.0026)	0.0215(0.0026)	10.2	10.2
2000	0.6002	0.3547	0.0151(0.0020)	0.0152(0.0019)	10.2	10.2
3000	1.0313	0.6238	0.0121(0.0014)	0.0121(0.0014)	10.2	10.2
4000	1.5504	0.9044	0.0106(0.0013)	0.0106(0.0013)	10.1	10.2
5000	2.0347	1.2557	0.0094(0.0012)	0.0094(0.0012)	10.3	10.3
Case 2: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.1.$						
1000	0.2617	0.1587	0.0183(0.0026)	0.0183(0.0026)	10.1	10.1
2000	0.5872	0.3613	0.0125(0.0014)	0.0126(0.0014)	10.2	10.2
3000	1.0263	0.6434	0.0103(0.0014)	0.0103(0.0014)	10.1	10.1
4000	1.5150	0.9267	0.0087(0.0011)	0.0087(0.0011)	10.2	10.2
5000	2.0175	1.2463	0.0080(0.0010)	0.0080(0.0010)	10.2	10.2
Case 3: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.5.$						
1000	0.2661	0.1770	0.0129(0.0026)	0.0130(0.0026)	10.3	10.3
2000	0.5917	0.3970	0.0091(0.0017)	0.0091(0.0017)	10.3	10.3
3000	0.9556	0.7156	0.0072(0.0012)	0.0072(0.0012)	10.2	10.2
4000	1.4386	1.0317	0.0064(0.0011)	0.0064(0.0011)	10.3	10.3
5000	1.8109	1.3429	0.0055(0.0010)	0.0055(0.0010)	10.2	10.3
$\mathbf{p = 100}$						
Case 1: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.$						
1000	0.4683	0.1578	0.0100(0.0012)	0.0100(0.0012)	10.2	10.2
2000	0.9979	0.3764	0.0069(0.0008)	0.0069(0.0009)	10.2	10.2
3000	1.6729	0.6897	0.0056(0.0007)	0.0056(0.0007)	10.3	10.3
4000	2.3985	1.0404	0.0049(0.0005)	0.0049(0.0005)	10.2	10.2
5000	3.0926	1.3766	0.0043(0.0005)	0.0043(0.0005)	10.2	10.2
Case 2: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.1.$						
1000	0.4356	0.1765	0.0070(0.0012)	0.0070(0.0012)	10.3	10.3
2000	0.9918	0.4366	0.0046(0.0008)	0.0046(0.0008)	10.3	10.3
3000	1.7206	0.8164	0.0038(0.0007)	0.0038(0.0007)	10.2	10.9
4000	2.6736	1.2287	0.0033(0.0006)	0.0033(0.0006)	10.3	10.3
5000	3.3316	1.6629	0.0029(0.0005)	0.0029(0.0005)	10.2	10.2
Case 3: $\Sigma_{i,j} = \rho^{\lvert i-j \rvert}, i, j = 1, 2, \cdots, p; \quad \rho = 0.5.$						
1000	0.5259	0.2441	0.0200(0.0032)	0.0203(0.0034)	10.6	10.5
2000	1.0344	0.6109	0.0145(0.0023)	0.0147(0.0024)	10.4	10.2
3000	1.7057	1.0219	0.0113(0.0019)	0.0115(0.0019)	10.5	10.8
4000	2.3733	1.7914	0.0100(0.0017)	0.0104(0.0018)	10.6	10.2
5000	3.1115	2.2958	0.0085(0.0014)	0.0088(0.0015)	10.6	10.3

Table 2.12: Estimation results of $\beta^{**}, \sigma = 1$

	TIME		MAD(sd)		DF	
n	LAD-IP	LAD-AZSD	LAD-IP	LAD-AZSD	LAD-IP	LAD-AZSD
$\mathbf{p=50}$						
Case 1: $\Sigma_{i,j}=\rho^{\lvert i-j\rvert}, i,j=1,2,\cdots,p;\quad \rho=0.$						
1000	0.2730	0.1186	0.0665(0.0078)	0.0666(0.0079)	9.9	9.9
2000	0.6129	0.2765	0.0449(0.0061)	0.0450(0.0062)	10.2	10.2
3000	1.0860	0.4766	0.0373(0.0047)	0.0373(0.0047)	10.2	10.2
4000	1.6264	0.7094	0.0311(0.0036)	0.0311(0.0036)	10.2	10.2
5000	2.1503	0.9311	0.0279(0.0034)	0.0279(0.0034)	10.2	10.2
Case 2: $\Sigma_{i,j}=\rho^{\lvert i-j\rvert}, i,j=1,2,\cdots,p;\quad \rho=0.1.$						
1000	0.2619	0.1215	0.0561(0.0083)	0.0562(0.0083)	10.1	10.0
2000	0.5834	0.2749	0.0382(0.0051)	0.0382(0.0051)	10.3	10.3
3000	1.0356	0.4769	0.0308(0.0037)	0.0308(0.0037)	10.2	10.2
4000	1.5174	0.7225	0.0266(0.0033)	0.0266(0.0033)	10.3	10.3
5000	2.1174	0.9541	0.0240(0.0026)	0.0240(0.0026)	10.2	10.2
Case 3: $\Sigma_{i,j}=\rho^{\lvert i-j\rvert}, i,j=1,2,\cdots,p;\quad \rho=0.5.$						
1000	0.2577	0.1475	0.0381(0.0068)	0.0380(0.0067)	10.3	10.3
2000	0.5659	0.3275	0.0266(0.0053)	0.0266(0.0053)	10.2	10.2
3000	0.9814	0.5521	0.0216(0.0041)	0.0216(0.0041)	10.3	10.3
4000	1.4555	0.8231	0.0192(0.0035)	0.0192(0.0035)	10.2	10.2
5000	1.9990	1.0890	0.0167(0.0029)	0.0167(0.0029)	10.2	10.2
$\mathbf{p=100}$						
Case 1: $\Sigma_{i,j}=\rho^{\lvert i-j\rvert}, i,j=1,2,\cdots,p;\quad \rho=0.$						
1000	0.5029	0.1311	0.0304(0.0038)	0.0304(0.0038)	10.1	10.2
2000	1.1100	0.3034	0.0209(0.0023)	0.0209(0.0023)	10.2	10.2
3000	1.8231	0.5500	0.0167(0.0020)	0.0167(0.0020)	10.3	10.3
4000	2.6039	0.8175	0.0145(0.0016)	0.0145(0.0016)	10.3	10.3
5000	3.4149	1.0947	0.0128(0.0016)	0.0128(0.0016)	10.3	10.2
Case 2: $\Sigma_{i,j}=\rho^{\lvert i-j\rvert}, i,j=1,2,\cdots,p;\quad \rho=0.1.$						
1000	0.4609	0.1432	0.0210(0.0036)	0.0210(0.0037)	10.3	10.3
2000	1.0352	0.3652	0.0141(0.0023)	0.0141(0.0023)	10.3	10.7
3000	1.7127	0.6214	0.0113(0.0020)	0.0113(0.0020)	10.2	10.7
4000	2.4797	0.9357	0.0096(0.0017)	0.0096(0.0017)	10.2	10.2
5000	3.1546	1.2587	0.0089(0.0016)	0.0089(0.0016)	10.3	10.3
Case 3: $\Sigma_{i,j}=\rho^{\lvert i-j\rvert}, i,j=1,2,\cdots,p;\quad \rho=0.5.$						
1000	0.4978	0.1562	0.0605(0.0099)	0.0611(0.0101)	10.4	10.2
2000	1.0346	0.4578	0.0427(0.0066)	0.0433(0.0067)	10.4	10.4
3000	1.6837	0.8665	0.0346(0.0054)	0.0354(0.0056)	10.7	10.8
4000	2.3693	1.1189	0.0296(0.0049)	0.0301(0.0048)	10.5	10.6
5000	3.0494	1.6296	0.0266(0.0039)	0.0270(0.0039)	10.5	10.8

Table 2.13: Estimation results of $\beta^{**}, \sigma = 3$

Experiment 2. This example focuses on the effect of the error distribution. Consider four error models $\varepsilon \sim N(0,1)\,, \mathrm{Dbexp}\,(0,1),\ t\,(3)\,, \mathrm{Cauchy}\,(0,1)$. The robustness of LAD Fused Lasso (AZSD, IP) and Least Square fused Lasso (LS-fuse, see [77]) are compared. For each $p = 50\,, 100$, and $n = 5000$, we generate 100 datasets from Model (2.44) with $\sigma = 1\,, X \in \mathcal{N}_p(0,\Sigma)\,, \Sigma_{i,j} = 0.5^{|i-j|}\,, i,j = 1,2,\cdots,p$, and

$$\beta = (\underbrace{1,\cdots,1}_{1-5},\underbrace{0,\cdots,0}_{6-15},\underbrace{2,\cdots,2}_{16-20},\underbrace{0,0,0,0}_{21-24},\underbrace{\quad 3\quad}_{25},\underbrace{0,\cdots,0}_{26-40},\underbrace{1,\cdots,1}_{41-45},\underbrace{0,\cdots,0}_{46-p}).$$

For each case, `TIME`, `DF` (the same as Experiment 1) and average mean absolute

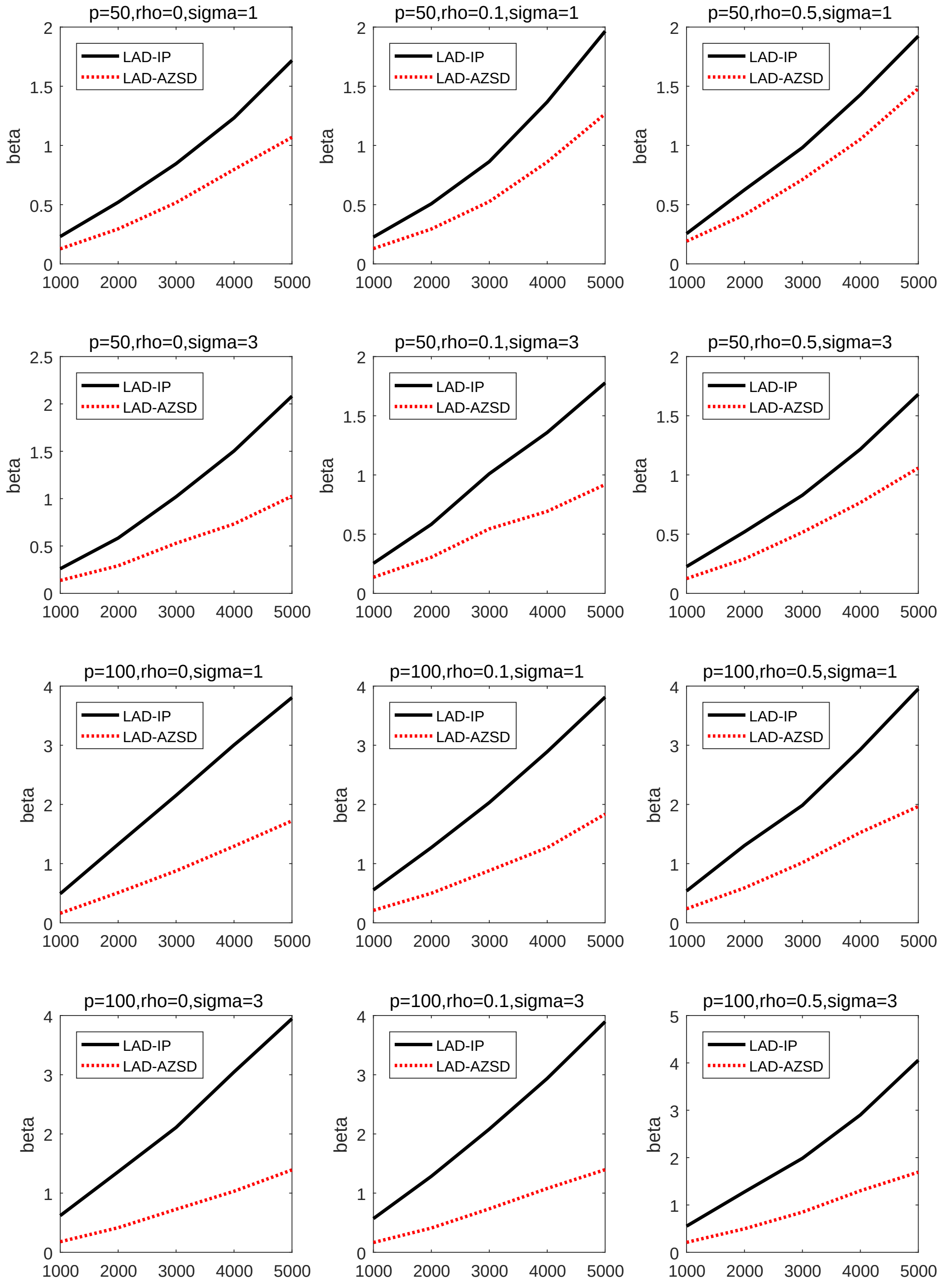

Figure 2.7: Running time tendency of β^* estimation with $p = 50, 100\,, \sigma = 1, 3\,, \rho = 0, 0.1, 0.5$.

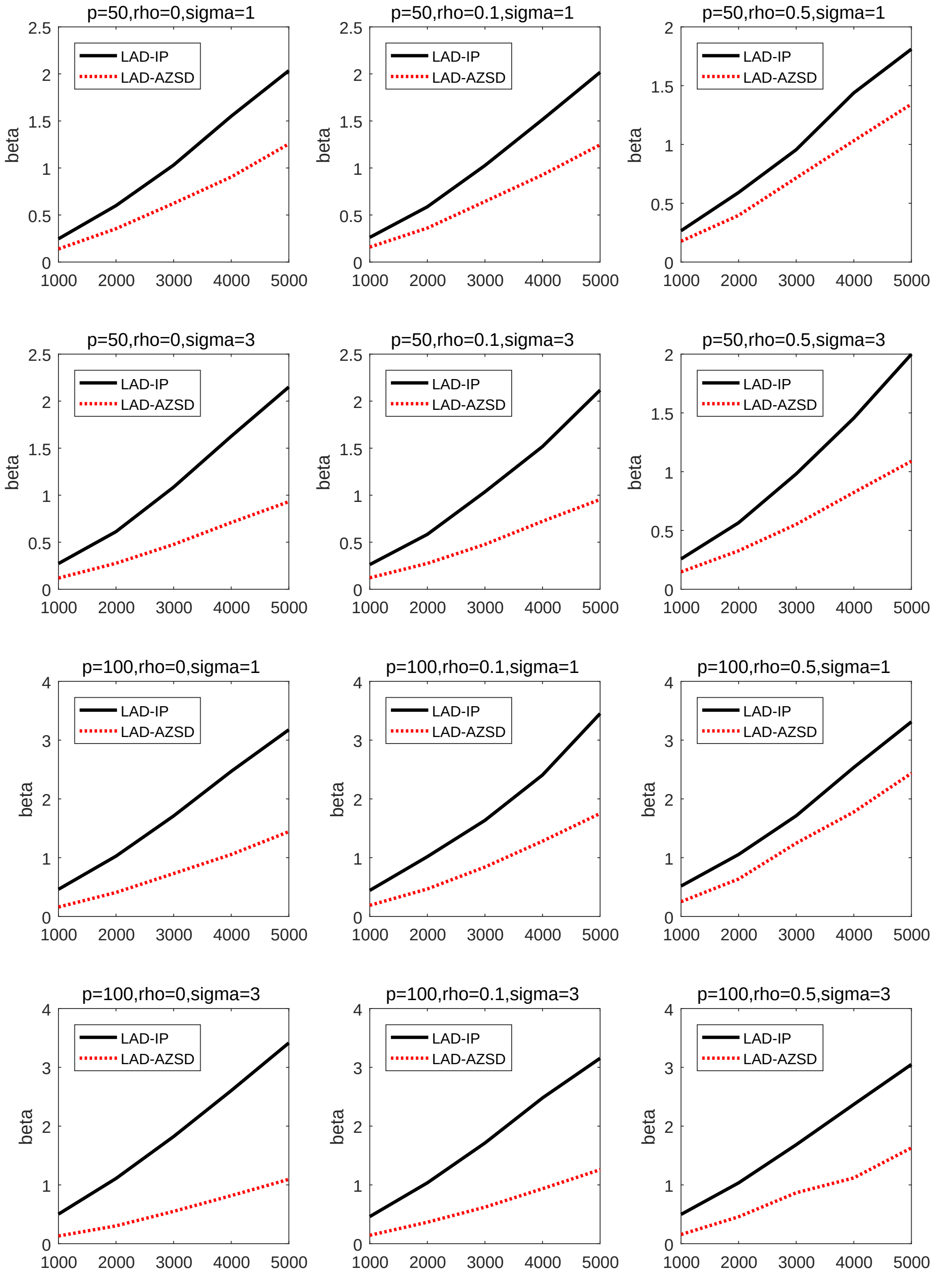

Figure 2.8: Running time tendency of β^{**} estimation with $p = 50, 100\,, \sigma = 1, 3\,, \rho = 0, 0.1, 0.5$.

deviation (`MAD`) $\sum_{i=1}^{100}(\hat{\beta}_i - \beta_i)^2$ are reported in Table 2.14, Figure 2.9, and 2.10. It can be seen that under the heavy-tailed case Cauchy (0,1), LAD Fused Lasso outperforms LS Fused Lasso in terms of MAD.

	$p = 50$				$p = 100$		
ε Distr. Type	LAD-IP	LAD-AZSD	LS-fuse		LAD-IP	LAD-AZSD	LS-fuse
	`TIME`						
N(0,1)	0.2261	0.1565	0.0011		0.4939	0.2129	0.0015
Dexp(0,1)	0.2450	0.1821	0.0010		0.5854	0.2412	0.0016
$t(3)$	0.2219	0.1533	0.0010		0.5003	0.2831	0.0016
Cauchy(0,1)	0.3490	0.1548	0.0012		0.6769	0.2191	0.0025
	`MAD`						
N(0,1)	1.2791	1.2772	1.1291		1.7249	1.7206	2.7586
Dexp(0,1)	1.4911	1.4918	2.2297		1.6815	1.6911	3.4561
$t(3)$	1.6499	1.6484	2.3727		2.3149	2.3125	6.2918
Cauchy(0,1)	2.2881	2.2840	58.0899		2.2811	2.2677	147.0171
	`DF`						
N(0,1)	16.0	16.0	50.0		16.0	16.0	100.0
Dexp(0,1)	17.0	17.0	50.0		17.0	17.0	100.0
$t(3)$	16.0	16.0	50.0		16.0	16.0	100.0
Cauchy(0,1)	16.0	16.0	50.0		16.0	16.0	100.0

Table 2.14: Estimation results under heavy tailed distributions for $p = 50, 100$.

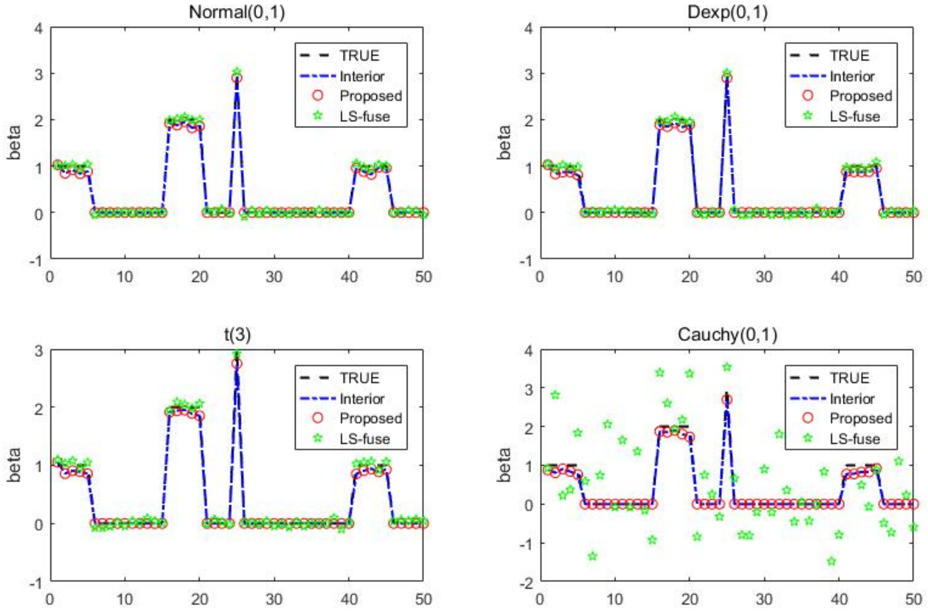

Figure 2.9: $p = 50\,, n = 5000$ estimation results of interior point method, proposed method and LS-fuse.

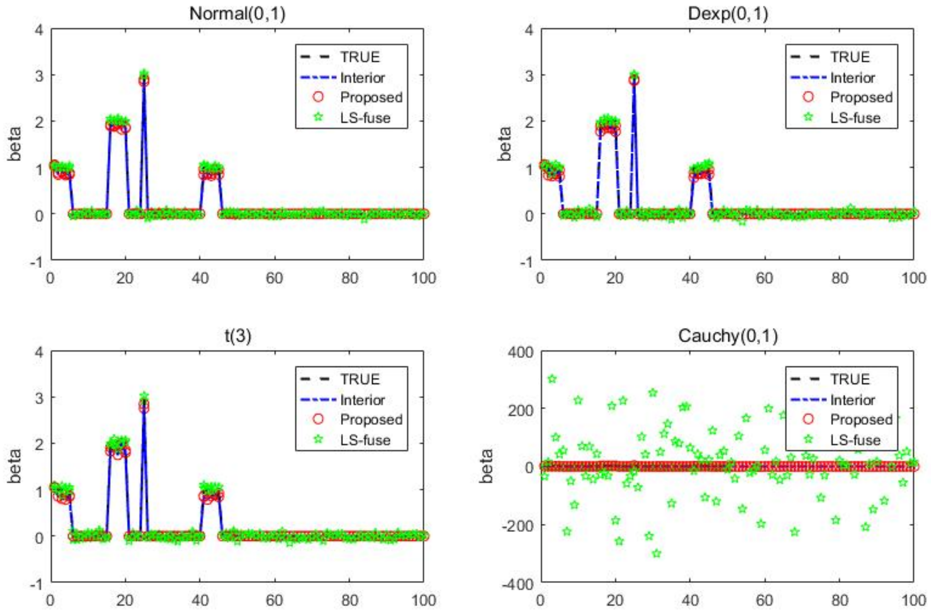

Figure 2.10: $p = 100\,, n = 5000$ estimation results of interior point method, proposed method and LS-fuse.

Real data analysis

In this section, we apply LAD Fused Lasso to soybean data in Davidian and Giltinan [26] (1995, §1.1.3, p.7), that is available in R package MEMSS. The experiment was carried out in three years, 1988, 1989, 1990. The average leaf weight (in grams) randomly chosen from 6 plants was measured at days after planting as $A = (A_1, A_2, \ldots, A_{25}) =$

$$(14, 15, 20, 21, 23, 27, 28, 30, 34, 35, 37, 41, 42, 43, 49, 51, 55, 56, 63, 64, 69, 70, 77, 79, 84)\,.$$

Eight plants were planted with each genotype in each planting year, giving a total of 48 plots in the study. Each plot is observed at only 8 to 10 days among the above-mentioned 25 days. Detailed information is illustrated in Figure 2.11.

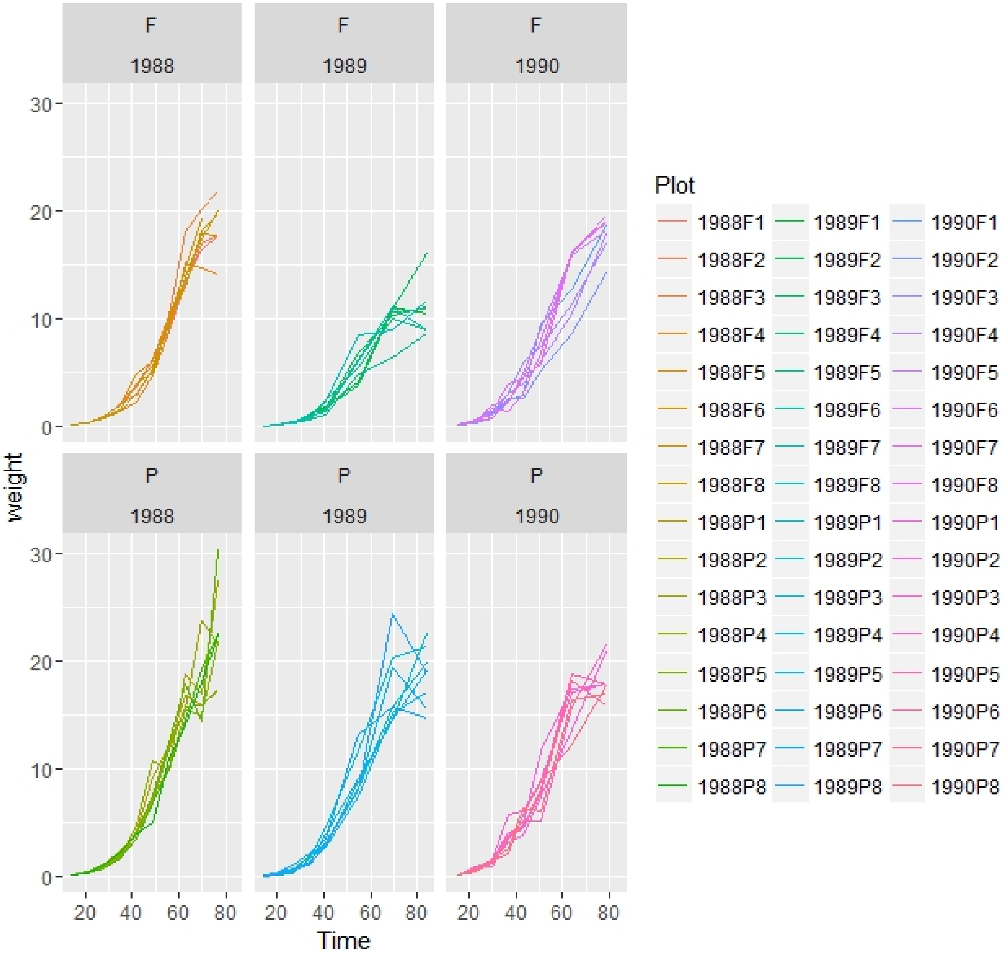

Figure 2.11: Plot of soybean data.

Let Y_{ij} be the differences between consecutive measurements of plot i at time $T_{j+1} - T_j$. Define the covariates $X_{ijk} = (A_{k+1} - A_k)I_{[(A_k, A_{k+1}) \subset (T_j, T_{j+1})]}$, where $I(\cdot)$ is the indicator function. Consider the model $Y = X\beta + \sigma\epsilon$. Here, the coefficients β_k, $k = 1, 2, \ldots, 25$ can be interpreted as the average growth rates between time A_{k-1} and A_k. For convenience, $A_0 = 0$ is defined. The changes in the growth rate can be detected using LAD Fused Lasso. Here, the tuning parameter is set as $\lambda = \sqrt{2n \log p}$ (see [114]).

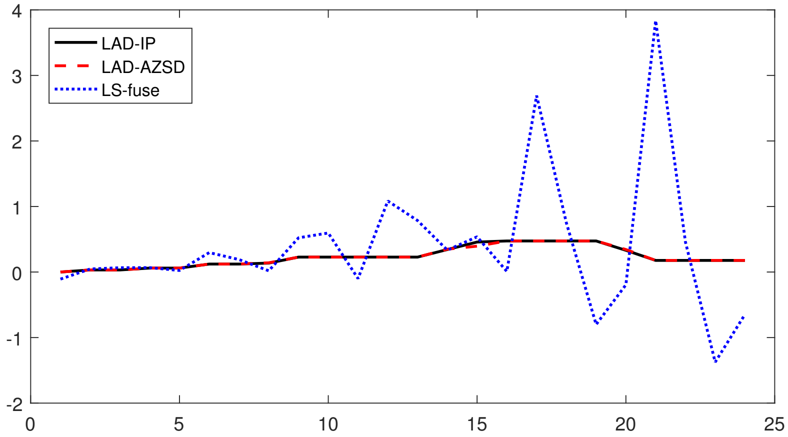

Figure 2.12: Plot of estimation results of soybean data with LAD-IP, LAD-AZSD and LS-fuse.

The estimated regression coefficients β under LAD-IP, LAD-AZSD, and LS-fuse as plotted in Figure 2.12. If least square is used, it is difficult to tell if there are changes in the growth rate. On the contrary, LAD is more robust to heavy-tailed cases and it allows us to see clearly that between A_{15} and A_{20} is a fast growing period. The estimated β of LAD-AZSD and LAD-IP models are more blocky and smoothly varying comparing to LS-fuse. The estimations from LAD-AZSD and LAD-IP are close to each other but significantly differ from that obtained by LS-fuse. To evaluate the performances of our algorithm, consider TIME and DF as described in Section 2.2.2). Table 2.15 shows that LAD-AZSD is considerably faster than LAD-IP. Table 2.16 indicates that both LAD-AZSD and LAD-IP give smooth and blocky estimators, whereas LS-fuse estimator is noisy.

	LAD-AZSD	LAD-IP	LS-fuse
TIME	0.9674	0.1533	0.0763
DF	23	23	24
Dimension	$(n, p) = (364, 24)$		

Table 2.15: Quantitative comparison of soybean data of 3 methods: LAD-IP, LAD-AZSD and LS-fuse models.

	X_{ij1}	X_{ij2}	X_{ij3}	X_{ij4}	X_{ij5}	X_{ij6}	X_{ij7}	X_{ij8}
lAD-IP	0.0000	0.0319	0.0319	0.0602	0.0602	0.1217	0.1217	0.1363
LAD-AZSD	0.0000	0.0319	0.0319	0.0602	0.0602	0.1217	0.1217	0.1412
LS-fuse	-0.1080	0.0465	0.0656	0.0669	0.0238	0.2967	0.1904	0.0185
	X_{ij9}	X_{ij10}	X_{ij11}	X_{ij12}	X_{ij13}	X_{ij14}	X_{ij15}	X_{ij16}
LAD-IP	0.2279	0.2279	0.2279	0.2279	0.2279	0.3432	0.4576	0.4741
LAD-AZSD	0.2279	0.2279	0.2279	0.2279	0.2279	0.3487	0.3947	0.4741
LS-fuse	0.5221	0.5956	-0.1000	1.0840	0.7852	0.3347	0.5377	0.0086
	X_{ij17}	X_{ij18}	X_{ij19}	X_{ij20}	X_{ij21}	X_{ij22}	X_{ij23}	X_{ij24}
LAD-IP	0.4741	0.4741	0.4741	0.3314	0.1772	0.1772	0.1772	0.1772
LAD-AZSD	0.4741	0.4741	0.4741	0.3435	0.1772	0.1772	0.1772	0.1772
LS-fuse	2.6922	0.7504	-0.8007	-0.1946	3.8369	0.4601	-1.3747	-0.6476

Table 2.16: Parameter estimation results of soybean data with LAD-IP, LAD-AZSD, LS-fuse.

Chapter 3

Constrained LAD Lasso models

In this chapter, we focus on Constrained LAD Lasso models with linear equality constraints, and extend the algorithm in Section 2.1 to linearly equality constraint case. Then we applied the algorithm to regularized Mean Absolute Deviation portfolio selection (MAD-Lasso) strategy.

3.1 MAD-Lasso model

Suppose that there are n securities $(\mathcal{S}_1, \cdots, \mathcal{S}_n)$ and their rate of returns are represented by the random vector $\boldsymbol{R} = (R_1\,,\, R_2\,, \cdots, R_n)$. The rate of returns at time t are $\boldsymbol{r}_t = (r_{t1}, r_{t2}, \cdots, r_{tn})\,, t = 1, 2, \cdots, T$, with mean $\boldsymbol{r} = (r_1, r_2, \cdots, r_n)'$ and covariance matrix $\mathbb{E}\big((\boldsymbol{r}_t - \boldsymbol{r})'(\boldsymbol{r}_t - \boldsymbol{r})\big) = \boldsymbol{\Sigma}$. The portfolio allocation weight vector $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)'$ satisfies $\sum_{i=1}^{n} x_i = 1$. Konno [67] proposed the Mean Absolute Deviation (MAD) risk measure, defined as

$$\mathrm{MAD}(\boldsymbol{x}) = \big|\sum_{i=1}^{n} x_i R_i - \mathbb{E}\sum_{i=1}^{n} x_i R_i\big| = \frac{1}{T}\sum_{t=1}^{T}\big|\sum_{i=1}^{n}(r_{ti} - r_i)x_i\big|.$$

The MAD-Lasso problem is formulated as

$$\begin{aligned} \min_{x} \quad & \frac{1}{T}\sum_{t=1}^{T}\Big|\sum_{i=1}^{n}(r_{ti}-r_i)x_i\Big| + \sum_{i=1}^{n}\lambda|x_i| \\ \text{s.t.} \quad & \boldsymbol{x}'\boldsymbol{r} = r_0, \\ & \boldsymbol{x}'\mathbf{1} = 1. \end{aligned} \tag{3.1}$$

Here, λ is the tuning parameter controlling the size of penalty. Brodie [16] penalizes Markowitz's model [87] with Lasso penalty. The Markowitz-Lasso problem can be described as

$$\begin{aligned} \arg\min_{\boldsymbol{x}} \quad & \mathbb{E}\big[|r_0\mathbf{1}_T - \boldsymbol{x}'\boldsymbol{r}_t|^2\big] + \lambda\|\boldsymbol{x}\|_1 \\ \text{s.t.} \quad & \boldsymbol{x}'\boldsymbol{r} = r_0, \\ & \boldsymbol{x}'\mathbf{1} = 1. \end{aligned} \tag{3.2}$$

The MAD-Lasso based method (3.1) has the following advantages:

1. It encourages sparsity. With appropriately chosen tuning parameter λ, some components in the portfolio weight vector $\boldsymbol{x}$ shrink towards zero, resulting in sparse portfolio selection strategies.

2. It controls the shorting level of portfolio selection model. The equivalent formulation is to minimize

$$\|r_0\mathbf{1}_T - \boldsymbol{r}'\boldsymbol{x}\|_1 + 2\lambda\sum_{i:x_i\leqslant 0}|x_i| + \lambda,$$

   where $\sum_{i:x_i\leqslant 0}|x_i|$ controls the shorting level. The last term does not affect the optimization problem.

3. It robustify the portfolio selection problem. The ℓ_1 norm penalty mitigate the computational difficulties related to the possible collinearity in the rates of

returns of different assets. Moreover, it ameliorates the influence of financial violations and extreme cases.

Proposition 3.1.1. *We have the followings.*

(1) For any two tuning parameters $\lambda_1 < \lambda_2$, let $\boldsymbol{x}^{(\lambda_1)}, \boldsymbol{x}^{(\lambda_2)}$ be the corresponding weight vectors. Then, we have

$$(\lambda_1 - \lambda_2)\big(\|\boldsymbol{x}^{(\lambda_2)}\|_1 - \|\boldsymbol{x}^{(\lambda_1)}\|_1\big) \geqslant 0.$$

This indicates that the greater is the penalty λ , the greater is the sparsity.

(2) Suppose that there exists λ_0 such that all entries in $\boldsymbol{x}^{(\lambda_0)}$ are non-negative. Then, for any $\lambda \geqslant \lambda_0$, all entries in the solution $\boldsymbol{x}^{(\lambda)}$ are non-negative too.

Proof. (1) Suppose there are two portfolio allocation vectors $\boldsymbol{x}^{(\lambda_1)}, \boldsymbol{x}^{(\lambda_2)}$ corresponding to the tuning parameter λ_1, λ_2 respectively in the MAD-Lasso problem (3.1). We have

$$\begin{aligned}
& \|r_0\mathbf{1}_T - \boldsymbol{r}'\boldsymbol{x}^{(\lambda_1)}\|_1 + \lambda_1\|\boldsymbol{x}^{(\lambda_1)}\|_1 \\
\leqslant\ & \|r_0\mathbf{1}_T - \boldsymbol{r}'\boldsymbol{x}^{(\lambda_2)}\|_1 + \lambda_1\|\boldsymbol{x}^{(\lambda_2)}\|_1 \\
=\ & \|r_0\mathbf{1}_T - \boldsymbol{r}'\boldsymbol{x}^{(\lambda_2)}\|_1 + \lambda_2\|\boldsymbol{x}^{(\lambda_2)}\|_1 + (\lambda_1 - \lambda_2)\|\boldsymbol{x}^{(\lambda_2)}\|_1 \\
\leqslant\ & \|r_0\mathbf{1}_T - \boldsymbol{r}'\boldsymbol{x}^{(\lambda_1)}\|_1 + \lambda_2\|\boldsymbol{x}^{(\lambda_1)}\|_1 + (\lambda_1 - \lambda_2)\|\boldsymbol{x}^{(\lambda_2)}\|_1 \\
=\ & \|r_0\mathbf{1}_T - \boldsymbol{r}'\boldsymbol{x}^{(\lambda_1)}\|_1 + \lambda_1\|\boldsymbol{x}^{(\lambda_1)}\|_1 + (\lambda_1 - \lambda_2)\big(\|\boldsymbol{x}^{(\lambda_2)}\|_1 - \|\boldsymbol{x}^{(\lambda_1)}\|_1\big).
\end{aligned}$$

This yields that

$$(\lambda_1 - \lambda_2)\big(\|\boldsymbol{x}^{(\lambda_2)}\|_1 - \|\boldsymbol{x}^{(\lambda_1)}\|_1\big) \geqslant 0. \tag{3.3}$$

(2) If all the entries of $\boldsymbol{x}^{(\lambda_0)}$ are nonnegative and some entries of $\boldsymbol{x}^{(\lambda)}$ are negative, we have $\|\boldsymbol{x}^{(\lambda)}\| \geqslant \sum_{i=1}^n |x_i^{(\lambda)}| = |\sum_{i=1}^n x_i^{(\lambda_0)}| = \sum_{i=1}^n |x_i^{(\lambda_0)}| = 1$. This yields that $\|\boldsymbol{x}^{(\lambda)}\| \geqslant \|\boldsymbol{x}^{(\lambda_0)}\|$. From (3.3), we have $\lambda_0 \geqslant \lambda$. This indicates that the all-nonnegative-entry case λ_0 corresponds to the sparest solution. The particular solution corresponding to λ_0 is the optimal solution among all solutions corresponding to $\lambda \geqslant \lambda_0$.

3.2 Constrained LAD Lasso model

To solve the MAD-Lasso portfolio selection problem, we generalize the descent algorithm of Shi et al. [101] to allow linearly equality constraints.

Problem definition

Problem (3.1) can be reformulated as the following Constrained LAD Lasso problem:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \|y - A\boldsymbol{x}\|_1 + \lambda\|\boldsymbol{x}\|_1 \\ \text{s.t.} \quad & C\boldsymbol{x} = b. \end{aligned} \tag{3.4}$$

Here,

$$A = \begin{pmatrix} r_{11} - r_1 & r_{12} - r_2 & \cdots & r_{1n} - r_n \\ r_{21} - r_1 & r_{22} - r_2 & \cdots & r_{2n} - r_n \\ \vdots & \vdots & \vdots & \vdots \\ r_{T1} - r_1 & r_{T2} - r_2 & \cdots & r_{Tn} - r_n \end{pmatrix}, y = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

$$C = \begin{pmatrix} r_1 & r_2 & \cdots & r_n \\ 1 & 1 & \cdots & 1 \end{pmatrix}, b = \begin{pmatrix} r_0 \\ 1 \end{pmatrix}.$$

Here, $y = (y_1, \cdots, y_T)' \in \mathbb{R}^T$, $A \in \mathbb{R}^{T\times n}$, $C \in \mathbb{R}^{q\times n}$, $b = (b_1, \cdots, b_q)' \in \mathbb{R}^q$, and $\boldsymbol{x} = (x_1, \cdots, x_n)' \in \mathbb{R}^n$. Denote the i-th row of A by A_i and the i-th element of y by y_i. Without loss of generality, we assume that C is full rank matrix, i.e., $\text{rank}\,(C) = q = 2$.

Remark 3.1. *An intuitive solution of Problem* (3.4) *is to be transformed to the following unconstrained optimization problem:*

$$\min_{\boldsymbol{x}} \|y - A\boldsymbol{x}\|_1 + \lambda_1\|\boldsymbol{x}\|_1 + \lambda_2\|C\boldsymbol{x} - b\|_1,$$

where we need two penalty parameters, λ_1 and λ_2, the computational complexity increases significantly. Hence we focus on Problem (3.4)*.*

Optimality conditions for feasible direction

Note that an arbitrary point $\boldsymbol{x}$ can be transformed to a feasible point as shown below. Suppose that $C\boldsymbol{x} - b = c_0 \neq 0$. Setting $\boldsymbol{x}_0 = \boldsymbol{x} - C^{\intercal}(CC^{\intercal})^{-1}c_0$, then,

$$C\boldsymbol{x}_0 - b = C\boldsymbol{x} - b - CC^{\intercal}(CC^{\intercal})^{-1}c_0 = c_0 - c_0 = 0.$$

The transformed point $\boldsymbol{x}_0$ is said to be the feasible point generated by $\boldsymbol{x}$. Thus, the initial point for the algorithm can be chosen as a feasible point. If $\boldsymbol{x}$ is a feasible point, we choose a direction h such that the cost function value decreases along this direction. The choice of the direction can not be arbitrary because the constraints must be satisfied along this direction. That is, it is required that

$$C(\boldsymbol{x} + h) - b = C\boldsymbol{x} - b + Ch = Ch = 0. \tag{3.5}$$

Definition 3.1. *The direction h fulfilling* (3.5) *is called a feasible direction. If h is a feasible direction, the corresponding directional derivative is a feasible directional derivative.*

First, we have the following assumptions.

Assumption 3.2.1. *For any $\boldsymbol{x}$ and $h \in \mathbb{R}^n$, $\lim_{\lambda \to \infty} f(\boldsymbol{x} + \lambda h) = \infty$.*

Assumption 3.2.2. *For any n indices $i_1, \cdots, i_n$ in $\{1, \cdots, T\}$, $\{A_{i_1}, \cdots, A_{i_n}\}$ are linearly independent.*

Denote $A_i\boldsymbol{x} - y_i = u_i$ and $\Omega = \{o_1, \cdots, o_m\} = \{o_i : u_{o_i} = 0, i = 1, 2, \cdots, m\}$ as the zero set. Then, the objective function can be rewritten as summation of smooth and nonsmooth part of $f(\boldsymbol{x})$:

$$\begin{aligned}
f(\boldsymbol{x}) &= S(\boldsymbol{x}) + N(\boldsymbol{x}), \\
S(\boldsymbol{x}) &\triangleq \sum_{i=1}^{T} \mathrm{I}\,(u_i > 0)\,(A_i\boldsymbol{x} - y_i) + \sum_{i=1}^{T} \mathrm{I}\,(u_i < 0)\,(-A_i\boldsymbol{x} + y_i) \triangleq c\boldsymbol{x} + z, \\
N(\boldsymbol{x}) &\triangleq \sum_{i=1}^{T} \mathrm{I}\,(u_i = 0)\,|A_i\boldsymbol{x} - y_i| = \sum_{i=1}^{m} |A_{o_i}\boldsymbol{x} - y_{o_i}|,
\end{aligned}$$

where $\mathrm{I}\,(\cdot)$ as the indicator function and

$$c \triangleq \sum_{i=1}^{T} \mathrm{I}\,(u_i > 0)A_i - \sum_{i=1}^{T} \mathrm{I}\,(u_i < 0)A_i\,, z = -\sum_{i=1}^{T} \mathrm{I}\,(u_i > 0)y_i + \sum_{i=1}^{T} \mathrm{I}\,(u_i < 0)y_i.$$

Since $f(\boldsymbol{x})$ is convex, its local minimizer must be the global minimizer. The optimality condition of the minimizer is that any feasible directional derivatives are greater than or equal to zero. That is, $\boldsymbol{x}^*$ is the optimal solution of (3.4) if and only if

$$\nabla_h f(\boldsymbol{x}^*) = \nabla_h S(\boldsymbol{x}^*) + \nabla_h N(\boldsymbol{x}^*) \geqslant 0, \quad \forall h \in \{h \mid h \in \mathbb{R}^n, Ch = 0\}. \tag{3.6}$$

However, it is not easy to verify the optimality condition (3.6) because there are infinitely-many feasible directions h. To obtain a finite representation of the optimality condition, consider the nonsmooth part $N(\boldsymbol{x})$ with

$$A_{o_i}\boldsymbol{x} = y_{o_i}, \quad i = 1, \cdots, m.$$

If $\{A_{o_i} : i = 1, \cdots, m\}$ are independent, then $m \leqslant n - q$. If $m > n - q$, then the equations above are overdetermined. Without loss of generality, we assume that $m \leqslant n - q$ and $\{A_{o_i} : i = 1, \cdots, m\} \cup \{C_1, C_2, \cdots, C_q\}$ are linearly independent. Let

$$D = \begin{pmatrix} C_1 \\ \vdots \\ C_q \\ A_{o_1} \\ \vdots \\ A_{o_m} \end{pmatrix}.$$

Generalized inverse matrix V_D can be obtained such that $DV_D = I_{m+q}$, where I_{m+q} is the $(m+q) \times (m+q)$ identity matrix and $V_D = (V_1, \cdots, V_{m+q})$. Consider the null space $\{V \in \mathbb{R}^n | DV = 0\}$. There exist $n - m - q$ linearly independent vectors $V_j\,, j = m + q + 1, \cdots, n$ that form the basis of the null space. Hence, we have $DV_j = 0\,, \forall j = m + q + 1, \cdots, n$. Then, an equivalent finite-representation of the optimality condition (3.6) is given by the following theorem.

Theorem 3.1. *Suppose rank$(D) = m + q$, then $\boldsymbol{x}^*$ is the optimal solution if and only if the feasible directional derivatives satisfy*

$$\begin{aligned}
&\nabla_{V_i} f(\boldsymbol{x}^*) = \nabla_{V_i} S(\boldsymbol{x}^*) + \nabla_{V_i} N(\boldsymbol{x}^*) \geqslant 0, \quad i = q+1, \cdots, q+m, \\
&\nabla_{V_i^-} f(\boldsymbol{x}^*) = \nabla_{V_i^-} S(\boldsymbol{x}^*) + \nabla_{V_i^-} N(\boldsymbol{x}^*) \geqslant 0, \quad i = q+1, \cdots, q+m, \\
&\nabla_{V_i} f(\boldsymbol{x}^*) = \nabla_{V_i} S(\boldsymbol{x}^*) = 0, \quad i = m+q+1, \cdots, n.
\end{aligned} \tag{3.7}$$

Proof. Note that the space of all feasible directions is spanned by $\{V_i\,, i = q + 1, \cdots, n\}$, condition (3.7) is a special case of (3.6) and the necessary condition is obvious. Next, we establish the sufficient condition, this means that if (3.7) are satisfied, (3.6) holds. If $\boldsymbol{x}^*$ is optimal, then we have

$$\begin{aligned}
&\nabla_{V_i} f(\boldsymbol{x}^*) \geqslant 0\,, \nabla_{V_i^-} f(\boldsymbol{x}^*) \geqslant 0, \quad i = q+1, \cdots, q+m, \\
&\nabla_{V_i} f(\boldsymbol{x}^*) = 0, \quad i = q+m+1, \cdots, n.
\end{aligned}$$

By orthonormality of $\{V_1, V_2, \cdots, V_n\}$, (3.7) can be simplified as

$$\begin{aligned}
&\nabla_{V_i} f(\boldsymbol{x}^*) = \nabla_{V_i} S(\boldsymbol{x}^*) = \frac{cV_i}{\|V_i\|} = 0, \quad i = m+q+1, \cdots, n, \\
&\nabla_{V_i} f(\boldsymbol{x}^*) = \nabla_{V_i} S(\boldsymbol{x}^*) + \nabla_{V_i} N(\boldsymbol{x}^*) = \frac{cV_i}{\|V_i\|} + \frac{1}{\|V_i\|} \geqslant 0, \quad i = q+1, \cdots, m+q, \\
&\nabla_{V_i^-} f(\boldsymbol{x}^*) = \nabla_{V_i^-} S(\boldsymbol{x}^*) + \nabla_{V_i^-} N(\boldsymbol{x}^*) = \frac{-cV_i}{\|V_i\|} + \frac{1}{\|V_i\|} \geqslant 0, \quad i = q+1, \cdots, m+q.
\end{aligned}$$

For any feasible direction h, there exists a weight vector $w = (w_{q+1}, w_{q+2}, \cdots, w_n)'$ such that

$$h = \sum_{i=q+1}^{n_1} w_i V_i + \sum_{i=n_1+1}^{n} w_i(-V_i).$$

Without loss of generality, we can set $w_i \geqslant 0, \forall i = 1, \cdots, n$. This is because when $w_i < 0$, we have $w_i V_i = (-w_i) \cdot (-V_i)$. Then, replacing V_i by $-V_i$ and w_i by $-w_i > 0$ yield that

$$\begin{aligned} \nabla_h N(\boldsymbol{x}^*) &= \frac{\sum_{i=1}^m |A_{o_i} h|}{\|h\|} = \frac{\sum_{i=1}^m \left| A_{o_i}\left(\sum_{j=q+1}^{n_1} w_j V_j + \sum_{j=n_1+1}^{n} w_j(-V_j)\right)\right|}{\|h\|} \\ &= \frac{\sum_{i=1}^m \left| w_i A_{o_i} V_{i+n}\right|}{\|h\|} = \frac{\sum_{i=1}^m w_i}{\|h\|}. \end{aligned}$$

We have

$$\begin{aligned} \nabla_h f(\boldsymbol{x}^*) &= \frac{\sum_{i=q+1}^n w_i c V_i}{\|h\|} + \frac{\sum_{i=1}^m w_i}{\|h\|} = \frac{1}{\|h\|}\Big(\sum_{i=1}^m t_{i+q}(cV_{i+q} + 1) + \sum_{i=m+q+1}^{n} w_i c V_i \Big) \\ &= \frac{1}{\|h\|} \sum_{i=1}^m w_{i+q} \nabla_{V_{i+q}} f(\boldsymbol{x}^*) \cdot \|V_{i+q}\| + \frac{1}{\|h\|} \sum_{i=m+q+1}^{n} w_i \nabla_{V_i} f(\boldsymbol{x}^*) \|V_i\| \geqslant 0. \end{aligned}$$

Then, for any feasible direction h, the feasible directional derivative is greater than or equals to zero. Hence, (3.6) holds and $\boldsymbol{x}^*$ is the optimal solution. □

Descent feasible directions

The design of the algorithm is as follows. Let $\boldsymbol{x}^{(k)}$ be the approximation at the k-th iteration. If the optimality condition (3.7) is satisfied, then $\boldsymbol{x}^{(k)}$ is the optimal solution, otherwise, there exists at least a feasible direction h such that the cost function decreases along this direction. These steps are repeated until (3.7) is satisfied.

Suppose that the i-th condition of (3.7) is not satisfied. Then, the following two statements

$$\nabla_{V_i} f(\boldsymbol{x}) \geqslant 0 \text{ and } \nabla_{V_i^-} f(\boldsymbol{x}) \geqslant 0$$

can not be satisfied at the same time and consequently at least one of V_i and V_i^- is the descent direction. For an iterative point $\boldsymbol{x}^{(k)}$, denote the zero set by $\Omega_k = \{o_{k1}, o_{k2}, \cdots, o_{kk_m}\} = \{o_{ki} : u_{ki} = A_{ki}\boldsymbol{x} - y_{ki} = 0\,, i = 1, 2, \cdots, k_m\}$. The cost function can be rewritten as

$$f(\boldsymbol{x}^{(k)}) = c^{(k)}\boldsymbol{x}^{(k)} + \sum_{o_{ki} \in \Omega_k} |A_{o_{ki}}\boldsymbol{x}^{(k)} - y_{o_{ki}}| + z^{(k)}. \tag{3.8}$$

Denote by Ω_k' the set of all the indexes ki so that (3.7) is not satisfied for V_i or V_i^-. To speed up the search, consider the indexes i so that the descent directional derivatives $\nabla_{V_i} f$ or $\nabla_{V_i^-} f$ are the greatest. Suppose that $\Lambda_1 \subset \Omega_k'$ contains a proportion α of the indexes in Ω_k' with slowest corresponding descent directional derivatives $\nabla_{V_i} f$ or $\nabla_{V_i^-} f$. Denote

$$\Omega_k^0 = \Omega_k \backslash \Lambda_1^k.$$

This means that the indexes in Λ_1^k are removed from the zero set. Choose the descent direction h in the space spanned by

$$\{V_{o_{ki}} : o_{ki} \in \Lambda_1^k\} \cup \{V_i : i = k_m + 1, \cdots, n\}$$

such that

$$h = \sum_{o_{ki} \in \Lambda_1^k} t_{o_{ki}} V_{o_{ki}} + \sum_{i=k_m+1}^{n} t_i V_i.$$

It can be verified that

$$A_{o_{ki}} h = 0, \quad \forall o_{ki} \in \Omega_k^0.$$

Such a choice guarantees that the descent direction keep the set Ω_k^0 unchanged. Set the descent direction $h^{(k)}$ as the optimal solution to

$$\begin{aligned} \max_{v \in \mathbb{R}^p} \quad & -c^{(k)} v \\ s.t. \quad & A_{o_{ki}} v = 0, \quad \forall o_{ki} \in \Omega_k^0. \end{aligned} \tag{3.9}$$

It means that the solution h is chosen as the vector nearest to the deepest descent direction $-c^{(k)}$. The optimal solution to Problem (3.9) is

$$\tilde{h} = -c^{(k)} - A_{0k}^{\intercal}(A_{0k}A_{0k}^{\intercal})^{-1}A_{0k} \cdot (-c^{(k)}), \tag{3.10}$$

where $A_{0k}^{\intercal}(A_{0k}A_{0k}^{\intercal})^{-1}A_{0k}(-c^{(k)})$ is the projected direction of $-c^{(k)}$ in the subspace $\{h : A_{o_{ki}}h = 0, o_{ki} \in \Omega_k^0\}$. Equivalently, $A_{0k}^{\intercal}(A_{0k}A_{0k}^{\intercal})^{-1}A_{0k}(-c^{(k)}) = \text{Proj}\{-c^{(k)}\big|h : A_{o_{ki}}h = 0, o_{ki} \in \Omega_k^0\}$. Without loss of generality, assume that $o_{k1}, o_{k2}, \cdots, o_{kl} \in \Omega_0^k$ and

$$A_{0k} = \begin{pmatrix} A_{o_{k1}} \\ \vdots \\ A_{o_{kl}} \end{pmatrix}, o_{k1}, \cdots, o_{kl} \in \Omega_k^0.$$

Then, the descent direction $h^{(k)}$ can be chosen as the normalized vector of $\tilde{h}$ with

$$h^{(k)} = \tilde{h}/\|\tilde{h}\|. \tag{3.11}$$

Optimal step length

The cost function decreases along the descent direction $h^{(k)}$. The next iteration point is generated by

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \gamma^{(k)}h^{(k)}, \quad \gamma^{(k)} > 0,$$

where $\gamma^{(k)}$ is the step length that is determined via the following optimization problem,

$$\min_{\lambda \geqslant 0} g(\gamma)$$

where

$$g(\gamma) = f(\boldsymbol{x}^{(k+1)}) = f(\boldsymbol{x}^{(k)} + \gamma h^{(k)}), \quad \gamma \geqslant 0.$$

Since f is convex, $g(\gamma)$ is also convex. Then, we can choose $\gamma^{(k)}$ as the optimal solution of the problem $\min\limits_{\gamma} g(\gamma)$. This problem is equivalent to the following problem

$$\begin{aligned} \max_{\gamma \geqslant 0} \quad & \gamma \\ \text{s.t.} \quad & \nabla_{h^{(k)}} f(\boldsymbol{x}^{(k)} + \gamma h^{(k)}) \geqslant 0, \\ & \nabla_{h^{(k)-}} f(\boldsymbol{x}^{(k)} + \gamma h^{(k)}) \geqslant 0. \end{aligned} \tag{3.12}$$

For this problem, we have the following observation (see Shi et al. [101]).

Lemma 3.1. *There exists an optimal solution $\gamma^{(k)} > 0$ and at least one i in $\{1, \cdots, n\}$ such that $A_i(\boldsymbol{x}^{(k)} + \gamma^{(k)} h^{(k)}) = y_i$, that is, i is in the zero set at the point $\boldsymbol{x}^{(k)} + \gamma^{(k)} h^{(k)}$ during the k-th iteration.*

Algorithm

Denote by $\gamma^{(k)}$ the optimum step length along the direction $h^{(k)}$ as described in section 3.4. The cost function is updated as

$$f(\boldsymbol{x}^{(k+1)}) = c^{(k+1)} \boldsymbol{x}^{(k+1)} + \sum_{o_{(k+1)i} \in \Omega_{k+1}} |A_{o_{(k+1)i}} \boldsymbol{x}^{(k+1)}| + z^{(k+1)}.$$

Remove the indexes in Λ_1^k from the zero set Ω_k and denote $\Omega_k^0 = \Omega_k \backslash \Lambda_1^k$. Let $\Lambda_2^k = \{o_{ki} | u_{o_{ki}} = 0\}$. Lemma 3.1 guarantees that Λ_2^k is non-empty. In the $(k+1)$-th iteration, set

$$\Omega_{k+1} = \Omega_k^0 \cup \Lambda_2^k,$$

and $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \gamma^{(k)} h^{(k)}$. Continue the above process until the optimal condition (3.7) is satisfied. To summarize, the algorithm is as follows:

Algorithm 3 Descent Algorithm for MAD-Lasso model

```
**Initialization:** Choose an initial point $\boldsymbol{x}^{(0)}$, compute the corresponding set $\Omega_0$, and compute the cost function $f(\boldsymbol{x}^{(0)})$. Set $k = 0$.
**Step 1: (Terminate)**
Generate the matrix $V$ for the zero set $\Omega_k$. If the condition (3.7) is satisfied, then stop and return the optimal solution and value. Otherwise, go to Step 2.
**Step 2: (Descent Direction)**
Find the $\alpha$ fastest descent directions as $\Lambda_1^k$, where $\alpha$ denotes the percentage of selected descent directions that decrease faster than the other $1-\alpha$ directions. Set $\Omega_k^0 = \Omega_k \backslash \Lambda_1^k$, and compute the descent direction $h^{(k)}$ using (3.10), (3.11).
**Step 3: (Optimal Step Length)** Find the best step length $\gamma^{(k)}$ by (3.12).
**Step 4: (Iteration)** Update $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \gamma^{(k)} h^{(k)}$. Find $\Lambda_2^k$ and update the zero set as $\Omega_{k+1} = \Omega_k^0 \cup \Lambda_2^k$. Then we compute the cost function $f(\boldsymbol{x}^{(k+1)})$ at $(k+1)$-th iteration, and then we go to Step 1.
```

3.3 Numerical experiments

In this section, simulation studies and real data analysis are carried out to compare (1) MAD, (2) MAD-Lasso with the proposed algorithm, and (3) MAD-Lasso with interior point method. The comparison is based on computational efficiency and the performance of the portfolio selection under the following risk measures:

(1) Expected Return (Mean): Mean = $\boldsymbol{w}'\boldsymbol{\mu}$.

(2) Sharpe Ratio: Sharpe = $\boldsymbol{w}'\boldsymbol{\mu}/\sqrt{\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}}$.

(3) Sparsity: number of nonzero entries of $\boldsymbol{w}$ (see [137]).

(4) Time: time consumption.

(5) Standard Deviation (StD): $\sigma = \sqrt{\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}}$.

(6) Mean Absolute Deviation (MAD): $\frac{1}{T}\sum_{t=1}^{T}\left|\sum_{i=1}^{n}(r_{ti} - r_i)w_i\right|$.

(7) Value at Risk (VaR): $\mathrm{VaR}_\alpha = \mu + \sigma\Phi^{-1}(1-\alpha)$.

(8) Expected Shortfall (ES): $\mathrm{ES}_\alpha = \mu + \sigma\frac{\psi(\Phi^{-1}(1-\alpha))}{\alpha}$.

Simulation studies

To investigate the performance of different portfolio selection method, consider two cases where the data are generated from the multivariate Gaussian distribution

	Gaussian Data						
Model	Time	Sparsity	Mean	StD	MAD	VaR	ES
MAD-Model	0.0576	10	0.3501	0.0663	0.0529	0.5042	0.5267
MAD-Lasso (IP)	0.1309	8	0.3501	0.0591	0.0470	0.4876	0.5076
MAD-Lasso (NEW)	0.0248	7	0.3501	0.0591	0.0470	0.4876	0.5076
	Asymmetric Laplace Data						
Model	Time	Sparsity	Mean	StD	MAD	VaR	ES
MAD-Model	0.0990	10	0.3497	0.3630	0.2730	1.1941	1.3171
MAD-Lasso (IP)	0.1277	6	0.3497	0.3540	0.2602	1.1733	1.2933
MAD-Lasso (NEW)	0.0138	6	0.3497	0.3540	0.2602	1.1733	1.2933

Table 3.1: Portfolio selection results of simulated datasets.

and multivariate Asymmetric Laplace distribution respectively, with the following parameter settings:

$$\boldsymbol{\mu} = (0.0001\,, 0.0002\,, 0.0003\,, 0.0004\,, 0.0005\,, 0.50\,, 0.60\,, 0.70\,, 0.80\,, 0.90).$$

$$\boldsymbol{\Sigma} = \mathrm{diag}\,(\boldsymbol{\mu}/10).$$

In both cases, (1) MAD, (2) MAD-Lasso with interior point method (IP), and (3) MAD-Lasso with the proposed method (NEW) as described in Section 3.2 are used for portfolio selection. The tuning parameter λ is chosen as $\lambda = \sqrt{2T \log n}$, as suggested in [114]. In both Gaussian case and asymmetric Laplace cases, 1000 replicates are performed in Matlab with an Intel (R) Core (TM) i7-4790 3.60 GHz Processor and 3.60 GHz memory. The interior point method is implemented using `linprog` provided in the MATLAB interface and the proposed method is programmed in MATLAB.

The results are shown in Table 3.1. MAD-Lasso outperforms MAD in terms of Sharpe Ratio, Sparsity and risk measures (StD, MAD, $\mathrm{VaR}_{0.01}$, $\mathrm{ES}_{0.01}$). Under the above-mentioned indicators, the performance of MAD-Lasso is similar for interior point method and the proposed descent algorithm. However, in terms of computational time, the proposed descent algorithm significantly outperforms the interior method. Figure 3.1-3.2 further display the boxplots of the computation results.

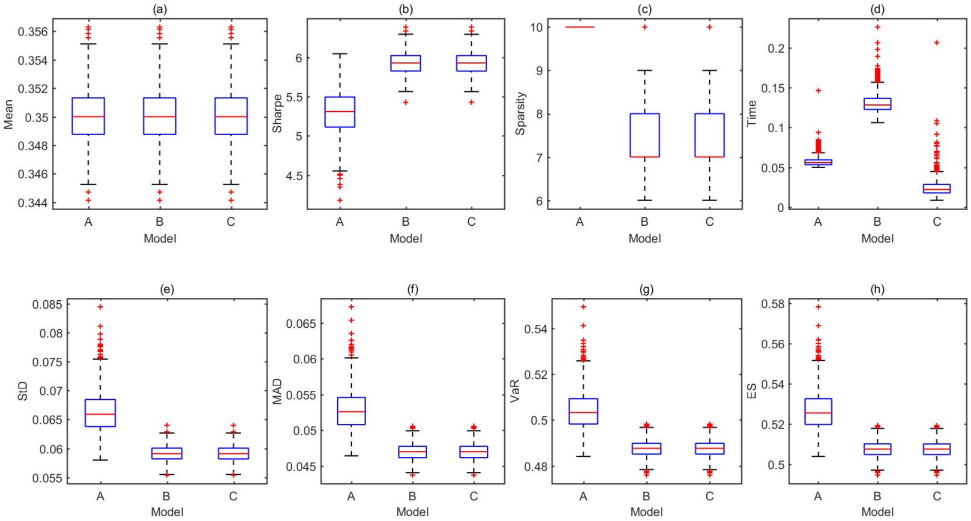

Figure 3.1: Portfolio selection results of Gaussian data A. MAD model; B. MAD-Lasso with interior point method (IP); C. MAD-Lasso with proposed method (NEW).

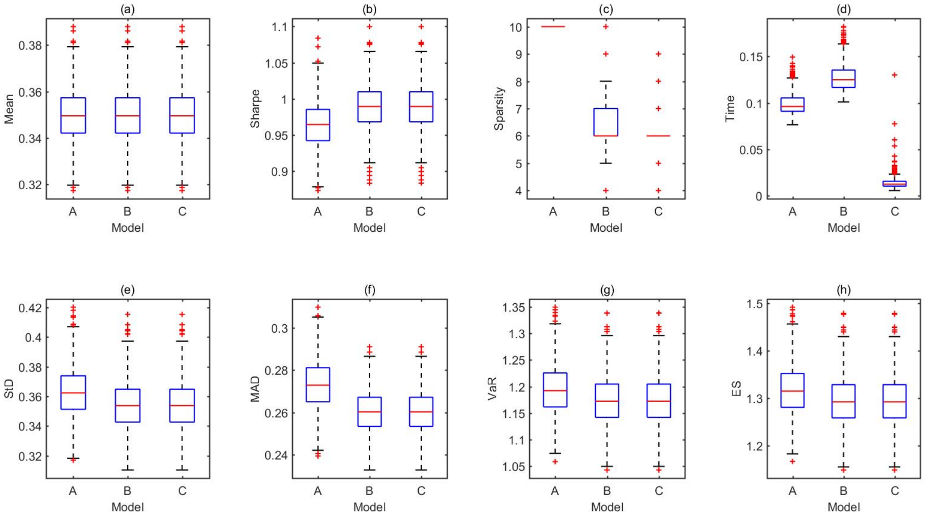

Figure 3.2: Portfolio selection results of Asymmetric Laplace data: A. MAD model; B. MAD-Lasso with interior point method (IP); C. MAD-Lasso with proposed method (NEW).

Real data analysis

Consider the datasets complied by Fama and French. Portfolios involving 48 industry sectors are obtained from both daily and monthly data (abbreviated to FF48d, FF48m) from June 1976 to June 2006. In both FF48d and FF48m datasets, the portfolios are constructed at the end of each June. Denote by $r_{i,t}$ the annualized return in time t of i-th industry, $i = 1, 2, \cdots, 48$.

Example 1. In this example, we compare the out-of-sample performances of MAD Model, Naive Evenly Model, and MAD-Lasso Models (IP, NEW).

For such a purpose, all portfolios are constructed by fixing the expected return at $r_0 = \bar{\boldsymbol{r}}$, where the target return r_0 as the average return achieved by the naive, evenly-weighted portfolio, computed from either the entire daily data or the entire monthly data. Consider the sequence of increasing tuning parameters $\lambda = 2^{-5:1:6}\sqrt{2T \log n}$ with $\lambda_1 = 2^{-5}\sqrt{2T \log n} = \frac{1}{32}\sqrt{2T \log n}\,, \lambda_2 = 2^{-4}\sqrt{2T \log n} = \frac{1}{16}\sqrt{2T \log n}\,, \cdots, \lambda_{12} = 2^{6}\sqrt{2T \log n} = 64\sqrt{2T \log n}$.

For both FF48d and FF48m datasets, we compare MAD Model (mad), Naive evenly-weighted model (naive), and MAD-Lasso (IP, NEW). The comparisons are based on computational time, Sparsity, Sharpe, MAD, $\mathrm{VaR}_{0.01}$ and $\mathrm{ES}_{0.01}$. Estimation results are reported in Table 3.2, 3.3. Results show that with increasing tuning parameter λ, we can achieve higher level of Sparsity and smaller Sharpe Ratio. Moreover, the values of MAD, $\mathrm{VaR}_{0.01}$, $\mathrm{ES}_{0.01}$ increase accordingly. For both datasets, MAD model outperforms Naive model in terms of Sharpe Ratio and risk. MAD-Lasso with smaller tuning parameters can achieve better performance than MAD Model. For MAD-Lasso models, results show that our proposed algorithm (NEW) is much more time efficient than interior point method (IP) with increasing tuning parameter λ. With properly chosen tuning parameters, MAD-Lasso models can achieve higher Sharpe Ratio and smaller risk than MAD model and Naive model. On the other hand, with increasing tuning parameter λ, MAD-Lasso models can achieve better performance than MAD Model by sacrificing a little bit Sharpe Ratio and risk.

	MAD and NAIVE Portfolio Selection Models					
	Time		Sharpe		Sparsity	
	mad	naive	mad	naive	mad	naive
–	3.8550	0.0064	0.0841	0.0632	47	48
	MAD-Lasso Portfolio Selection Models					
	Time		Sharpe		Sparsity	
	IP	NEW	IP	NEW	IP	NEW
λ_1	2.9822	2.3658	0.0852	0.0852	47	47
λ_2	3.4228	2.0967	0.0852	0.0852	47	47
λ_3	3.3466	2.0494	0.0851	0.0851	45	45
λ_4	3.4036	2.1634	0.0848	0.0848	36	36
λ_5	3.3688	2.4207	0.0843	0.0843	32	32
λ_6	3.6359	1.8651	0.0830	0.0830	25	24
λ_7	3.9089	2.8070	0.0811	0.0811	18	18
λ_8	4.2888	2.8444	0.0801	0.0801	13	13
λ_9	4.1552	2.8232	0.0801	0.0801	13	13
λ_{10}	4.6611	2.8770	0.0801	0.0801	13	13
λ_{11}	5.2082	2.2390	0.0801	0.0801	13	13
λ_{12}	4.7706	2.1307	0.0801	0.0801	13	11
	MAD and NAIVE Portfolio Selection Models					
	MAD		$\mathrm{VaR}_{0.01}$		$\mathrm{ES}_{0.01}$	
	mad	naive	mad	naive	mad	naive
–	0.4617	0.6069	1.5746	2.0784	1.7960	2.3732
	MAD-Lasso Portfolio Selection Models					
	MAD		$\mathrm{VaR}_{0.01}$		$\mathrm{ES}_{0.01}$	
	IP	NEW	IP	NEW	IP	NEW
λ_1	0.4457	0.4457	1.5556	1.5556	1.7742	1.7742
λ_2	0.4458	0.4458	1.5559	1.5559	1.7745	1.7745
λ_3	0.4461	0.4460	1.5576	1.5576	1.7765	1.7765
λ_4	0.4471	0.4471	1.5620	1.5620	1.7815	1.7815
λ_5	0.4501	0.4501	1.5716	1.5716	1.7925	1.7925
λ_6	0.4576	0.4576	1.5948	1.5948	1.8191	1.8191
λ_7	0.4696	0.4696	1.6313	1.6313	1.8609	1.8609
λ_8	0.4770	0.4770	1.6517	1.6517	1.8843	1.8843
λ_9	0.4770	0.4770	1.6517	1.6517	1.8843	1.8843
λ_{10}	0.4770	0.4770	1.6517	1.6517	1.8843	1.8843
λ_{11}	0.4770	0.4770	1.6517	1.6517	1.8843	1.8843
λ_{12}	0.4770	0.4771	1.6517	1.6513	1.8843	1.8839

Table 3.2: Portfolio selection results of FF48d Data:$(T, n) = (7573, 48)$; ExpRet: $r_0 = 0.0550$.

MAD and NAIVE Portfolio Selection Models						
	Time		Sharpe		Sparsity	
	mad	naive	mad	naive	mad	naive
–	0.9094	0.0001	0.4332	0.2450	48	48
MAD-Lasso Portfolio Selection Models						
	Time		Sharpe		Sparsity	
	IP	NEW	IP	NEW	IP	NEW
λ_1	0.3121	0.2698	0.4311	0.4324	48	48
λ_2	0.2449	0.1782	0.4317	0.4304	47	48
λ_3	0.2245	0.1363	0.4314	0.4316	47	48
λ_4	0.2763	0.1646	0.4308	0.4308	42	44
λ_5	0.2742	0.1154	0.4260	0.4265	40	39
λ_6	0.2485	0.0977	0.4189	0.4177	29	27
λ_7	0.2599	0.1042	0.3986	0.3985	19	19
λ_8	0.2856	0.1217	0.3734	0.3734	15	15
λ_9	0.2343	0.0932	0.3473	0.3473	10	10
λ_{10}	0.2590	0.0945	0.3473	0.3481	10	10
λ_{11}	0.3257	0.0867	0.3473	0.3481	10	10
λ_{12}	0.2760	0.0837	0.3473	0.3476	10	10
MAD and NAIVE Portfolio Selection Models						
	MAD		$\text{VaR}_{0.01}$		$\text{ES}_{0.01}$	
	mad	naive	mad	naive	mad	naive
–	2.1255	3.7957	8.0304	13.2301	9.0166	14.9736
MAD-Lasso Portfolio Selection Models						
	MAD		$\text{VaR}_{0.01}$		$\text{ES}_{0.01}$	
	IP	NEW	IP	NEW	IP	NEW
λ_1	2.0243	2.0247	8.0625	8.0419	9.0533	9.0298
λ_2	2.0251	2.0261	8.0542	8.0738	9.0438	9.0662
λ_3	2.0262	2.0281	8.0583	8.0557	9.0485	9.0455
λ_4	2.0341	2.0347	8.0675	8.0686	9.0590	9.0603
λ_5	2.0604	2.0600	8.1440	8.1359	9.1467	9.1374
λ_6	2.1144	2.1236	8.2615	8.2806	9.2813	9.3032
λ_7	2.2383	2.2389	8.6172	8.6197	9.6888	9.6916
λ_8	2.4150	2.4150	9.1145	9.1146	10.2586	10.2586
λ_9	2.6162	2.6162	9.7053	9.7053	10.9354	10.9354
λ_{10}	2.6162	2.6169	9.7053	9.6843	10.9354	10.9113
λ_{11}	2.6162	2.6169	9.7053	9.6843	10.9354	10.9113
λ_{12}	2.6162	2.6163	9.7053	9.6966	10.9354	10.9254

Table 3.3: Portfolio selection results of FF48m data: $(T, n) = (361, 48)$; ExpRet: $r_0 = 1.2606$.

Example 2. In this example, MAD-Lasso with interior point method (IP) and our proposed method (NEW) are compared. For FF48d Dataset ($[0.05 : 0.05 : 8]\sqrt{2T \log n}$), the tuning parameter is chosen from $0.05\sqrt{2T \log n}$ to $4\sqrt{2T \log n}$. For FF48m Dataset ($[0.05 : 0.05 : 8]\sqrt{2T \log n}$), the tuning parameter is chosen from $0.05\sqrt{2T \log n}$ to $8\sqrt{2T \log n}$. The interval length is chosen as $\Delta = 0.05\sqrt{2T \log n}$.

The results of the MAD-Lasso methods are displayed in Figure 3.3-3.4. Results show that the sparsity measure increases with larger tuning parameter λ and our proposed method is much more time efficient than the interior point method. Moreover, after some point when λ is large, the curves of Sharpe ratio, StD, VaR_α and CVaR_α with respect to λ behave like horizontal lines.

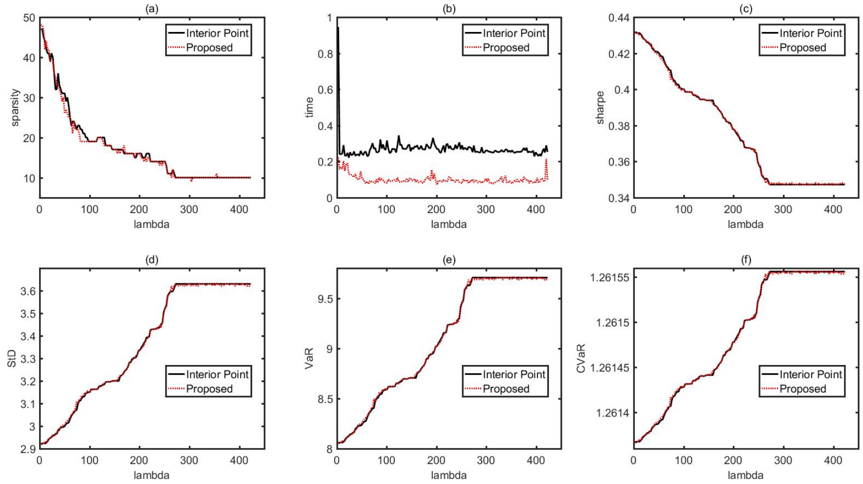

Figure 3.3: Portfolio selection tendency of FF48d data with increasing tuning parameter.

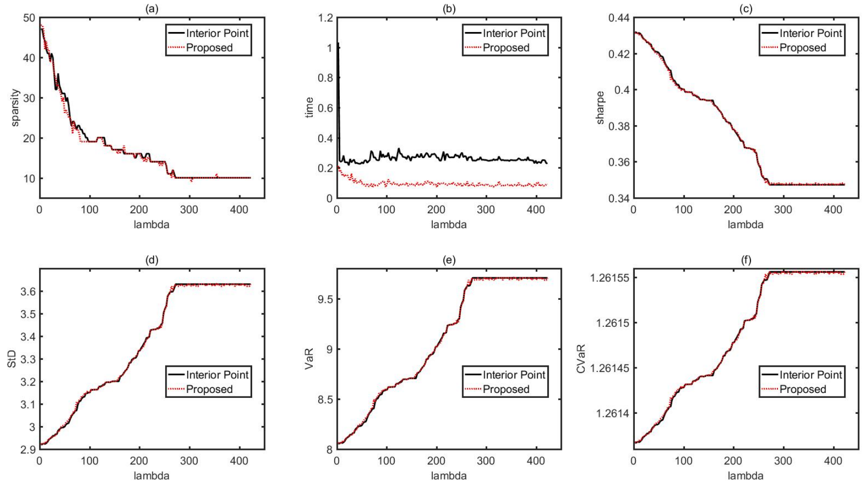

Figure 3.4: Portfolio selection tendency of FF48m data with increasing tuning parameter.

Chapter 4

Signal Processing

In this chapter, we derived the two-level optimization of penalty parameter selection for compressive sensing in signal processing problems.

4.1 Signal model background

Let the noisy signal be

$$x(n) = s(n) + v(n), \tag{4.1}$$

where $s(n)$ and $v(n)$ are the speech and noise signals, respectively. Its corresponding L-point STFT is given as

$$X(\omega, k) = \sum_{n=0}^{L-1} x(n)w(n - kR)e^{-j\omega n} = S(\omega, k) + V(\omega, k), \tag{4.2}$$

where $w(n - kR)$ is a time-limited window function with a hop size of R and length L, $\omega \in \omega_0, \cdots, \omega_{L-1}$ and k is the time index. The k-th instant data envelope of (4.2) is $|X(\omega, k)|$, where $| \cdot |$ denotes the absolute value operator.

Consider a $N \times N$ matrix $\mathbf{\Psi}$ whose columns form an orthonormal basis. The K-sparse signal, $\mathbf{x}(\omega, k) \in \mathbb{R}^N$ can then be given as

$$\mathbf{x}(\omega, k) = \mathbf{\Psi}(\omega)\theta(\omega, k), \tag{4.3}$$

where the N-length envelope vector $\mathbf{x}(\omega, k) = [|X(\omega, k)|, |X(\omega, k-1)|, \cdots, |X(\omega, k-N+2)|, |X(\omega, k-N+1)|\,]^\intercal$, the symbol $[\cdot]^\intercal$ is the transposition operator and $\theta(\omega, k) \in \mathbb{R}^N$ has K non-zero entries. The compressed measurement vector is given as

$$\mathbf{y}(\omega, k) = \mathbf{\Phi}(\omega)\mathbf{x}(\omega, k), \tag{4.4}$$

where $\mathbf{\Phi}(\omega)$ is a $M \times N$ sensing matrix/linear mapping matrix. In this instant, the sensing matrix compresses the signal's envelope for each frequency ω. Since $M \ll N$, this means that the dimension of $\mathbf{y}(\omega, k)$ is considerably smaller than $\mathbf{x}(\omega, k)$, hence the term "compressed". Equation (4.4) represents an alternative sampling procedure, which samples sparse signals close to their intrinsic information rate rather than their Nyquist rate. It has been shown that the tractable recovery of K-sparse signal, $\mathbf{x}(\omega, k)$ from the measurements, $\mathbf{y}(\omega, k)$ requires the sensing matrix, $\mathbf{\Phi}(\omega)$ to obey the restricted isometry property (RIP) [18]. Here, a sensing matrix, $\mathbf{\Phi}(\omega)$ is said to satisfy RIP of order K for all K-sparse signal, $\mathbf{x}(\omega, k)$, if there exists a constant, $\delta_K \in (0, 1)$ such that

$$(1 - \delta_K) \parallel \mathbf{x}(\omega, k) \parallel^2 \leqslant ||\mathbf{\Phi}(\omega)\mathbf{x}(\omega, k)||^2 \leqslant (1 - \delta_K) \parallel \mathbf{x}(\omega, k) \parallel^2, \tag{4.5}$$

CS recovery

One solution to ensure sparse recovery is to solve the following:

$$\hat{\mathbf{x}}(\omega, k) = \arg \min_{\mathbf{x}(\omega,k)} \|\mathbf{x}(\omega, k)\|_0 \qquad \text{s.t.} \qquad \mathbf{y}(\omega, k) = \mathbf{\Phi}(\omega)\mathbf{x}(\omega, k), \tag{4.6}$$

where $\|\mathbf{x}(\omega, k)\|_0$ is the number of non-zero components of $\mathbf{x}(\omega, k)$. However, solving (4.6) requires a combinatorial search, which is NP-hard [44]. A computational tractable solution to (4.6) is the widely known basis pursuit method as follows

$$\hat{\mathbf{x}}(\omega, k) = \arg \min_{\mathbf{x}(\omega,k)} \|\mathbf{x}(\omega, k)\|_1 \qquad \text{s.t.} \qquad \mathbf{y}(\omega, k) = \mathbf{\Phi}(\omega)\mathbf{x}(\omega, k), \tag{4.7}$$

where $\|\cdot\|_1$ is the ℓ_1 norm. Whilst the basis pursuit is a weaker formulation compared to (4.6), it allows efficient solution via linear programming techniques [44, 61]. A more flexible formulation, which allows for a trade-off between the exact congruence of $\mathbf{y}(\omega, k) = \mathbf{\Phi}(\omega)\mathbf{x}(\omega, k)$ and a sparser $\mathbf{x}(\omega, k)$ is the popular basis pursuit denoising [61] given as

$$\hat{\mathbf{x}}(\omega, k) = \arg \min_{\mathbf{x}(\omega,k)} \|\mathbf{y}(\omega, k) - \mathbf{\Phi}(\omega)\mathbf{x}(\omega, k)\|^2 + \lambda(\omega)\|\mathbf{x}(\omega, k)\|_1, \tag{4.8}$$

where $\|\cdot\|_2$ is the L_2-norm and $\lambda(\omega)$ is the regularization parameter. The formulation in (4.6) is a simple least-squares minimization process with a L_1-norm penalizer and the dictionary matrix $\mathbf{\Phi}(\omega)$. It is worth noting that since L_1-norm is non-differentiable, the optimization then leads to a decomposition which is sparser [21]. Simply, the first term in Eqn. (4.8) is to reduce the mean square area whilst the regulator seeks a sparser solution.

Note that the optimal solution tends to trivial as $\lambda(\omega) \to \infty$ [61]. A higher value of $\lambda(\omega)$ would generally result in a sparser solution since the ℓ_1-norm is being penalized more heavily. This means that the regularizer, $\lambda(\omega)$, penalizes the sum of the observed signal. In other words, the solution to (4.8) is indeed a function of $\lambda(\omega)$, i.e., fixing $\lambda(\omega)$ is equivalent to setting it to a particular subset of sparse solution for the least squares to be performed on [109]. Simply, the optimization problem is a trade-off between a quadratic misfit error (mean square error) against the sparsity of the data, i.e., ℓ_1-norm [22]. Clearly, if the incoming signal is already sparse, then $\lambda(\omega)$ can be relaxed and vice versa. Since the sparsity of the signal varies as a function of frequency, the regularizer should ideally vary according to the signal's profile.

A good choice of $\lambda(\omega)$ should provide a reasonable trade-off between the smoothness of the reconstructed signal and similarity to the original signal [84]. Nevertheless, it remains not so straightforward to set the regularization parameter $\lambda(\omega)$ and thus far, $\lambda(\omega)$ has been empirically determined. In practice, $\lambda(\omega)$, should be set according to the sparsity of the actual signal as $\lambda(\omega)$ controls the amount of regularization that can be imposed. It is precisely this quality control that this paper seeks to establish, i.e., by linking sparsity to quality. Since a larger value of $\lambda(\omega)$ yields a sparser solution, then more noise would be suppressed. However, how much can $\lambda(\omega)$ be set before the signal quality is compromised.

Quality measures

In a big data setting such as speech signals, this paper seeks to subsume the affective design by hyper-parameterizing λ via the Gini index and the model selectors, Akaike information criterion (AIC) and Bayesian information criterion (BIC). The set of solutions is then evaluated with respect to PESQ. In particular, λ is to be optimized in such a way that the sparsest solution yields the one with the best quality in terms of noise suppression and target distortion. In this case, the noise suppression and speech distortion can be viewed as the engineering requirement and the affective design attribute, respectively. The idea is to incorporate affective design via the influence of the key design parameter on the aforementioned PESQ measure. By doing so, the parameter can be translated to consumer reactions (via the PESQ measure).

We propose a two-level optimization strategy to optimize $\lambda(\omega)$ to affective measure. In the inner level, the big data is first compressed via the sensing matrix, $\mathbf{\Phi}(\omega)$. In the outer level or the sparse reconstruction stage, the hyperparameter is optimally chosen to incorporate the overall signal quality. Quality measures such as the AIC, BIC and Gini index are used to optimize the value of the hyperparameter. These measures are explicitly used to determine the relationship between key design parameters with the consumer reactions from the processed signal. The following sections explain each of the chosen optimization criteria, namely Gini index, the

AIC and BIC model selection methods.

Gini index

As mentioned, the actual sparsity of the signal affects the performance sparse recovery. As an effective measure of sparsity, Zonoobi [135] concluded that the Gini index can induce a significantly improved performance in reconstruction from compressive samples. A signal is considered most sparse if a signal can be represented by only one non-zero coefficient with the rest being zero [54]. Similarly, if a signal has only one high value non-zero coefficient amidst a low non-zero coefficients, then the signal can be said to be most sparse. In essence, sparsity is a measure of disparity, i.e., the relative distribution of the coefficients of a signal is. A non-sparse signal on the other hand is described as having a uniform non-zero coefficients throughout. Of the many sparsity measures, it has been shown that Gini index remains the most consistent and fulfil all of the desirable sparsity criteria [54, 135].

Consider a M long ordered vector, $\mathbf{w} = [w_1, \cdots, w_M]$ such that $w_M \geqslant w_{M-1}, \cdots,$ $w_2 \geqslant w_1$, then the Gini coefficient is defined as

$$\mathsf{GI}(\mathbf{w}) = 1 - 2 \sum_{m=1}^{M} \frac{w_m}{\|\mathbf{w}\|_1} \left(\frac{M - m + 0.5}{M} \right). \tag{4.9}$$

A zero-valued Gini represents perfect equality whilst a close to unity value shows the opposite. In sparsity terms, a larger Gini coefficient shows a sparser signal. As such, Gini coefficient can be used as a measure to ascertain if a signal is sparse. Table 4.1 tabulates the Gini coefficients for three types of noise, speech and the noisy speech at different SNR levels. The coefficients show speech indeed is the sparsest signal in comparison with the other noise signals. Note that, of all the noise signals, babble noise has the highest Gini coefficient, owing to its speech-like nature. For the case of noisy speech signals, it can be seen that as the SNR increases, the Gini

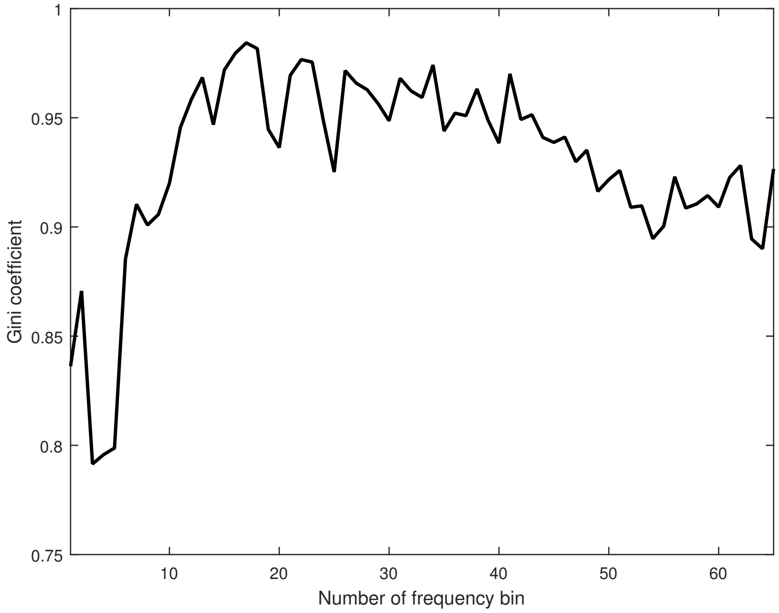

Figure 4.1: *The Gini coefficients for different frequency bins.*

coefficient approaches unity. As the SNR decreases, the value of the Gini coefficient drops accordingly. This simple example demonstrates that a sparser signal tends to have a higher SNR and as the signal becomes more noisy, sparsity reduces. Figure 4.1 shows that the sparsity of a speech signal varies as a function of frequency. It can be seen that the mid to high frequency range of a speech signal tend to be sparser compared to the low frequency components. Thus, by properly optimizing $\lambda(\omega)$ based on the Gini coefficient, the sparse reconstruction could potentially lead to better SNR improvement, as appropriate tuning parameter can be set according to the sparsity of the signal in question. As speech is highly non-stationary across time and frequency, its sparsity level would also vary accordingly. From Figure 4.1 the Gini index for the three noisy speech varies as a function of speakers and frequency, thus the λ will need to be re-estimated every N samples.

Signal	Gini coefficient	SNR	Speech + Babble	Speech + White	Speech + Destroyerops
Speech	0.9266	0	0.7522	0.7382	0.7372
Babble	0.6634	5	0.8302	0.8234	0.8239
White	0.6352	10	0.8848	0.8823	0.8828
Destroyerops	0.6243	15	0.9108	0.9099	0.9104

Table 4.1: The Gini coefficients for speech and different types of noise and at different SNRs.

Selection of $\lambda(\omega)$ based on Gini

Consider an N-length signal, $\mathbf{x}(\omega, k)$, then from Eqn. (4.8), its sparse reconstruction is given as

$$\hat{\mathbf{x}}(\omega, k) = \arg \min_{\mathbf{x}(\omega,k)} \|\mathbf{y}(\omega, k) - \mathbf{\Phi}(\omega)\mathbf{x}(\omega, k)\|^2 + \lambda(\omega)\|\mathbf{x}(\omega, k)\|_1. \tag{4.10}$$

For each given value of $\lambda(\omega)$ value, an estimation of $\hat{\mathbf{x}}(\omega, k)$ is denoted as $\hat{\mathbf{x}}_{\lambda(\omega)}(\omega, k)$. The Gini coefficient of $\hat{\mathbf{x}}_{\lambda(\omega)}(\omega, k)$ is then defined as

$$\mathsf{GI}(\hat{\mathbf{x}}_{\lambda}(\omega, k)) = 1 - 2\sum_{n=1}^{N} \frac{\hat{x}_{\lambda(\omega)}(\omega, k, n)}{\|\hat{\mathbf{x}}_{\lambda(\omega)}(\omega, k)\|_1} \left(\frac{N - n + 0.5}{N}\right), \tag{4.11}$$

where $\hat{x}_{\lambda(\omega)}(\omega, k, n)$ is the n-th ordered value of vector $\hat{\mathbf{x}}_{\lambda(\omega)}(\omega, k)$ in a descending order. The corresponding optimization problem of maximizing the GI coefficients can be written as

$$\lambda_{\mathsf{maxGini}}(\omega) = \arg \max_{\lambda(\omega)} \mathsf{GI}(\hat{\mathbf{x}}_{\lambda(\omega)}(\omega, k)), \tag{4.12}$$

where $\mathsf{GI}(\hat{\mathbf{x}}_{\lambda(\omega)}(\omega, k))$ is given in Eqn. (4.11). Equivalently, the optimization formulation for finding $\lambda(\omega)$ for the minimum Gini index is

$$\lambda_{\mathsf{minGini}}(\omega) = \arg \min_{\lambda(\omega)} \mathsf{GI}(\hat{\mathbf{x}}_{\lambda(\omega)}(\omega, k)). \tag{4.13}$$

Eqns. (4.12) and (4.13) can be viewed as the extreme ends of compressive speech enhancement, as Eqn. (4.12) recovers the sparsest signal it could possibly tuned and vice versa for Eqn. (4.13). In the numerical experiments to follow, we will show that both the optimization above behaves very differently for the PESQ and segmental SNR measures, with Eqn. (4.12) leaning towards noise suppression and Eqn. (4.13) acting towards more on speech preservation.

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

Whilst the tuning parameter selection based on Gini criterion is intuitive, it is by no means the only approach. For any regularization method, finding the best regularization parameter is essential. As explained by Dicker et al. [28], the estimators are typically found to correspond to a range of tuning parameter values, which is referred to as a solution path. Subsequently, the preferred estimator is identified along the solution path as the estimator, which fits the optimization criteria. In the same vein, this paper considers the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) based approach for the selection of the tuning parameter, $\lambda(\omega)$ [99]. It is well known that AIC and BIC are popular model selection criteria. As shown in Zou [137], AIC and BIC possess different asymptotic optimality. AIC converges at the minimax optimal rate to the true regression mode, whereas BIC is consistent in selecting the true model. In this case, we ascertain the heuristics usefulness of both the AIC and BIC in tuning $\lambda(\omega)$, for compressive speech enhancement. The major difference between AIC and BIC is that they possess different asymptotic optimality [137]. For AIC ([2]), it seeks the model with the least average squared error irrespective of whether the true model is in the candidate list. BIC, on the other hand, guarantees in selecting the true model, should the true model be selectable. Readers may refer to [28, 137, 86] for in-depth view of the two approaches.

Let us define the residual sum of squares (RSS) as

$$\mathsf{RSS} = \left\|\mathbf{y}(\omega, k) - \mathbf{\Phi}(\omega)\mathbf{x}(\omega, k)\right\|^2 . \tag{4.14}$$

From [137], given an estimator $\hat{\mathbf{x}}$, the number of nonzero entries of an estimator $\hat{\boldsymbol{x}}$

is an unbiased estimate of the degree of freedom (df), that is

$$\mathsf{df} = \text{number of nonzero entries of } \hat{\mathbf{x}}(\omega, k). \tag{4.15}$$

AIC and BIC are usually used to make model selection and predict models, both of them could be represented as a combination of a likelihood term and a penalty term. Thus from Eqn. (4.14) and (4.15), the corresponding AIC and BIC can be formulated as

$$\begin{aligned} \mathsf{AIC} &= \ell \log(\mathsf{RSS}/\ell) + 2\mathsf{df}, && (4.16) \\ \mathsf{BIC} &= \ell \log(\mathsf{RSS}/\ell) + \mathsf{df} \cdot \log(\ell), && (4.17) \end{aligned}$$

where ℓ is the length of estimator $\hat{\mathbf{x}}$. The tuning parameter selection procedure can be reduced to the minimization of AIC or BIC, and as discussed previously, AIC is comparatively more conservative in its variable selection. Inserting Eqn. (4.14) into (4.16) and (4.17), respectively, yields the $\lambda(\omega)$ selection as follows:

$$\begin{aligned} \lambda_{\mathsf{AIC}}(\omega) &= \arg\min_{\lambda(\omega)} n \log(\|\mathbf{y}(\omega, k) - \mathbf{\Phi}(\omega)\hat{\mathbf{x}}_{\lambda}(\omega, k)\|^2/n) + 2\mathsf{df}, && (4.18) \\ \lambda_{\mathsf{BIC}}(\omega) &= \arg\min_{\lambda(\omega)} n \log(\|\mathbf{y}(\omega, k) - \mathbf{\Phi}(\omega)\hat{\mathbf{x}}_{\lambda}(\omega, k)\|^2/n) + \mathsf{df} \log(n). && (4.19) \end{aligned}$$

Perceptual Evaluation of Speech Quality

Broadly, the assessment of speech quality can be classified as subjective and objective evaluation. As the name implies, subjective evaluation involves subjective listening test by some listeners. Objective evaluation on the hand, measures the numerical distance between the reference signal and the processed signals [79]. One established method of evaluating how good the enhancement process is via the use perceptual evaluation of speech quality (PESQ). PESQ is an automated computation algorithm developed by the International Telecommunications Union (ITU) to replace human subjects in the evaluation of the mean opinion score (MOS). The PESQ model considers how human perceive speech and it has been used widely in the evaluation of speech quality [12]. PESQ is defined mathematically as [80]

$$\text{PESQ} = a_0 + a_1 d_{sym} + a_2 d_{asym}, \tag{4.20}$$

where $a_0 = 4.5$, $a_1 = -0.1$ and $a_2 = -0.0309$. The variables d_{sym} and d_{asym} are the average disturbance values for the symmetrical and asymmetrical components. The former measures the distortion due to noise and the latter describes the omission of the actual speech.

PESQ bypasses the need for human subjects to take part in the evaluation process and can be used as part of the affective design process. Numerous studies have shown that PESQ consistently rated to be the most reliable objective measure for speech quality assessment [78, 50]. In fact, PESQ has also been shown to be consistent in measuring speech intelligibility [85]. As PESQ gives the overall speech quality score, consequently, it is regarded as an affective indicator as to how "pleased" the consumers are with the processed speech.

4.2 Proposed two-level optimization process

This section details the proposed two-level optimization strategy to optimize $\lambda(\omega)$ with respect to the quality measures. In the first level optimization, the big data is first compressed via the sensing matrix, $\mathbf{\Phi}(\omega)$. The second level then optimizes the hyperparameter through the quality measures, which then improves the overall signal affective's quality.

First level optimization: compressive sensing

The first step entails the compressive sensing matrix selection. The data compression from Equation (4.4) is reproduced here for convenience

$$\mathbf{y}(\omega, k) = \mathbf{\Phi}(\omega)\mathbf{x}(\omega, k), \tag{4.21}$$

where $\mathbf{y}(\omega, k) \in \mathbb{R}^M, \mathbf{x}(\omega, k) \in \mathbb{R}^N$, and $\mathbf{\Phi}(\omega) \in \mathbb{R}^{M\times N}$ is the compressive sensing matrix, which compresses the signal dimension by projecting the signal from $\mathbb{R}^N$ into $\mathbb{R}^M$, where $M \ll N$. The sensing matrix is typically generated by using a random Gaussian matrix or a partial DCT matrix [84].

Under the Restricted Isometry Property condition (4.5), the solution to (4.21) can be solved by using the popular basis pursuit as follows ([61]):

$$\hat{\mathbf{x}}(\omega, k) = \arg \min_{\mathbf{x}(\omega,k)} \|\mathbf{x}(\omega, k)\|_1 \text{ s.t. } \mathbf{y}(\omega, k) = \mathbf{\Phi}(\omega)\mathbf{x}(\omega, k). \tag{4.22}$$

Alternatively, Equation (4.22) can be viewed as a linear regression

$$\mathbf{y}(\omega, k) = \mathbf{\Phi}(\omega)\mathbf{x}(\omega, k) + \varepsilon, \text{ s.t. } \|\mathbf{x}(\omega, k)\|_1 \leqslant \nu, \tag{4.23}$$

where ν is a constant relating to the sparsity constraint and $\varepsilon \in \mathbb{R}^M$ is the intercept or error. Thus Equation (4.22) can be reposed as the following

$$\min_{\mathbf{x}(\omega,k)} \|\mathbf{y}(\omega, k) - \mathbf{\Phi}(\omega)\mathbf{x}(\omega, k)\|^2 + \lambda(\omega)\|\mathbf{x}(\omega, k)\|_1, \tag{4.24}$$

where $\lambda(\omega)$ is the tuning hyperparameter. The solution to Equation (4.24) is the key to finding the best affective solution to the problem in question. Here, the $\lambda(\omega)$ plays a key role in mapping the solution to the affective measures. The following section explains how the solution to (4.24) is optimized with respect to the affective measures as discussed in the previous section.

Second level optimization: hyperparameter selection

To solve model (4.24), we implement the interior point method for large-scale l_1 regularized least squares algorithm in [61] with the following properties:

(i) When $\lambda(\omega) \to 0$, the estimator has the limiting behavior with (4.24), satisfying $\mathbf{\Phi}(\omega)^\intercal[\mathbf{\Phi}(\omega)\mathbf{x}(\omega, k) - \mathbf{y}(\omega, k)] = 0$.

(ii) As $\lambda(\omega) \to \infty$, the estimator shrinks to the zero vector, $\mathbf{0}$. The convergence occurs for a finite value of $\lambda(\omega)$, i.e., $\lambda(\omega) \geqslant \lambda_{max}(\omega) = \|2\mathbf{\Phi}(\omega)^\intercal \mathbf{y}(\omega, k)\|_\infty$, where $\|\mathbf{x}\|_\infty = \max_i |x_i|$ is the l_∞ norm of vector $\mathbf{x}$. However, for $\lambda(\omega) > \lambda_{max}(\omega)$, the optimal solution of (4.24) is trivial, i.e., $\mathbf{0}$.

(iii) As λ varies across $(0, \infty)$, the solution path of $\mathbf{x}$ is piecewise linear. That is, with tuning parameters satisfy $0 = \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_k = \lambda_{max}$, the regularization path of $\mathbf{x}$ is a piecewise linear curve on $\mathbb{R}^N$:

$$\mathbf{x} = \frac{\lambda_{i+1} - \lambda}{\lambda_{i+1} - \lambda_i}\mathbf{x}^{(i)} + \frac{\lambda - \lambda_i}{\lambda_{i+1} - \lambda_i}\mathbf{x}^{(i)}, \lambda_i \leqslant \lambda \leqslant \lambda_{i+1}, i = 1, 2, \cdots, k-1.$$

(iv) Clearly as a general rule, with properly chosen $\lambda(\omega)$, Equation (4.24) will result in a sparse solution.

(v) The computational complexity of this algorithm is determined by the product of the total number of Preconditioned Conjugate Gradient (PCG) steps during all iterations and the cost of a PCG step. As noted in [61], extensive testing suggest that the total number of PCG steps vary from a few tens to several hundreds to compute a solution. The computational complexity of a PCG step is $\mathcal{O}(NM)$, where M, N are the dimensions of sensing matrix $\mathbf{\Phi}(\omega)$. Then the total computational complexity is at most $\mathcal{O}(cNM)$, where c is the number of iterations in the order of hundreds.

We propose a grid search tuning parameter selection based on minimizing/maximizing the AIC, the BIC and the Gini index. Here, a set of $\lambda(\omega)$ is set as in interval length of 0.01 as $\lambda(\omega) = \{\lambda_1(\omega), \lambda_2(\omega), \cdots, \lambda_{100}(\omega)\}$ where $\lambda_1(\omega) = 0.01, \lambda_2(\omega) = 0.02, \cdots, \lambda_{100}(\omega) = 1$. For each fixed $\lambda_i(\omega)$, we can obtain $\hat{\mathbf{x}}_{\lambda_i(\omega)}$ by optimizing (4.24). Note that for a high-dimensional least squares Lasso problem, it is computationally expensive to implement through the Newton system. In order to balance between computation and convergence rate we propose to use the iterative method to solve the Newton system by using the truncated Newton method combined with interior point method [61]. From Eqn. (4.11), (4.18) and (4.19), we have

$$\begin{aligned}
\mathsf{AIC}(\lambda_i(\omega)) &= \ell \log(\|\mathbf{y}(\omega,k) - \mathbf{\Phi}(\omega)\hat{\mathbf{x}}_{\lambda_i(\omega)}(\omega,k)\|^2/\ell) + 2\mathsf{df}, && (4.25)\\
\mathsf{BIC}(\lambda_i(\omega)) &= \ell \log(\|\mathbf{y}(\omega,k) - \mathbf{\Phi}(\omega)\hat{\mathbf{x}}_{\lambda_i(\omega)}(\omega,k)\|^2/\ell) + \mathsf{df}\cdot\log(\ell), && (4.26)\\
\mathsf{GI}(\lambda_i(\omega)) &= 1 - 2\sum_{n=1}^{N} \frac{\hat{x}_{\lambda_i(\omega)}(\omega,k,n)}{\|\hat{\mathbf{x}}_{\lambda_i(\omega)}(\omega,k)\|_1}\left(\frac{N-n+0.5}{N}\right). && (4.27)
\end{aligned}$$

From the above, each optimized parameter can be found as $\lambda_i(\omega) \in \lambda(\omega)$ as follows

$$\begin{aligned}
\lambda_{\mathsf{MinAIC}} &= \arg\min_{\lambda(\omega)} \mathsf{AIC}\{\lambda(\omega)\}, && (4.28)\\
\lambda_{\mathsf{MinBIC}} &= \arg\min_{\lambda(\omega)} \mathsf{BIC}\{\lambda(\omega)\}, && (4.29)\\
\lambda_{\mathsf{MinGI}} &= \arg\min_{\lambda(\omega)} \mathsf{GI}\{\lambda(\omega)\}, && (4.30)\\
\lambda_{\mathsf{MaxGI}} &= \arg\max_{\lambda(\omega)} \mathsf{GI}\{\lambda(\omega)\}. && (4.31)
\end{aligned}$$

Finally, the corresponding optimal estimators are obtained as

$$\hat{\mathbf{x}}(\lambda(\omega)_{\mathsf{MinAIC}}), \hat{\mathbf{x}}(\lambda(\omega)_{\mathsf{MinBIC}}), \hat{\mathbf{x}}(\lambda(\omega)_{\mathsf{MinGI}}), \hat{\mathbf{x}}(\lambda(\omega)_{\mathsf{MaxGI}}). \quad (4.32)$$

Each optimal estimator is then evaluated against the affective measures, i.e., PESQ and segSNR. As mentioned the proposed approach is a grid based ratio selection method to optimize $\lambda(\omega)$. Here, the optimized $\lambda(\omega)$ is chosen based on the optimization of either on the Gini index, AIC and BIC criterion as shown above. In the following numerical study, we investigate the influence of hyperparameterizing $\lambda(\omega)$ on the results of compressive speech enhancement in terms of perceptual evaluation of speech quality (PESQ) and the segmental SNR (segSNR). Generally speaking, PESQ measures the overall improvement in the perceptibility of the speech signal, whereas segmental SNR rests more heavily on the suppression of noise in the observation.

4.3 Numerical experiments

Experiment settings: Four different types of noise sources from the NOISEX database, namely, babble, subway, destroyer and car noise were tested over a wide range of SNR, from 0dB to 20 dB, with similar SNR setting as in [84]. The noise types were chosen to represent the different degree of non-stationarity noise encountered in the real world. Five female and five male speech signals from the TIMIT database were used as stimuli. The performance was evaluated by using the segmental SNR and the PESQ measure with a total of five female and five male speech signals from the TIMIT database. As mentioned in the introduction, PESQ measure is an automated evaluation process, which in this case a key measure for the inclusion of affective design. The PESQ score reveals how good or bad the perceptual quality of the audio signal to a human listener. This paper also includes the objective measure segmental SNR as a comparison. The number of frequency points was fixed at 256 with 50% oversampling and the compressive ratio, M/N was set to 0.9.

Hyperparameterizing λ based on Gini, AIC and BIC criterion

Four criteria based on Equations (4.12), (4.13), (4.18) and (4.19) were used to examine the influence of $\lambda(\omega)$ on compressive speech enhancement for a range of SNRs. In this case, each of the criteria is evaluated in each frequency band via grid search. We take fixed $\lambda(\omega) = 0.1$ for comparison purposes as the same implementation in [84]. Figures 4.2, 4.3, 4.4 and 4.5 show the PESQ and segmental SNR performance of the four model selection criterion for babble noise, car noise, subway noise and destroyer noise, respectively. Evidently, the role of $\lambda(\omega)$ is crucial as its variation results in a very different performance across the SNRs.

In terms of PESQ, the minimization of the Gini and AIC criterion provide a consistent performance across the SNR range for the different types of noise. Both the criterion achieves higher PESQ values over the performance of having a fixed value of $\lambda(\omega)$ e.g., $\lambda(\omega) = 0.1$ (see [84]) and the unprocessed observation. Note that minimization of the Gini index results in the most non-sparse solution in the set of sparse solution. This means that the recovery process emphasizes on maintaining the speech signal as opposed to the reduction of noise (via a sparser solution). Interestingly, the minimization of BIC does not provide much improvement when the SNR> 10dB. Also, when compared to the AIC criterion, BIC obtains lower PESQ improvement but a higher segmental SNR improvement. This corroborates with the fact that in general, BIC tends to choose a parsimonious model compared to AIC. Hence for compressive speech enhancement, AIC is more inclined to select a model with less sparsity. This explains why AIC criterion results in a higher PESQ score but a lower segmental SNR compared to the BIC criterion.

In terms of segmental SNR improvement, the maximization of the Gini index gains the highest improvement with an approximately 4dB gain over the range of SNRs and the different types of noise. This is because the maximization of the Gini index results in the sparsest representation, which as shown in Section 4.1 is often the ones with the highest SNR. However, having an SNR improvement does not necessarily translate to overall speech intelligibility improvement. This is shown by the corresponding results in terms of the PESQ, where the maximization of Gini index attains the lowest PESQ improvement. This indicates that maxGINI maximally suppresses noise at the expense of the perceptual aspects of the output. This may be suitable for applications such as speech recognition where noise is the main issue. However, for hearing instruments such as assistive listening devices, SNR may not be the primary factor as improving SNR does not necessarily improve the perceptual part of speech as measured by PESQ. The proposed method allows such tuning by choosing the different criterion for the application in question. In a way it effectively parameterizes the sparse reconstruction through $\lambda(\omega)$ to allow for an engineering trade-off between noise suppression and perceptual preservation. Informal listening test confirms the improvement with respect to the different criteria used.

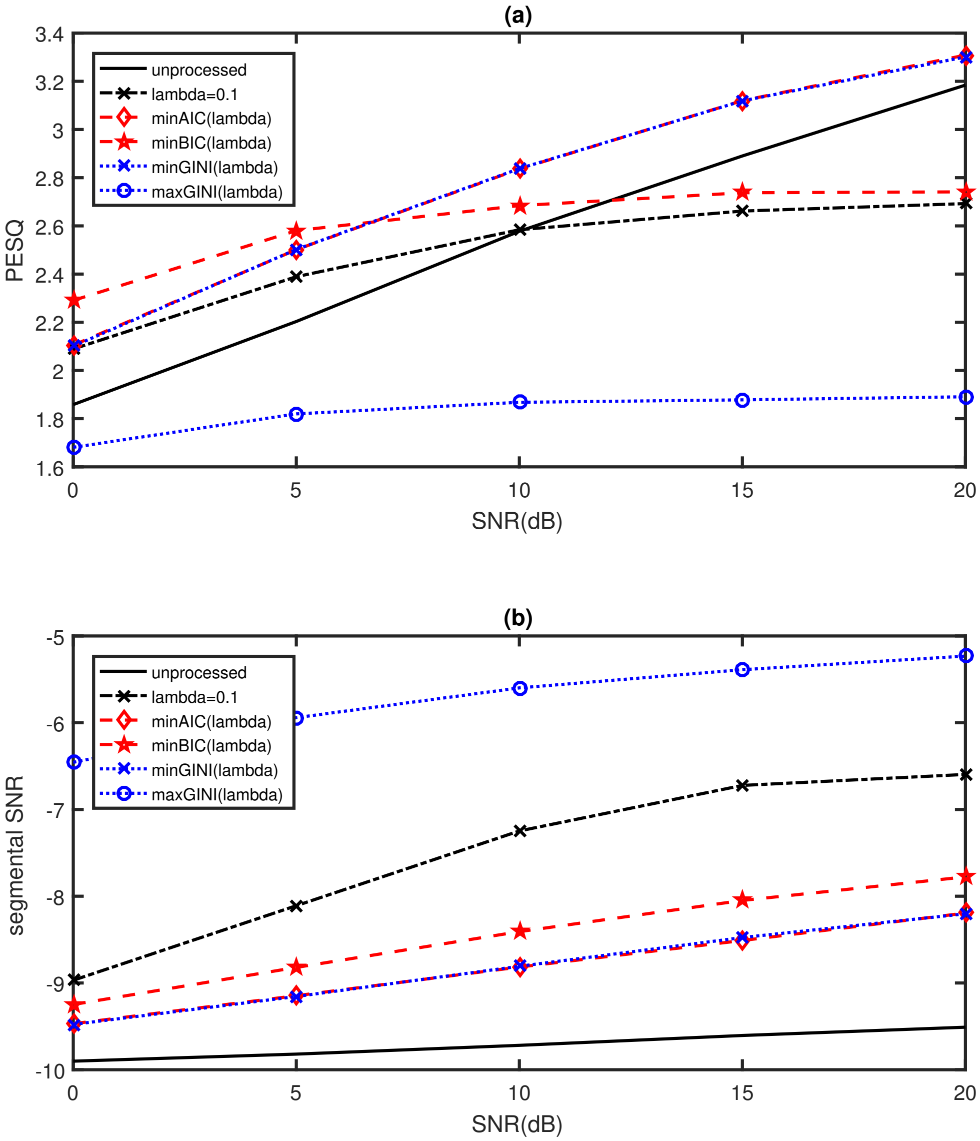

Figure 4.2: The (a) PESQ and (b) segmental SNR of the different hyperparameter optimization methods as a function of SNRs for babble noise.

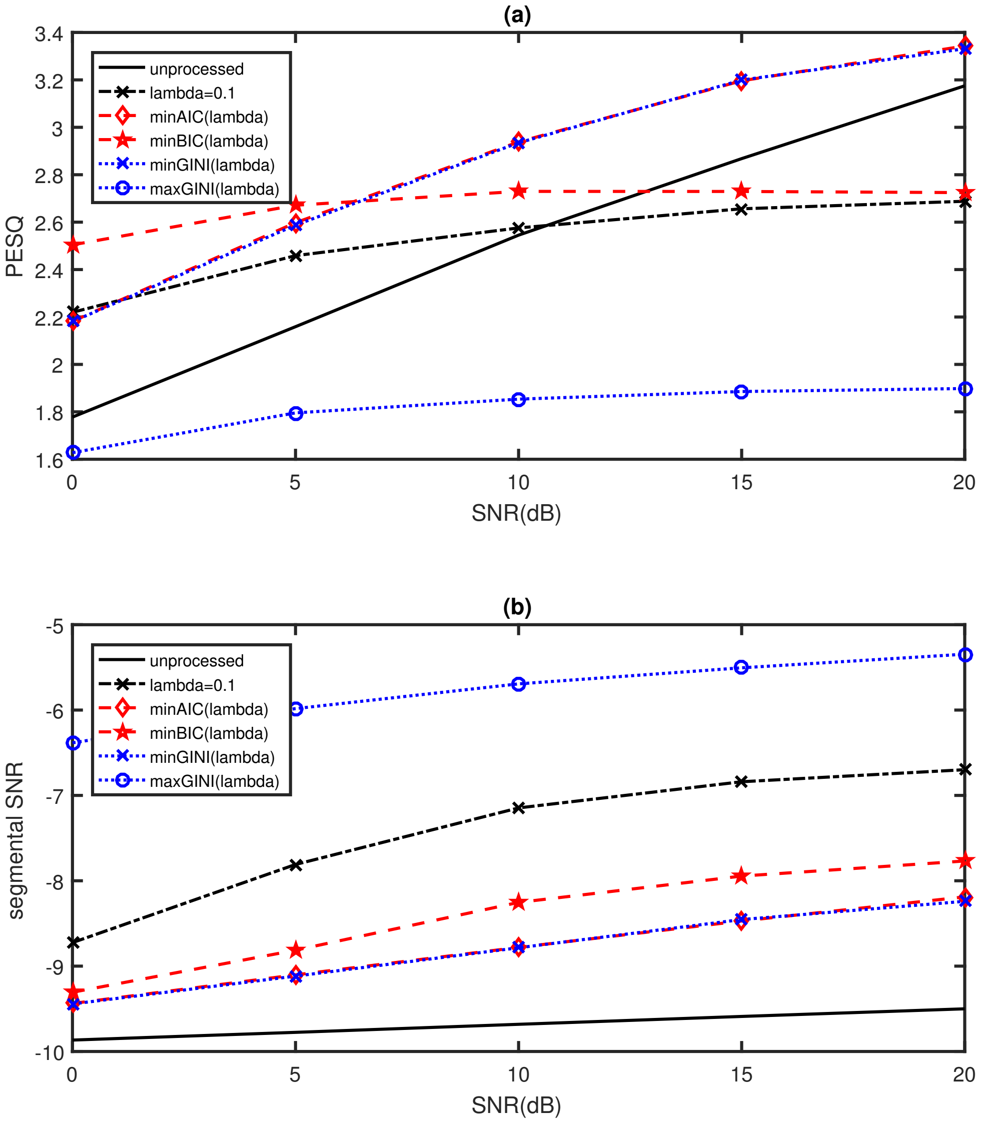

Figure 4.3: The (a) PESQ and (b) segmental SNR of the different hyperparameter optimization methods as a function of SNRs for car noise.

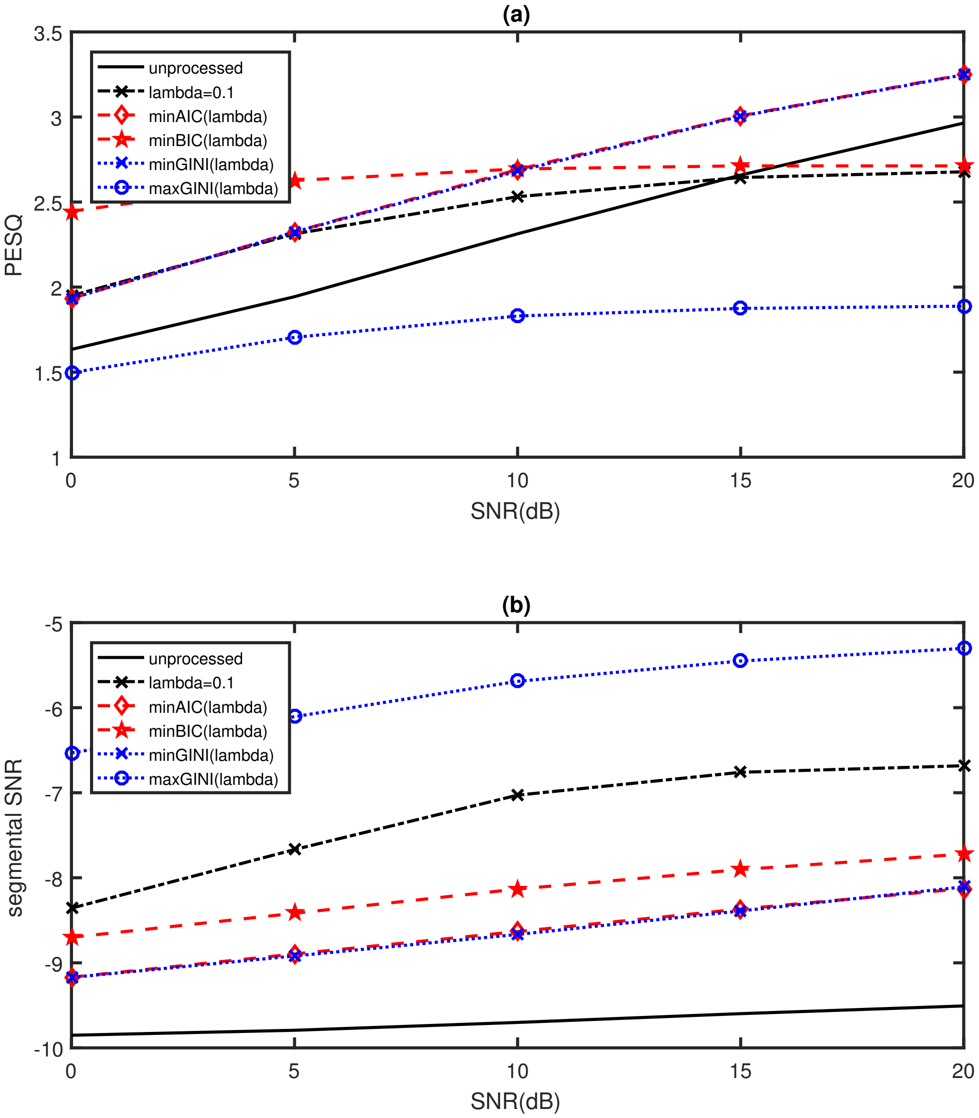

Figure 4.4: The (a) PESQ and (b) segmental SNR of the different hyperparameter optimization methods as a function of SNRs for subway noise.

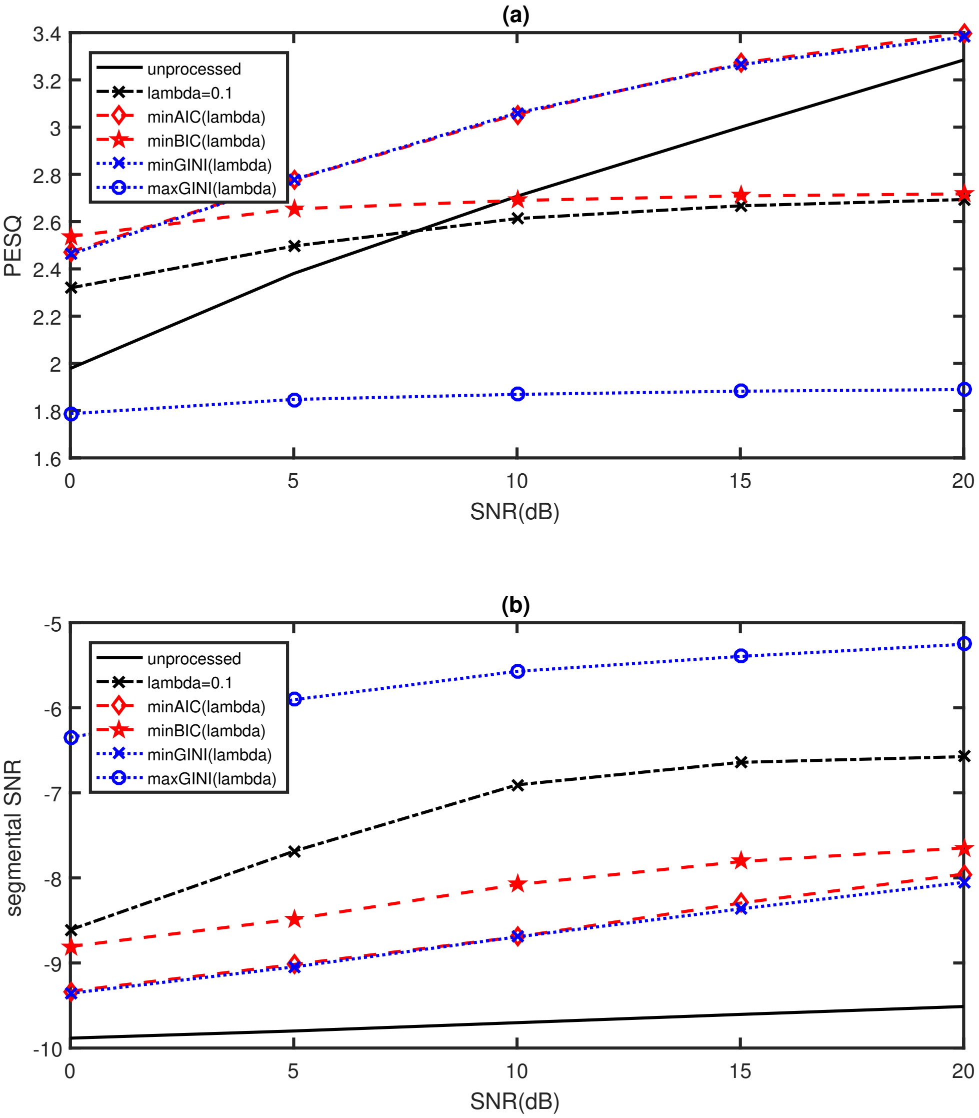

Figure 4.5: The (a) PESQ and (b) segmental SNR of the different hyperparameter optimization methods as a function of SNRs for destroyer noise.

Chapter 5

Two Models under Asymmetric Laplace Distributions

In this chapter, we focus on two models under Asymmetric Laplace Distributions. These two models are: mixture linear regression model and robust portfolio selection model.

5.1 Mixture linear regression under ALDs

The Quantile Regression estimator is equivalent to maximizing the likelihood function of a linear regression model with random errors following the ALD (see [129]):

$$f(t;\mu,\sigma,\tau)=\frac{\tau(1-\tau)}{\sigma}\exp\Big\{-\rho_\tau\Big(\frac{t-\mu}{\sigma}\Big)\Big\}, \tag{5.1}$$

where $\rho_\tau(t)=t(\tau-I(t<0))$ is the check function with $I(\cdot)$ as the indicator function, $-\infty<\mu<\infty$ is the location parameter, $\sigma>0$ is the scalar parameter, $0<\tau<1$ is the asymmetric (skewness) parameter. Hereafter, we refer to this distribution as ALD (μ,σ,τ). From Eqn. (5.1), it is easy to calculate that its cumulative distribution function (CDF) is

$$F(t;\mu,\sigma,\tau)=\begin{cases}\tau\exp\Big\{\rho_\tau\Big(\frac{t-\mu}{\sigma}\Big)\Big\}, & t<\mu,\\ 1-(1-\tau)\exp\Big\{\rho_\tau\Big(\frac{t-\mu}{\sigma}\Big)\Big\}, & t\geqslant\mu.\end{cases} \tag{5.2}$$

Obviously, the τth quantile of ALD (μ, σ, τ) is μ, and ALD (μ, σ, τ) reduces to the standard Laplace distribution or double-exponential distribution when $\tau = 1/2$. This important property of ALD (μ, σ, τ) makes it more popular than ALDs, as it can be generally applied to quantile regression. Another property of ALD (μ, σ, τ) is that the ALD (μ, σ, τ) can be represented as a normal-variance-mean mixture with an exponential mixing distribution as follows [72].

Lemma 1. If a random variable X follows the ALD (μ, σ, τ), then it holds that

$$X|Z \sim N(\mu + \kappa Z, \nu^2 \sigma Z), \quad \text{and} \quad Z \sim \operatorname{Exp}(\sigma^{-1}), \tag{5.3}$$

where $\kappa = \frac{1-2\tau}{\tau(1-\tau)}$ and $\nu^2 = \frac{2}{\tau(1-\tau)}$, $Exp(\sigma^{-1})$ is the exponential distribution with mean σ.

Remark 1. Random numbers from ALD $(0, 1, \tau)$ can be generated via the simple linear combination $\frac{U_1}{\tau} - \frac{U_2}{1-\tau}$ of two independent exponential random variables U_1 and U_2 each with mean 1 [128]. By location-scale transformation, we can generate random variables from ALD (μ, σ, τ). The expectation and variance of X is $\mathrm{E}(X) = \mu + \kappa\sigma$, and $\mathrm{Var}(X) = \psi^2\sigma^2$ with $\psi^2 = \kappa^2 + \nu^2$.

5.1.1 Methodology

Given the mixture structure and the objective function in (1.4), the special link between quantile regression and Asymmetric Laplace distribution motivate us to link the error distribution with mixture ALDs. Thus we seek to conduct a regression with linear regression based on mixture Laplace distribution, and advocate EM algorithm for solutions.

The model

For linear regression model with mixture Laplace error, we assume that for each component k, $k = 1, \cdots, K$, ε_k follows an Asymmetric Laplace distribution with location 0, scale ψ^{-1}, which results in the variance of ε_k being 1, and the asymmetric parameter τ, i.e., $\varepsilon_k \sim \mathrm{ALD}\,(0, \psi^{-1}, \tau)$. With sample observation (X_i, Y_i), our model becomes

$$Y_i = X_i^{\intercal}\boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \sum_{k=1}^{K} \pi_k \mathrm{ALD}(\mu, \sigma_k/\phi, \tau), \tag{5.4}$$

where $\pi_k > 0, \sum_{k=1}^{K} \pi_k = 1$, the scale ϕ satisfy $\phi^2 = \kappa^2 + \nu^2$, with $\kappa = \frac{1-2\tau}{\tau(1-\tau)}, \nu^2 = \frac{2}{\tau(1-\tau)}$. When $k = 1$, this reduces to the usual quantile regression.

Under model (5.4), the conditional distribution of $Y_i|X_i$ can be written as

$$Y_i|X_i \sim \sum_{k=1}^{K} \pi_k f(Y_i - X_i^{\intercal}\boldsymbol{\beta}, 0, \sigma_k/\phi, \tau),$$

where $f(Y_i - X_i^{\intercal}\boldsymbol{\beta}, 0, \sigma_k/\phi, \tau)$ is the density function of $\mathrm{ALD}\,(x; \mu, \sigma, \tau)$ evaluated at $Y_i - X_i^{\intercal}\boldsymbol{\beta}$.

Then it is easily seen that for a sample $\mathcal{O} = \{X_i, Y_i\}_{i=1}^{n}$ form model (1.3), the log-likelihood function of $\theta = (\pi^{\intercal}, \boldsymbol{\beta}^{\intercal}, \sigma^{\intercal})^{\intercal}$ with $\pi = (\pi_1, \cdots, \pi_K)^{\intercal}$, $\boldsymbol{\beta} = (\beta_1^{\intercal}, \cdots, \beta_K^{\intercal})^{\intercal}$, and $\sigma = (\sigma_1, \cdots, \sigma_K)^{\intercal}$, can be written as

$$l_{obs}(\theta; \mathcal{O}) = \sum_{i=1}^{n} \log \Big[\sum_{k=1}^{K} \pi_k \frac{\psi\tau(1-\tau)}{\sigma_k} \exp\Big\{ -\rho_\tau\Big(\frac{\psi(Y_i - X_i^{\intercal}\beta_k)}{\sigma_k}\Big)\Big\}\Big]. \tag{5.5}$$

Usually no explicit MLE is available. In the following, two missing component will be incorporated into the log-likelihood function (5.5), so that the maximizer can be obtained via a standard use of EM algorithm.

We try to estimate parameter $\beta_k, \pi_k, k \in \{1, 2, \cdots, K\}$ using EM algorithm, with two level latent variables taken into consideration.

First level latent variable

Denote the unobservable information G_{ik} with

$$G_{ik} = \begin{cases} 1, & \text{if i}-th \text{ observation is generated from the k}-th \text{ component} \\ 0, & \text{otherwise} \end{cases}$$

where $i = 1, \cdots, n$ and $k = 1, \cdots, K$, and denote the K-dimensional vector $G_i = (G_{i1}, \cdots, G_{iK})^\intercal$ as the component of origin of (Y_i, X_i), respectively. Then the complete log-likelihood function $l_c(\theta; \mathcal{O}, G)$ of model (1.3) can be easily obtained as

$$\begin{aligned} l_c(\theta; \mathcal{O}, G) &= \sum_{i=1}^{n} \sum_{k=1}^{K} G_{ik} \log \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} G_{ik} \log f(Y_i - X_i^\intercal \beta_k; 0, \sigma_k/\phi, \tau) \\ &:= \sum_{i=1}^{n} \sum_{k=1}^{K} G_{ik} \log \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} G_{ik} \log \Big[\frac{\psi\tau(1-\tau)}{\sigma_k} \exp\Big\{ -\rho_\tau \Big(\frac{\psi(Y_i - X_i^\intercal \beta_k)}{\sigma_k} \Big) \Big\} \Big] \\ &:= l_{c_1}(\pi; G) + l_{c_2}(\beta, \sigma; \mathcal{O}, G), \end{aligned} \tag{5.6}$$

where $G = (G_1, \cdots, G_n)$ are the first missing variables.

Second level latent variable

According to the normal-variance-mean mixture representation of ALD given in Eqn. (5.3), denote z_i, coupled with (Y_i, X_i), as the second latent scalar variable, $i = 1, \cdots, n$, then the complete log-likelihood function of θ, based on $\mathcal{D} = \{\mathcal{O}, G, z\}$

with $z = (z_i, i = 1, \cdots, n)$ has the form

$$l_c(\theta; \mathcal{D}) = \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\Big[\log \pi_k + \tag{5.7}$$

$$\log\Big\{(2\pi\nu^2 z_i \sigma_k^2 \psi^{-1})^{-1/2} \exp\Big\{-\frac{(Y_i - X_i^\intercal \beta_k - \kappa\sigma_k z_i)^2}{2\nu^2 z_i \sigma_k^2 \psi^{-1}}\Big\}\psi \exp\Big(-\psi z_i\Big)\Big\}\Big]$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\log \pi_k - \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\frac{\psi(Y_i - X_i^\intercal \beta_k - \kappa\sigma_k z_i)^2}{2\nu^2 z_i \sigma_k^2} - \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\log \sigma_k$$

$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\log\Big(2\pi\nu^2 z_i\Big) - \psi\sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik} z_i + \frac{3n}{2}\log\psi$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\log \pi_k - \psi\sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\frac{(Y_i - X_i^\intercal \beta_k)^2}{2\nu^2\sigma_k^2} z_i^{-1} + \psi\sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\frac{\kappa}{\nu^2\sigma_k}\Big(Y_i - X_i^\intercal \beta_k\Big)$$

$$-\psi\sum_{i=1}^{n}\frac{\kappa^2}{2\nu^2} z_i - \frac{1}{2}\sum_{i=1}^{n}\log\Big(2\pi\nu^2 z_i\Big) - \psi\sum_{i=1}^{n} z_i + \frac{3n}{2}\log\psi \tag{5.8}$$

$$\propto \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\log \pi_k - \psi\sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\frac{(Y_i - X_i^\intercal \beta_k)^2}{2\nu^2\sigma_k^2} z_i^{-1} +$$

$$\psi\sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\frac{\kappa}{\nu^2\sigma_k}\Big(Y_i - X_i^\intercal \beta_k\Big) \tag{5.9}$$

$$:= l_{c_1}(\pi; G) + l_{c_2}(\beta, \sigma; \mathcal{O}, z). \tag{5.10}$$

By noticing that the last three terms (5.8) do not involve the unknown parameters, we can simply drop them from the analysis, and obtain Eq. (5.9). And notations in

Eqn. (5.10) are

$$l_{c_1}(\pi; G) = \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik} \log \pi_k, \tag{5.11}$$

$$l_{c_2}(\beta, \sigma; \mathcal{O}, z) = -\psi \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik} \frac{(Y_i - X_i^{\intercal}\beta_k)^2}{2\nu^2\sigma_k^2} z_i^{-1} + \psi \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik} \frac{\kappa}{\nu^2\sigma_k}\Big(Y_i - X_i^{\intercal}\beta_k\Big). \tag{5.12}$$

Remark 2. Eqn. (5.9) can be obtained as follows According to the normal-variance-mean mixture representation of ALD given in Eqn. (5.3), model (1.3) can be rewritten as

$$Y_i = X_i^{\intercal}\beta_k + \sigma_k(\kappa z_i + \nu\psi^{-1/2}u_i), \tag{5.13}$$

where z_is be the latent variable with independently and identically distributed Exp (ψ), and u_is are independently distributed as $N(0, z_i)$ given z_is. Thus, Given $G_{ik} = 1$, the i-th complete log-likelihood function based on $(\mathcal{O}_i, z_i)$ is

$$\log\Big[(2\pi\nu^2\sigma_k^2\psi^{-1}z_i)^{-1/2} \exp\Big\{-\frac{(Y_i - X_i^{\intercal}\beta_k - \kappa\sigma_k z_i)^2}{2\nu^2\sigma_k^2\psi^{-1}z_i}\Big\}\psi \exp\Big(-\psi z_i\Big)\Big],$$

thus the complete log-likelihood function of θ, based on $\mathcal{D} = \{\mathcal{O}, G, z\}$ with $z = (z_i, i = 1, \cdots, n)$ has the form as $l_c(\theta; \mathcal{D})$ in Eqn. (5.9) after omitting terms that are not dependent on θ.

E-Step

Based on the EM algorithm principle, in the E-step, we have to calculate the conditional expectation $E(l_c(\theta, b; \mathcal{D})|\mathcal{O}, \theta^{(m)})$. Since the last three terms (5.8) do not involve the unknown parameters, we can simply drop them from the analysis, and obtain the conditional expectation for Eqn. (5.9) under the observation $\mathcal{O}$ and the current estimate $\theta^{(m)}$. Thus, we only have to calculate the following two terms

$$\delta_{ik}^{(m)} = E\Big(G_{ik}|\mathcal{O};\theta^{(m)}\Big) = \frac{\pi_k^{(m)} f(Y_i - X_i^\intercal \beta_k^{(m)}; 0, \sigma_k^{(m)}/\psi, \tau)}{\sum_{k=1}^K \pi_k^{(m)} f(Y_i - X_i^\intercal \beta_k^{(m)}; 0, \sigma_k^{(m)}/\psi, \tau)}, \tag{5.14}$$

$$\omega_{ik}^{(m)} = E\Big(z_i^{-1}|\mathcal{O}, G_{ik} = 1;\theta^{(m)}\Big) = \frac{\sigma_k^{(m)}}{\tau(1-\tau)|Y_i - X_i^\intercal \beta_k^{(m)}|}, \tag{5.15}$$

where $f(Y_i - X_i^\intercal \beta_k^{(m)}; 0, \sigma_k^{(m)}/\psi, \tau)$ is the pdf of the distribution ALD $(0, \sigma_k^{(m)}/\psi, \tau)$ evaluated at $Y_i - X_i^\intercal \beta_k^{(m)}$ as in Eqn. (5.1). For expectation in Eqn. (5.15), see detail in the end of this section. With these expectations in Eqn.(5.14), (5.15), it follows that

$$\tilde{l}_{c_1}(\pi) = \sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \log \pi_k, \tag{5.16}$$

$$\begin{aligned}\tilde{l}_{c_2}(\beta, \sigma) &= -\psi \sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \omega_{ik}^{(m)} \frac{(Y_i - X_i^\intercal \beta_k)^2}{2\nu^2 \sigma_k^2} + \psi \sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \frac{\kappa}{\nu^2 \sigma_k}\Big(Y_i - X_i^\intercal \beta_k\Big) \\ &= \frac{\psi}{2}\Big[-\sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \breve{\omega}_{ik}^{(m)} \frac{(Y_i - X_i^\intercal \beta_k)^2}{2\sigma_k^2} + \sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \frac{1-2\tau}{\sigma_k}\Big(Y_i - X_i^\intercal \beta_k\Big)\Big], \end{aligned} \tag{5.17}$$

where the second equality in Eqn. (5.17) holds due to the fact that ν^2, κ and ω_{ik} all contains the same term $\tau(1-\tau)$, and define $\breve{\omega}_{ik}^{(m)} = \frac{2\omega_{ik}}{\nu^2} = \sigma^{(m)}|Y_i - X_i^\intercal \beta_k^{(m)}|^{-1}$.

M-Step

At the M-step, $\pi_k^{(m+1)} = \frac{1}{n}\sum_{i=1}^n \delta_{ik}^{(m)}$. And update $\beta_k^{(m+1)}$ and $\sigma_k^{(m+1)}$ via maximizing the following equations

$$-\sum_{i=1}^n \delta_{ik}^{(m)} \breve{\omega}_{ik}^{(m)} \frac{(Y_i - X_i^\intercal \beta_k)^2}{2\sigma_k^2} + \sum_{i=1}^n \delta_{ik}^{(m)} \frac{1-2\tau}{\sigma_k}\Big(Y_i - X_i^\intercal \beta_k\Big). \tag{5.18}$$

With σ_k fixed at $\sigma_k^{(m)}$ in Eqn. (5.18), and on differentiation with respect to β_k, it holds that

$$\sum_{i=1}^{n} \delta_{ik}^{(m)} \tilde{\omega}_{ik} X_i (Y_i - X_i^{\intercal} \beta_k) - \sum_{i=1}^{n} \delta_{ik}^{(m)} (1 - 2\tau) X_i = 0,$$

where $\tilde{\omega}_{ik}^{(m)} = |Y_i - X_i^{\intercal} \beta_k^{(m)}|^{-1}$, and the updating formulae for β_k is

$$\beta_k^{(m+1)} = \Big(\sum_{i=1}^{n} \delta_{ik}^{(m)} \tilde{\omega}_{ik}^{(m)} X_i X_i^{\intercal} \Big)^{-1} \Big\{ \sum_{i=1}^{n} \delta_{ik}^{(m)} \Big(\tilde{\omega}_{ik}^{(m)} Y_i - (1 - 2\tau) \Big) X_i \Big\}, \ k = 1, \cdots, K. \tag{5.19}$$

Denote $\mathbf{Y} = (Y_1, \cdots, Y_n)^{\intercal}$, $\mathbf{X} = (X_1, \cdots, X_n)^{\intercal}$, $\mathbf{W}_k^{(m)} = \operatorname{diag}\{\delta_{1k}^{(m)} \tilde{\omega}_{1k}^{(m)}, \cdots, \delta_{nk}^{(m)} \tilde{\omega}_{nk}^{(m)}\}$, and $\Delta_k^{(m)} = (\delta_{1k}^{(m)}, \cdots, \delta_{nk}^{(m)})^{\intercal}$, then the updating formulae for β_k in Eqn.(5.19) can be rewritten as

$$\beta_k^{(m+1)} = \Big(\mathbf{X}^{\intercal} \mathbf{W}_k^{(m)} \mathbf{X} \Big)^{-1} \mathbf{X}^{\intercal} \Big(\mathbf{W}_k^{(m)} \mathbf{Y} - (1 - 2\tau) \Delta_k^{(m)} \Big). \tag{5.20}$$

The estimation of $\boldsymbol{\beta}^{(m+1)}$ can be viewed as reweighted least squared procedure, as shown in Schlossmacher [98] for one group situation.

After obtaining the updated estimates $\beta_k^{(m+1)}, k = 1, \cdots, K$, we can update $\sigma_k^{(m+1)}$ as follows. For the second term in Eqn. (5.6), take expectation with respect to G_{ik} based on $\mathcal{O}$ and the current estimate, it follows that

$$\tilde{l}_{c_2}(\beta, \sigma) = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{ik}^{(m)} \log \Big[\frac{\psi \tau (1 - \tau)}{\sigma_k} \exp \Big\{ - \rho_\tau \Big(\frac{\psi (Y_i - X_i^{\intercal} \beta_k)}{\sigma_k} \Big) \Big\} \Big]. \tag{5.21}$$

With β_k fixed at $\beta^{(m+1)}$ in Eqn. (5.21), and on differentiation $\tilde{l}_{c_2}(\boldsymbol{\beta}^{(m+1)}, \sigma)$ with respect to σ_k, it follows that

$$\sigma_k^{(m+1)} = \frac{\psi \sum_{i=1}^{n} \delta_{ik}^{(m)} \rho_\tau \Big(Y_i - X_i^{\intercal} \beta_k^{(m+1)} \Big)}{\sum_{i=1}^{n} \delta_{ik}^{(m)}}, \ k = 1, \cdots, K. \tag{5.22}$$

If we further assume that all σ_k's are equal, i.e., σ, then a common updated value for σ should be used in Eqn. (5.22) as follows

$$\sigma^{(m+1)} = \frac{\psi \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{ik}^{(m)} \rho_\tau\Big(Y_i - X_i^\intercal \beta_k^{(m+1)}\Big)}{n}. \tag{5.23}$$

These updated estimate $\theta^{(m+1)}$ can be substituted into Eqns. (5.14) and (5.15) for the implementation of the next E-step, until convergence is obtained.

Remark 3. We can explain the updated estimates for $\boldsymbol{\beta}^{(m+1)}$ in Eqn. (5.19) and $\sigma_k^{(m+1)}$ in Eqn. (5.22) from another viewpoint as follows. Given $G_{ik} = 1$, model (1.3) can be rewritten as

$$Y_i = X_i^\intercal \beta_k + \kappa z_{ik} + \nu \sigma_k^{1/2} \psi^{-1/2} u_{ik}, \tag{5.24}$$

where z_{ik}s be the latent variable with independently and identically distributed $\text{Exp}(\psi/\sigma_k)$, and u_{ik}s are independently distributed as $N(0, z_{ik})$ given z_{ik}s. Thus, Given $G_{ik} = 1$, the ith complete log-likelihood function is

$$\begin{aligned} & l_{cik}(\beta_k, \sigma_k; \mathcal{O}_i, z_{ik}, G_{ik} = 1) \\ & = \log\Big[(2\pi\nu^2\sigma_k\psi^{-1}z_{ik})^{-1/2} \exp\Big\{-\frac{(Y_i - X_i^\intercal\beta_k - \kappa z_{ik})^2}{2\nu^2\sigma_k\psi^{-1}z_{ik}}\Big\} \frac{\psi}{\sigma_k} \exp\Big(-\frac{\psi z_{ik}}{\sigma_k}\Big)\Big], \end{aligned}$$

thus after omitting terms that are not dependent on θ, the complete log-likelihood function of θ, based on $\mathcal{D} = \{\mathcal{O}, G, \tilde{z}\}$ with $\tilde{z} = (z_{ik}, i = 1, \cdots, n, k = 1, \cdots, K)$ has the form as

$$\begin{aligned} l_c(\theta; \mathcal{D}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik} \log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik} l_{cik}(\beta_k, \sigma_k; \mathcal{O}_i, z_{ik}, G_{ik} = 1) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik} \log \pi_k - \frac{3}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik} \log \sigma_k - \psi \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik} \frac{(Y_i - X_i^\intercal\beta_k)^2}{2\nu^2\sigma_k} z_{ik}^{-1} \\ & \quad +\psi \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik} \frac{\kappa}{\nu^2\sigma_k}(Y_i - X_i^\intercal\beta_k) - \psi \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik} \frac{\kappa^2}{2\nu^2\sigma_k} z_{ik} - \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik} \frac{\psi z_{ik}}{\sigma_k} \\ &:= l_{c_1}(\pi; G) + l_{c_2}(\beta, \sigma; \mathcal{D}) + l_{c_3}(\sigma; G, \tilde{z}), \end{aligned} \tag{5.25}$$

where

$$l_{c_1}(\pi; G) = \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\log \pi_k, \tag{5.26}$$

$$l_{c_2}(\beta, b, \sigma; \mathcal{D}) = \psi\Big[-\sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\frac{(Y_i - X_i^{\intercal}\beta_k)^2}{2\nu^2\sigma_k} z_{ik}^{-1} + \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\frac{\kappa}{\nu^2\sigma_k}(Y_i - X_i^{\intercal}\beta_k)\Big], \tag{5.27}$$

$$l_{c_3}(\sigma; G, z) = -\frac{3}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\log\sigma_k - \psi\sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\frac{\kappa^2}{2\nu^2\sigma_k}z_{ik} - \sum_{i=1}^{n}\sum_{k=1}^{K} G_{ik}\frac{\psi z_{ik}}{\sigma_k} \tag{5.28}$$

Based on the EM algorithm principle, in the E-step, we have to calculate the conditional expectation $E(l_c(\theta; \mathcal{D})|\mathcal{O}, \theta^{(m)})$. Since the third term $l_{c_3}(\sigma; G, \tilde{z})$ in Eqn. (5.25) do not involve the unknown regression parameters $\beta_k, k = 1, \cdots, K$, we can simply drop them from the following analysis. Thus, to find $E(l_c(\theta; \mathcal{D})|\mathcal{O}, \theta^{(m)})$, we only have to calculate the following two terms

$$\delta_{ik}^{(m)} = E\Big(G_{ik}|\mathcal{O}; \theta^{(m)}\Big) = \frac{\pi_k^{(m)} f(Y_i - X_i^{\intercal}\beta_k^{(m)}; 0, \sigma_k^{(m)}/\psi, \tau)}{\sum_{k=1}^{K}\pi_k^{(m)} f(Y_i - X_i^{\intercal}\beta_k^{(m)}; 0, \sigma_k^{(m)}/\psi, \tau)}, \tag{5.29}$$

$$\breve{\omega}_{ik}^{(m)} = E\Big(z_{ik}^{-1}|\mathcal{O}, G_{ik} = 1; \theta^{(m)}\Big) = \frac{1}{\tau(1-\tau)|Y_i - X_i^{\intercal}\beta_k^{(m)}|}, \tag{5.30}$$

where $f(Y_i - X_i^{\intercal}\beta_k^{(m)}; 0, \sigma_k^{(m)}/\psi, \tau)$ is the pdf of the distribution ALD $(0, \sigma_k^{(m)}/\psi, \tau)$ evaluated at $Y_i - X_i^{\intercal}\beta_k^{(m)}$ as in Eqn. (5.1). Note that the conditional expectation $\breve{\omega}_{ik}^{(m)}$ can be obtained similar to the calculation of ω_{ik} in Eqn. (5.15). With these expectations in Eqn.(5.29) and (5.30), it follows that

$$\breve{l}_{c_1}(\pi) = \sum_{i=1}^{n}\sum_{k=1}^{K} \delta_{ik}^{(m)} \log \pi_k, \tag{5.31}$$

$$\begin{aligned}\breve{l}_{c_2}(\beta,\sigma) &= \psi\Big[-\sum_{i=1}^{n}\sum_{k=1}^{K} \delta_{ik}^{(m)}\breve{\omega}_{ik}^{(m)} \frac{(Y_i - X_i^\intercal\beta_k)^2}{2\nu^2\sigma_k} + \sum_{i=1}^{n}\sum_{k=1}^{K} \delta_{ik}^{(m)} \frac{\kappa}{\nu^2}\Big(Y_i - X_i^\intercal\beta_k\Big)\Big] \\ &= \frac{\psi}{2}\Big[-\sum_{i=1}^{n}\sum_{k=1}^{K} \delta_{ik}^{(m)}\tilde{\omega}_{ik}^{(m)} \frac{(Y_i - X_i^\intercal\beta_k)^2}{2\sigma_k} + \sum_{i=1}^{n}\sum_{k=1}^{K} \delta_{ik}^{(m)}(1-2\tau)\Big(Y_i - X_i^\intercal\beta_k\Big)\Big]\end{aligned} \tag{5.32}$$

At the M-step, $\pi_k^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n}\delta_{ik}^{(m)}$. By solving

$$\frac{\partial \breve{l}_{c_2}(\beta,\sigma)}{\partial \beta_k} = \frac{\psi}{2\sigma_k}\sum_{i=1}^{n}\delta_{ik}^{(m)} X_i\Big\{\tilde{\omega}_{ik}^{(m)}\Big(Y_i - X_i^\intercal\beta_k\Big) - (1-2\tau)\Big\} = 0, \quad k = 1,\cdots,K,$$

we obtain the following updating formulae for $\beta_k, k = 1,\cdots,K$ as

$$\beta_k^{(m+1)} = \Big(\sum_{i=1}^{n}\delta_{ik}^{(m)}\tilde{\omega}_{ik}^{(m)} X_i X_i^\intercal\Big)^{-1}\Big\{\sum_{i=1}^{n}\delta_{ik}^{(m)}\Big(\tilde{\omega}_{ik}^{(m)} Y_i - (1-2\tau)\Big)X_i\Big\}, \tag{5.33}$$

which coincides with Eqn. (5.19) with σ fixed at $\sigma_k^{(m)}$ in Eqn. (5.18). From Eqn. (5.33), we can see that the updating formulae of β_k is independent of the updating value of σ_k, thus, after the updating value of β_k is obtained, the updating value of σ_k can then be got by maximizing Eqn. (5.21) with β_k fixed at $\beta_k^{(m)}$, which produces

$$\sigma_k^{(m+1)} = \frac{\psi\sum_{i=1}^{n}\delta_{ik}^{(m)}\rho_\tau\Big(Y_i - X_i^\intercal\beta_k^{(m+1)}\Big)}{\sum_{i=1}^{n}\delta_{ik}^{(m)}}, \ k = 1,\cdots,K, \tag{5.34}$$

which coincides with Eqn. (5.22).

If we further assume that all σ_k's are equal, i.e., σ, then in the above EM algorithm, a common value for σ should be used, and it can be updated in the M-step via

$$\sigma^{(m+1)} = \frac{\psi\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{ik}^{(m)}\rho_\tau\Big(Y_i - X_i^\intercal\beta_k^{(m+1)}\Big)}{n}, \tag{5.35}$$

the robustness of the above EM procedure follows from the adoption of composite QR regression. It is also obvious from the formulas of the updated β_k in each iteration. Note that the factor $\tilde{\omega}_{ik}^{(m)}$ is reciprocally related to the term $|Y_i - X_i^\intercal \beta_k^{(m)}|$, implying that the larger residuals gives smaller values of $\tilde{\omega}_{ik}^{(m)}$, and hence impose less weight of the corresponding observations on the updating estimates. Moreover, the above EM algorithm for updating β_k is an iterated re-weighted least square (IRLS) procedure, as the one proposed in [98].

Extra attention should be paid when programming the above EM algorithm. On one hand, the regression quantile satisfies that $|Y_i - X_i^\intercal \beta_k^{(m)}|$ is equal to zero for a subset of observations [62, 63] if a perfect QR fits occurs, as a result, $\tilde{\omega}_{ik}^{(m)}$ will be very large, and numerical instability would occur. To overcome this problem, Similar to [92], one can apply the following modified weighting strategy: one can choose a small $\epsilon > 0$, and if $|Y_i - X_i^\intercal \beta_k^{(m)}| \geqslant \epsilon$ for all observations, set $\tilde{\omega}_{ik}^{(m)} = |Y_i - X_i^\intercal \beta_k^{(m)}|^{-1}$; otherwise, set $\tilde{\omega}_{ik}^{(m)} = 1$ for $|Y_i - X_i^\intercal \beta_k^{(m)}| < \epsilon$, and $\tilde{\omega}_{ik}^{(m)} = \frac{\epsilon}{|Y_i - X_i^\intercal \beta_k^{(m)}|}$ for all other cases. These adjusted weights are still consistent with the original ones in the sense that those cases with more smaller residuals should be weighted more heavily. Here, another adjusted weighting strategy is applied, similar as [133], and simplifies the above adjusted weights. For the pre-assigned $\epsilon > 0$, a rather small but not too small positive value, set $\tilde{\omega}_{ik}^{(m)} = \{|Y_i - X_i^\intercal \beta_k^{(m)}| + \epsilon\}^{-1}$, and we set $\epsilon = 10^{-6}$ in our simulation.

On another hand, numerical instability could also occur if the weights $\delta_{ik}^{(m)}$ are very small. A common way to deal with this issue is to impose a hard threshold on $\delta_{ik}^{(m)}$ in Eqn. (5.14). That is, For the pre-assigned $\tilde{\epsilon} > 0$, a rather small but not too small positive value, set $\tilde{\delta}_{ik}^{(m)} = \tilde{\epsilon}$ if $\delta_{ik}^{(m)} < \epsilon$, and $\tilde{\delta}_{ik}^{(m)} = \delta_{ik}^{(m)}$, otherwise. And replace $\delta_{ik}^{(m)}$ in Eqn. (5.14) with $\tilde{\delta}_{ik}^{(m)}$ for the iteration, which is similar as [121] and [102]. In our simulations, $\tilde{\epsilon} = 10^{-6}$ is adopted.

In the end, we simply show the calculation for $\omega_{ik}^{(m)}$ in Eqn. (5.15). In fact, it is easy to calculate from (5.13) that the conditional distribution of z_i is proportional to

$$z_i^{-1/2} \exp\Big[-\frac{1}{2}\Big\{ \varsigma_{ik}^2 z_i^{-1} + \gamma^2 z_i \Big\}\Big], \tag{5.36}$$

where $\varsigma_{ik}^2 = \psi(Y_i - X_i^\intercal \beta_k)^2/(\nu^2\sigma_k^2)$ and $\gamma^2 = \psi(2+\kappa^2/\nu^2)$. Note that Eqn. (5.36) is the kernel of a generalized inverse Gaussian (GIG) distribution, thus

$$\Big[z_i | \mathcal{O}_i, G_{ik} = 1, \beta_k, \sigma_k \Big] \sim \text{GIG}\Big(\frac{1}{2}, \varsigma_{ik}, \gamma\Big).$$

For the general GIG$(u; \upsilon, \varsigma, \gamma)$ with $u > 0, -\infty < \upsilon < \infty, \varsigma > 0$, and $\gamma > 0$, [59] showed that the moments around the original of the GIG $(u; \upsilon, \varsigma, \gamma)$ distribution are given by

$$E\Big(z^r\Big) = \Big(\frac{\varsigma}{\gamma}\Big)^r \frac{K_{\upsilon+r}(\varsigma\gamma)}{K_\upsilon(\varsigma\gamma)},$$

where $K_\upsilon(\cdot)$ is a modified Bassel function of the third kind (See detail in [59, 72]). For $\upsilon = 1/2$ in our setting, it holds that

$$E(z^{-1}) = \Big(\frac{\varsigma}{\gamma}\Big)^{-1} \frac{K_{-1/2}(\varsigma\gamma)}{K_{1/2}(\varsigma\gamma)} = \frac{\gamma}{\varsigma}, \tag{5.37}$$

where the last equality holds due to property of $K_\upsilon(\cdot)$, i.e., $K_\upsilon(\cdot) = K_{-\upsilon}(\cdot)$ (see [1]). Substitute $\kappa = \frac{1-2\tau}{\tau(1-\tau)}, \nu^2 = \frac{2}{\tau(1-\tau)}$, $\varsigma_{ik} = \psi^{1/2}|Y_i - X_i^\intercal \beta_k|/(\nu\sigma_k)$ and $\gamma = \psi^{1/2}(2 + \kappa^2/\nu^2)^{1/2}$ into Eqn. (5.37), it follows that

$$E\Big(z_i^{-1} | \mathcal{O}, G_{ik} = 1; \theta^{(m)}\Big) = \frac{\sigma_k^{(m)}}{\tau(1-\tau)|Y_i - X_i^\intercal \beta_k^{(m)}|},$$

as shown in Eqn. (5.15). In the same way, we can calculate the expectation $\breve{\omega}_{ik}^{(m)}$ in Eqn. (5.30). Readjust the iteration formula, we have the EM algorithm as follows:

Algorithm 4 EM Algorithm for Mixture Laplace Distribution

```
1. Choose an initial value for $\theta^{(0)} = (\pi^{(0)\intercal}, \beta^{(0)\intercal}, \sigma^{(0)})^\intercal$;

2. **E-Step:** at the $(m+1)$-th iteration, calculate $\delta_{ik}^{(m)}$ from Eqs. (5.14), and
       $\tilde{\delta}_{ik}^{(m)} = \max\{\delta_{ik}^{(m)}, 10^{-6}\};$
       $\tilde{\omega}_{ik}^{(m)} = \{|Y_i - X_i^\intercal \beta_k^{(m)}| + 10^{-6}\}^{-1}.$

3. **M-Step:** at the $(m+1)$-th iteration, using the following formulas to calculate
   the updated estimates of $\theta$. For $k = 1, \cdots, K$,
       $\pi_k^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\delta}_{ik}^{(m)},$
       $\beta_k^{(m+1)} = \Big(\sum_{i=1}^{n} \tilde{\delta}_{ik}^{(m)} \tilde{\omega}_{ik}^{(m)} X_i X_i^\intercal\Big)^{-1} \Big\{\sum_{i=1}^{n} \tilde{\delta}_{ik}^{(m)} \Big(\tilde{\omega}_{ik}^{(m)} Y_i - (1-2\tau)\Big) X_i\Big\},$
                        $= \Big(\mathbf{X}^\intercal \mathbf{W}_k^{(m)} \mathbf{X}\Big)^{-1} \mathbf{X}^\intercal \Big(\mathbf{W}_k^{(m)} \mathbf{Y} - (1-2\tau)\Delta_k\Big),$
   where $\mathbf{W}_k^{(m)} = \mathrm{diag}\{\tilde{\delta}_{1k}^{(m)} \tilde{\omega}_{1k}^{(m)}, \cdots, \tilde{\delta}_{nk}^{(m)} \tilde{\omega}_{nk}^{(m)}\}$, $\Delta_k^{(m)} = (\tilde{\delta}_{1k}^{(m)}, \cdots, \tilde{\delta}_{nk}^{(m)})^\intercal$, and
       $\begin{cases} \sigma_k^{(m+1)} = \frac{\psi \sum_{i=1}^{n} \tilde{\delta}_{ik}^{(m)} \rho_\tau \Big(Y_i - X_i^\intercal \beta_k^{(m+1)}\Big)}{\sum_{i=1}^{n} \tilde{\delta}_{ik}^{(m)}}., & \text{if } \sigma_k \text{ are unequal;} \\ \sigma^{(m+1)} = \frac{\psi \sum_{i=1}^{n} \sum_{k=1}^{K} \tilde{\delta}_{ik}^{(m)} \rho_\tau \Big(Y_i - X_i^\intercal \beta_k^{(m+1)}\Big)}{n}, & \text{if } \sigma_k \text{ are equal.} \end{cases}$

4. Repeat **E-Step** and **M-Step** until convergence is obtained.
```

5.1.2 Numerical experiments

In this simulation study, we carry out several numerical experiments to assess the estimation performance of the proposed approaches described above. Simulated data sample $(X_i, y_i)_{i=1}^n$ are generated from the following two-component mixture regression models with mixing proportion $\pi_1 = \pi_2 = 0.5$,

$$Y = \begin{cases} 0 + 2X_1 + 2X_2 + \varepsilon_1(\tau), & \text{if } G = 1; \\ 0 - 2X_1 - 2X_2 + \varepsilon_2(\tau), & \text{if } G = 2. \end{cases}$$

Here, G is the group indicator, the true values for the regression coefficients of two components $\beta_1 = (0, 2, 2)'$ and $\beta_2 = (0, -2, -2)'$. The predictors $X = (1, X_1, X_2)^\intercal$ with X_1 and X_2 being simulated independently from uniform distribution $U(0, 1)$, the noised level $s = 0.2, 0.4$ corresponds to SNR ratio as SNR=4:1 and SNR=2:1. We set the sample size $n = 200$ and 400, and for each sample size, we generate 500 data sets. Once the simulated data were generated, we fit the proposed model with $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$ for the QR methods. Here, we consider equal variance for these two components, and the random error $\varepsilon_1(\tau)$ and $\varepsilon_2(\tau)$ are independent and have the same distribution as $\varepsilon(\tau)$, where $\varepsilon(\tau) = \varepsilon - F^{-1}(\tau)$ with F being the common CDF of ε, thus $F^{-1}(\tau)$ is subtracted from ε to make the τ-th quantile of $\varepsilon(\tau)$ zero for identifiability purpose. Generally, the τ-th quantile for each case is fasten to zero. We consider five cases for generating ε:

Case 1 (Normal distribution). The error term $\varepsilon \sim N(0, 1)$;

Case 2 (Chisquare distribution with 2 degrees of freedom). The error term is chi-square distribution with two degrees of freedom;

Case 3 (T-distribution with 3 degrees of freedom). The error distribution is student t-distribution with three degrees of freedom;

Case 4 (Heteroscedastic Normal distribution). The error term $\varepsilon \sim (1+X)\, N(0, 1)$, $X \sim U(0, 1)$.

Case 5 (Asymmetric Laplace distribution). The error term $\varepsilon \sim ALD\,(0, 1/\phi, \tau)$.

We display the Bias (MSE) of each estimated parameters together with the total Bias (MSE) in Table 5.1-5.4.

We tuned the error term with τ-quantile quantity to guarantee zero location condition. EM algorithm based on mixture of Asymmetric Laplace Distribution is considered. In all the simulation studies, the iteration terminated when change in log likelihood is less than $1e^{-6}$, and the maximal iteration step is 10000. Of all the error cases, our proposed method performs well.

Simulation results are presented in Table 5.1-5.2. For estimation consistency, Table 5.1, 5.3 and Table 5.2, 5.4 show that the estimation error of $n = 400$ is overall smaller than $n = 200$ case; In terms of SNR, it is obvious that Table 5.2 ,5.4 perform worse than Table 5.1 ,5.3, respectively. Simulation results trade off between estimation efficiency and accuracy, and that our proposed method perform well for skewed error cases with SNR ratio less or equal to 2:1.

TRUE	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
Case I: $\varepsilon \sim N(0,1) - N(\tau, 0, 1)$					
β_{10}:0	0.0873(0.1067)	0.0695(0.0942)	0.0909(0.0291)	0.0291(0.6073)	0.0329(0.0407)
β_{11}:2	0.0369(0.0450)	0.0404(0.0320)	0.0320(0.2974)	0.1521(0.1536)	0.1299(0.1751)
β_{12}:2	0.1901(0.0357)	0.0357(1.0444)	0.1172(0.1199)	0.0981(0.1347)	0.1405(0.0306)
β_{20}:0	0.0306(0.8111)	0.0205(0.0172)	0.0198(0.0150)	0.0129(0.0308)	0.0308(0.1629)
β_{21}:-2	0.0117(0.0174)	0.0075(0.0145)	0.0128(0.0013)	0.0013(0.0823)	0.0016(0.0026)
β_{22}:-2	0.0026(0.0034)	0.0028(0.0016)	0.0016(0.0184)	0.0314(0.0361)	0.0257(0.0541)
pr	0.0523(0.0019)	0.0019(0.2512)	0.0198(0.0218)	0.0185(0.0287)	0.0323(0.0015)
Case II: $\varepsilon \sim \chi(2) - \chi(\tau, 2)$					
β_{10}:0	0.0643(0.0744)	0.0609(0.0677)	0.0828(0.0282)	0.0282(0.4808)	0.0583(0.0704)
β_{11}:2	0.0513(0.0595)	0.0713(0.0307)	0.0307(0.4405)	0.0698(0.0931)	0.0694(0.0955)
β_{12}:2	0.0915(0.0243)	0.0243(0.5661)	0.0987(0.1203)	0.0885(0.1135)	0.1115(0.0268)
β_{20}:0	0.0268(0.7025)	0.0152(0.0151)	0.0164(0.0145)	0.0161(0.0295)	0.0295(0.1514)
β_{21}:-2	0.0066(0.0088)	0.0060(0.0078)	0.0103(0.0012)	0.0012(0.0508)	0.0055(0.0085)
β_{22}:-2	0.0042(0.0056)	0.0075(0.0014)	0.0014(0.0414)	0.0077(0.0132)	0.0088(0.0148)
pr	0.0146(0.0010)	0.0010(0.0762)	0.0147(0.0210)	0.0115(0.0193)	0.0199(0.0011)
Case III: $\varepsilon \sim t(3) - t(\tau, 3)$					
β_{10}:0	0.0559(0.0681)	0.0566(0.0730)	0.0701(0.0245)	0.0245(0.4411)	0.0771(0.0975)
β_{11}:2	0.0885(0.1164)	0.1070(0.0317)	0.0317(0.6614)	0.0614(0.0707)	0.0708(0.0756)
β_{12}:2	0.0848(0.0279)	0.0279(0.5027)	0.0825(0.1113)	0.0925(0.0984)	0.1145(0.0298)
β_{20}:0	0.0298(0.6596)	0.0207(0.0232)	0.0172(0.0248)	0.0205(0.0316)	0.0316(0.1943)
β_{21}:-2	0.0052(0.0077)	0.0047(0.0082)	0.0070(0.0009)	0.0009(0.0419)	0.0093(0.0151)
β_{22}:-2	0.0119(0.0210)	0.0184(0.0015)	0.0015(0.0980)	0.0056(0.0081)	0.0082(0.0096)
pr	0.0124(0.0012)	0.0012(0.0574)	0.0106(0.0199)	0.0133(0.0154)	0.0208(0.0013)
Case IV: $\varepsilon \sim (1+X)(N(0,1) - N(\tau, 0, 1)), X \sim U(0,1)$					
β_{10}:0	0.0522(0.0680)	0.0544(0.0776)	0.0819(0.0260)	0.0260(0.4626)	0.1569(0.1789)
β_{11}:2	0.1973(0.2150)	0.2215(0.0472)	0.0472(1.2521)	0.0833(0.1010)	0.0831(0.1072)
β_{12}:2	0.1013(0.0289)	0.0289(0.6366)	0.0924(0.1273)	0.0895(0.1086)	0.1162(0.0296)
β_{20}:0	0.0296(0.7085)	0.0175(0.0184)	0.0152(0.0155)	0.0141(0.0255)	0.0255(0.1490)
β_{21}:-2	0.0044(0.0077)	0.0046(0.0101)	0.0102(0.0011)	0.0011(0.0479)	0.0378(0.0562)
β_{22}:-2	0.0535(0.0705)	0.0660(0.0034)	0.0034(0.3446)	0.0106(0.0164)	0.0114(0.0183)
pr	0.0170(0.0013)	0.0013(0.0935)	0.0135(0.0246)	0.0152(0.0197)	0.0227(0.0013)
Case V: $\varepsilon \sim ALD(0, 1/\phi, \tau)$					
β_{10}:0	0.0726(0.1054)	0.0892(0.1049)	0.1092(0.0281)	0.0281(0.6385)	0.2855(0.3602)
β_{11}:2	0.3200(0.3193)	0.3400(0.0599)	0.0599(2.0531)	0.1314(0.1495)	0.1459(0.1412)
β_{12}:2	0.1545(0.0401)	0.0401(0.9797)	0.1218(0.1395)	0.1318(0.1355)	0.1524(0.0326)
β_{20}:0	0.0326(0.8871)	0.0194(0.0166)	0.0197(0.0153)	0.0165(0.0301)	0.0301(0.1625)
β_{21}:-2	0.0079(0.0169)	0.0118(0.0157)	0.0177(0.0012)	0.0012(0.0892)	0.1211(0.2030)
β_{22}:-2	0.1746(0.1724)	0.2141(0.0046)	0.0046(1.0470)	0.0278(0.0358)	0.0337(0.0388)
pr	0.0377(0.0023)	0.0023(0.2312)	0.0260(0.0333)	0.0277(0.0305)	0.0384(0.0017)

Table 5.1: Simulation results of $n = 200, s = 0.2$, SNR=4:1

TRUE	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
Case I: $\varepsilon \sim N(0,1) - N(\tau,0,1)$					
β_{10}:0	0.1930(0.2269)	0.1443(0.1951)	0.1812(0.0356)	0.0356(1.2149)	0.0717(0.0845)
β_{11}:2	0.0564(0.0786)	0.0613(0.0320)	0.0320(0.4872)	0.2769(0.2970)	0.2600(0.2807)
β_{12}:2	0.3248(0.0673)	0.0673(1.8450)	0.3093(0.3085)	0.2000(0.2426)	0.2680(0.0493)
β_{20}:0	0.0493(1.7768)	0.0395(0.0360)	0.0431(0.0297)	0.0361(0.0284)	0.0284(0.2758)
β_{21}:-2	0.0587(0.0815)	0.0332(0.0605)	0.0528(0.0019)	0.0019(0.3551)	0.0078(0.0115)
β_{22}:-2	0.0055(0.0107)	0.0069(0.0016)	0.0016(0.0548)	0.1136(0.1289)	0.0976(0.1382)
pr	0.1584(0.0062)	0.0062(0.7794)	0.1486(0.1530)	0.0707(0.0977)	0.1147(0.0036)
Case II: $\varepsilon \sim \chi(2) - \chi(\tau,2)$					
β_{10}:0	0.1361(0.1586)	0.1256(0.1409)	0.1561(0.0295)	0.0295(0.9436)	0.1234(0.1563)
β_{11}:2	0.0966(0.1227)	0.1242(0.0322)	0.0322(0.8268)	0.1957(0.2181)	0.1680(0.2062)
β_{12}:2	0.2079(0.0473)	0.0473(1.2785)	0.2270(0.2190)	0.1610(0.2219)	0.1815(0.0390)
β_{20}:0	0.0390(1.3543)	0.0307(0.0359)	0.0297(0.0379)	0.0315(0.0289)	0.0289(0.2579)
β_{21}:-2	0.0303(0.0396)	0.0252(0.0322)	0.0382(0.0014)	0.0014(0.2073)	0.0220(0.0402)
β_{22}:-2	0.0146(0.0242)	0.0258(0.0016)	0.0016(0.1600)	0.0557(0.0750)	0.0461(0.0690)
pr	0.0650(0.0036)	0.0036(0.3774)	0.0730(0.0720)	0.0449(0.0740)	0.0641(0.0025)
Case III: $\varepsilon \sim t(3) - t(\tau,3)$					
β_{10}:0	0.1455(0.1664)	0.1290(0.1508)	0.1516(0.0300)	0.0300(0.9806)	0.1706(0.2156)
β_{11}:2	0.2193(0.2064)	0.2617(0.0516)	0.0516(1.3796)	0.1176(0.1631)	0.1359(0.1574)
β_{12}:2	0.1708(0.0282)	0.0282(0.9525)	0.1776(0.2022)	0.1518(0.2101)	0.1977(0.0298)
β_{20}:0	0.0298(1.1853)	0.0386(0.0489)	0.0389(0.0530)	0.0608(0.0267)	0.0267(0.3428)
β_{21}:-2	0.0330(0.0470)	0.0254(0.0353)	0.0341(0.0015)	0.0015(0.2273)	0.0427(0.0690)
β_{22}:-2	0.0714(0.0653)	0.1071(0.0039)	0.0039(0.4340)	0.0220(0.0454)	0.0309(0.0402)
pr	0.0463(0.0013)	0.0013(0.2230)	0.0506(0.0644)	0.0376(0.0685)	0.0578(0.0014)
Case IV: $\varepsilon \sim (1+X)(N(0,1) - N(\tau,0,1)), X \sim U(0,1)$					
β_{10}:0	0.1101(0.1382)	0.1239(0.1610)	0.1592(0.0314)	0.0314(0.8857)	0.2770(0.4008)
β_{11}:2	0.5102(0.3540)	0.4008(0.1205)	0.1205(2.5294)	0.1605(0.2055)	0.1873(0.1865)
β_{12}:2	0.2088(0.0415)	0.0415(1.2149)	0.2096(0.2584)	0.2279(0.2538)	0.2615(0.0356)
β_{20}:0	0.0356(1.5122)	0.0327(0.0334)	0.0313(0.0346)	0.0336(0.0317)	0.0317(0.2620)
β_{21}:-2	0.0184(0.0316)	0.0264(0.0415)	0.0441(0.0015)	0.0015(0.1936)	0.1232(0.2481)
β_{22}:-2	0.3480(0.1925)	0.2962(0.0162)	0.0162(1.4464)	0.0333(0.0614)	0.0511(0.0651)
pr	0.0654(0.0026)	0.0026(0.3290)	0.0690(0.1006)	0.0780(0.0980)	0.1119(0.0021)
Case V: $\varepsilon \sim ALD(0,1/\phi,\tau)$					
β_{10}:0	0.1428(0.2196)	0.1770(0.1921)	0.2107(0.0341)	0.0341(1.1728)	0.5189(0.6733)
β_{11}:2	0.6361(0.7884)	0.9530(0.1281)	0.1281(4.3967)	0.2525(0.3038)	0.2854(0.2830)
β_{12}:2	0.3197(0.0641)	0.0641(1.8996)	0.2166(0.2697)	0.2782(0.2921)	0.3246(0.0510)
β_{20}:0	0.0510(1.7365)	0.0406(0.0303)	0.0381(0.0296)	0.0286(0.0297)	0.0297(0.2610)
β_{21}:-2	0.0306(0.0694)	0.0465(0.0547)	0.0725(0.0018)	0.0018(0.3210)	0.4093(0.6971)
β_{22}:-2	0.5444(1.3672)	2.2585(0.0195)	0.0195(5.8243)	0.0996(0.1473)	0.1218(0.1294)
pr	0.1617(0.0054)	0.0054(0.8356)	0.0728(0.1185)	0.1146(0.1304)	0.1607(0.0038)

Table 5.2: Simulation results of $n = 200, s = 0.4$, SNR=2:1

TRUE	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
Case I: $\varepsilon \sim N(0,1) - N(\tau,0,1)$					
β_{10}:0	0.0643(0.0718)	0.0485(0.0621)	0.0660(0.0213)	0.0213(0.4274)	0.0239(0.0231)
β_{11}:2	0.0184(0.0273)	0.0267(0.0243)	0.0243(0.1965)	0.1300(0.1209)	0.0777(0.1076)
β_{12}:2	0.1023(0.0336)	0.0336(0.7014)	0.1001(0.1119)	0.0684(0.0990)	0.0942(0.0257)
β_{20}:0	0.0257(0.6331)	0.0169(0.0100)	0.0182(0.0108)	0.0115(0.0210)	0.0210(0.1214)
β_{21}:-2	0.0067(0.0084)	0.0037(0.0066)	0.0064(0.0007)	0.0007(0.0416)	0.0008(0.0009)
β_{22}:-2	0.0006(0.0011)	0.0011(0.0009)	0.0009(0.0076)	0.0234(0.0218)	0.0096(0.0199)
pr	0.0178(0.0016)	0.0016(0.1123)	0.0152(0.0193)	0.0070(0.0145)	0.0135(0.0009)
Case II: $\varepsilon \sim \chi(2) - \chi(\tau,2)$					
β_{10}:0	0.0423(0.0520)	0.0438(0.0554)	0.0588(0.0218)	0.0218(0.3449)	0.0363(0.0435)
β_{11}:2	0.0384(0.0469)	0.0421(0.0195)	0.0195(0.2927)	0.0584(0.0633)	0.0592(0.0736)
β_{12}:2	0.0725(0.0231)	0.0231(0.4392)	0.0638(0.0803)	0.0583(0.0700)	0.0852(0.0174)
β_{20}:0	0.0174(0.4724)	0.0126(0.0105)	0.0132(0.0132)	0.0134(0.0226)	0.0226(0.1187)
β_{21}:-2	0.0030(0.0044)	0.0028(0.0050)	0.0053(0.0007)	0.0007(0.0259)	0.0020(0.0029)
β_{22}:-2	0.0024(0.0035)	0.0030(0.0006)	0.0006(0.0184)	0.0053(0.0075)	0.0056(0.0079)
pr	0.0078(0.0009)	0.0009(0.0429)	0.0066(0.0100)	0.0056(0.0084)	0.0110(0.0005)
Case III: $\varepsilon \sim t(3) - t(\tau,3)$					
β_{10}:0	0.0410(0.0529)	0.0407(0.0482)	0.0544(0.0207)	0.0207(0.3302)	0.0699(0.0817)
β_{11}:2	0.0657(0.0868)	0.0737(0.0229)	0.0229(0.5029)	0.0434(0.0552)	0.0464(0.0568)
β_{12}:2	0.0599(0.0203)	0.0203(0.3554)	0.0584(0.0662)	0.0570(0.0748)	0.0762(0.0190)
β_{20}:0	0.0190(0.4514)	0.0126(0.0151)	0.0098(0.0153)	0.0124(0.0196)	0.0196(0.1213)
β_{21}:-2	0.0026(0.0044)	0.0028(0.0042)	0.0047(0.0007)	0.0007(0.0240)	0.0069(0.0097)
β_{22}:-2	0.0064(0.0117)	0.0094(0.0008)	0.0008(0.0563)	0.0029(0.0047)	0.0032(0.0048)
pr	0.0054(0.0006)	0.0006(0.0266)	0.0060(0.0071)	0.0050(0.0085)	0.0082(0.0005)
Case IV: $\varepsilon \sim (1+X)(N(0,1) - N(\tau,0,1)), X \sim U(0,1)$					
β_{10}:0	0.0452(0.0613)	0.0424(0.0567)	0.0590(0.0176)	0.0176(0.3543)	0.1123(0.1392)
β_{11}:2	0.1774(0.1617)	0.1432(0.0431)	0.0431(0.9505)	0.0506(0.0691)	0.0519(0.0617)
β_{12}:2	0.0611(0.0199)	0.0199(0.3945)	0.0623(0.0846)	0.0667(0.0876)	0.0672(0.0229)
β_{20}:0	0.0229(0.4918)	0.0130(0.0117)	0.0114(0.0114)	0.0116(0.0207)	0.0207(0.1111)
β_{21}:-2	0.0030(0.0055)	0.0029(0.0051)	0.0053(0.0005)	0.0005(0.0273)	0.0181(0.0311)
β_{22}:-2	0.0434(0.0389)	0.0341(0.0024)	0.0024(0.1965)	0.0039(0.0074)	0.0041(0.0059)
pr	0.0061(0.0006)	0.0006(0.0344)	0.0065(0.0110)	0.0067(0.0121)	0.0077(0.0008)
Case V: $\varepsilon \sim ALD(0,1/\phi,\tau)$					
β_{10}:0	0.0572(0.0786)	0.0557(0.0758)	0.0617(0.0207)	0.0207(0.4343)	0.1726(0.2244)
β_{11}:2	0.3488(0.1987)	0.2221(0.0810)	0.0810(1.5519)	0.0854(0.1040)	0.1274(0.1148)
β_{12}:2	0.1207(0.0342)	0.0342(0.7247)	0.0831(0.1170)	0.0915(0.0921)	0.1005(0.0240)
β_{20}:0	0.0240(0.6255)	0.0198(0.0113)	0.0184(0.0106)	0.0108(0.0204)	0.0204(0.1221)
β_{21}:-2	0.0050(0.0098)	0.0051(0.0089)	0.0056(0.0006)	0.0006(0.0421)	0.0485(0.0873)
β_{22}:-2	0.1589(0.0630)	0.0779(0.0076)	0.0076(0.5321)	0.0109(0.0175)	0.0228(0.0205)
pr	0.0226(0.0016)	0.0016(0.1158)	0.0102(0.0199)	0.0123(0.0138)	0.0155(0.0008)

Table 5.3: Simulation results of $n = 400, s = 0.2$, SNR=4:1

TRUE	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
Case I: $\varepsilon \sim N(0,1) - N(\tau,0,1)$					
β_{10}:0	0.1450(0.1498)	0.1140(0.1544)	0.1247(0.0299)	0.0299(0.8916)	0.0442(0.0499)
β_{11}:2	0.0411(0.0537)	0.0548(0.0214)	0.0214(0.3406)	0.2564(0.2018)	0.1933(0.2392)
β_{12}:2	0.2147(0.0805)	0.0805(1.4656)	0.2356(0.2511)	0.1750(0.2280)	0.1895(0.0568)
β_{20}:0	0.0568(1.4148)	0.0399(0.0244)	0.0411(0.0234)	0.0228(0.0176)	0.0176(0.2084)
β_{21}:-2	0.0334(0.0337)	0.0209(0.0379)	0.0264(0.0013)	0.0013(0.1901)	0.0030(0.0040)
β_{22}:-2	0.0024(0.0042)	0.0045(0.0007)	0.0007(0.0246)	0.0923(0.0616)	0.0648(0.0915)
pr	0.0812(0.0073)	0.0073(0.4755)	0.0894(0.0958)	0.0460(0.0813)	0.0597(0.0041)
Case II: $\varepsilon \sim \chi(2) - \chi(\tau,2)$					
β_{10}:0	0.1068(0.1211)	0.0726(0.1067)	0.1071(0.0226)	0.0226(0.6852)	0.0853(0.0974)
β_{11}:2	0.0759(0.0969)	0.0911(0.0268)	0.0268(0.5914)	0.1571(0.1441)	0.1142(0.1397)
β_{12}:2	0.1428(0.0393)	0.0393(0.9278)	0.1452(0.1709)	0.1263(0.1573)	0.1757(0.0320)
β_{20}:0	0.0320(1.0031)	0.0258(0.0210)	0.0243(0.0231)	0.0221(0.0206)	0.0206(0.1814)
β_{21}:-2	0.0166(0.0237)	0.0093(0.0179)	0.0189(0.0008)	0.0008(0.1126)	0.0108(0.0151)
β_{22}:-2	0.0088(0.0144)	0.0135(0.0010)	0.0010(0.0781)	0.0338(0.0326)	0.0221(0.0310)
pr	0.0317(0.0022)	0.0022(0.1911)	0.0340(0.0469)	0.0264(0.0370)	0.0507(0.0016)
Case III: $\varepsilon \sim t(3) - t(\tau,3)$					
β_{10}:0	0.0914(0.1025)	0.0843(0.1010)	0.1118(0.0193)	0.0193(0.6301)	0.1338(0.1682)
β_{11}:2	0.1627(0.1545)	0.1640(0.0519)	0.0519(1.0447)	0.0960(0.0921)	0.0877(0.1134)
β_{12}:2	0.1180(0.0250)	0.0250(0.6907)	0.1141(0.1379)	0.1324(0.1561)	0.1506(0.0202)
β_{20}:0	0.0202(0.8550)	0.0193(0.0314)	0.0228(0.0303)	0.0292(0.0184)	0.0184(0.1974)
β_{21}:-2	0.0127(0.0189)	0.0114(0.0156)	0.0207(0.0006)	0.0006(0.0960)	0.0266(0.0448)
β_{22}:-2	0.0360(0.0360)	0.0422(0.0033)	0.0033(0.2301)	0.0141(0.0132)	0.0122(0.0206)
pr	0.0210(0.0009)	0.0009(0.1086)	0.0204(0.0330)	0.0273(0.0354)	0.0364(0.0007)
Case IV: $\varepsilon \sim (1+X)(N(0,1) - N(\tau,0,1)), X \sim U(0,1)$					
β_{10}:0	0.0805(0.1102)	0.0948(0.1096)	0.1038(0.0249)	0.0249(0.6425)	0.2198(0.2437)
β_{11}:2	0.4583(0.2738)	0.2799(0.1308)	0.1308(2.0025)	0.1105(0.1427)	0.1502(0.1406)
β_{12}:2	0.1399(0.0374)	0.0374(0.8964)	0.1209(0.1651)	0.1795(0.1879)	0.1733(0.0274)
β_{20}:0	0.0274(1.0163)	0.0252(0.0215)	0.0257(0.0208)	0.0229(0.0192)	0.0192(0.1758)
β_{21}:-2	0.0100(0.0191)	0.0128(0.0186)	0.0169(0.0009)	0.0009(0.0935)	0.0737(0.0952)
β_{22}:-2	0.2769(0.1055)	0.1457(0.0183)	0.0183(0.8467)	0.0177(0.0314)	0.0331(0.0304)
pr	0.0308(0.0020)	0.0020(0.1752)	0.0238(0.0400)	0.0501(0.0525)	0.0479(0.0012)
Case V: $\varepsilon \sim ALD(0,1/\phi,\tau)$					
β_{10}:0	0.1108(0.1381)	0.1692(0.1728)	0.1635(0.0292)	0.0292(0.9299)	0.3488(0.4261)
β_{11}:2	0.5099(0.5776)	0.5798(0.1352)	0.1352(3.1584)	0.1646(0.2170)	0.2596(0.1986)
β_{12}:2	0.2246(0.0782)	0.0782(1.4152)	0.1448(0.2125)	0.2186(0.2081)	0.2421(0.0518)
β_{20}:0	0.0518(1.3349)	0.0354(0.0243)	0.0339(0.0222)	0.0216(0.0190)	0.0190(0.1998)
β_{21}:-2	0.0209(0.0305)	0.0481(0.0455)	0.0453(0.0013)	0.0013(0.2164)	0.1862(0.2941)
β_{22}:-2	0.3851(0.5004)	0.5422(0.0190)	0.0190(2.2423)	0.0431(0.0763)	0.0939(0.0686)
pr	0.0730(0.0069)	0.0069(0.4248)	0.0325(0.0684)	0.0706(0.0678)	0.0947(0.0035)

Table 5.4: Simulation results of $n = 400, s = 0.4$, SNR=2:1

Tone perception data

A typical example for mixture regression is the tune perception data collected by Cohen [23]. In the experiment, we implement the proposed method to tone perception data. The experiment record 150 trails with the same musician. The overtones were determined by a stretching ratio.

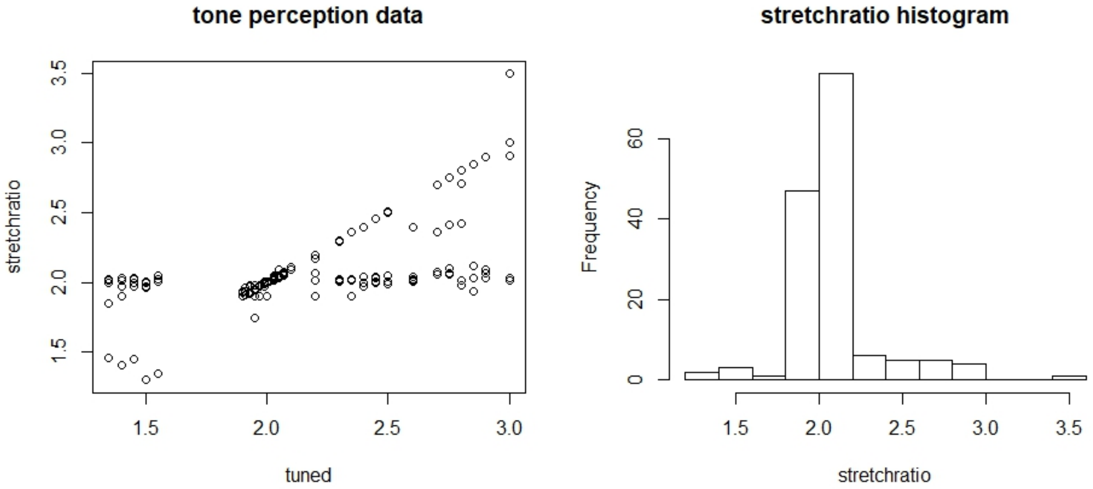

Figure 5.1: (a) scatter plot of tonedata; (b) histogram of perceived tune ratio.

The data is displayed in Figure 5.1. Figure 5.1 (a) indicate that the tone data can be modelled by two linear regression lines; Figure 5.1 (b) display clear non-normality and tail heaviness of the data, we fit the data according to Mixture of Normal distribution (MixN), Mixture of t distribution (MixT) and Mixture of Laplace distribution (MixLa), as shown in Figure 5.2. To better illustrate the robustness of the proposed estimation procedure, we conduct several outlier settings to evaluate the estimation performance.

We contaminate the datasets with extreme outlier cases: (a). 5 similar outliers $(1, 3)$; (b) 5 outliers $(3, 1)$; (c) 4 outliers (1,3), 4 outliers (3,1). As shown in Figure 5.2, we fit the data using Mixture normal distribution, Mixture t distribution and Mixture Laplace distribution with $\tau = 0.5$. Results show that for all these cases, MixN model perform worst, while MixT and MixLa are comparable in fitting performance. Figure 5.2 (a) show that when outliers deviate significantly from the population, MixLa method perform better as well; Figure 5.2 (b) show that when outliers deviate slightly from the population, MixLa and MixT all perform well whereas MixNormal fails; Figure 5.2 (c) indicate that with higher level outliers, MixN and MixT fail while MixLa still performs well.

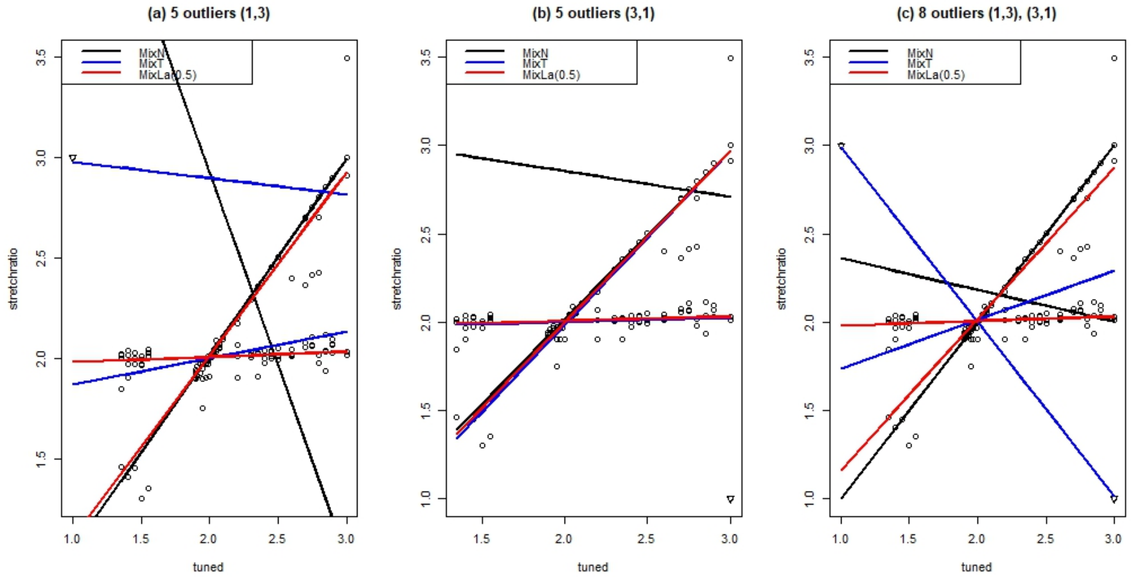

Figure 5.2: Outlier cases of tone perception data

5.2 Portfolio selection under ALD framework

Asymmetric Laplace Distribution

Kotz [68] proposed the Asymmetric Laplace Distribution with density function

$$f(\boldsymbol{x}) = \frac{2e^{\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}\big(\frac{\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}{2+\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}\big)^{v/2}K_v\big(\sqrt{(2+\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})(\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x})}\big), \tag{5.38}$$

denoted as $\boldsymbol{X} \sim AL_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here, n is the dimension of random vector $\boldsymbol{X}$, $v = (2-n)/2$ and $K_v(u)$ is the modified Bessel function of the third kind with the following two popular representations:

$$K_v(u) = \frac{1}{2}\big(\frac{u}{2}\big)^v \int_0^\infty t^{-v-1}\exp\Big\{-t-\frac{u^2}{4t}\Big\}dt, \quad u > 0, \tag{5.39}$$

$$K_v(u) = \frac{(u/2)^v\Gamma(1/2)}{\Gamma(v+1/2)}\int_1^\infty e^{-ut}(t^2-1)^{v-1/2}dt, \quad u > 0, v \geqslant -1/2. \tag{5.40}$$

When $\boldsymbol{\mu} = \boldsymbol{0}_n$, we can obtain Symmetric Laplace distribution $SL\,(\boldsymbol{\Sigma})$ with density

$$f(\boldsymbol{x}) = 2(2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\Big(\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}/2\Big)^{\nu/2}K_\nu\Big(\sqrt{2\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}\Big).$$

When $n = 1$, we have $\boldsymbol{\Sigma} = \sigma_{11} = \sigma$. In such cases, (5.38) becomes the univariate Laplace distribution $AL_1(\mu, \sigma)$ distribution with parameters μ and σ. The corresponding density function is

$$f(x) = \frac{1}{\gamma}\exp\Big\{-\frac{|x|}{\sigma^2}\big[\gamma - \mu\cdot\mathrm{sign}(x)\big]\Big\} \quad \text{with} \quad \gamma = \sqrt{\mu^2+2\sigma^2}. \tag{5.41}$$

The symmetric case ($\mu = 0$) leads to the univariate Laplace distribution $SL_1\,(0, \sigma)$.

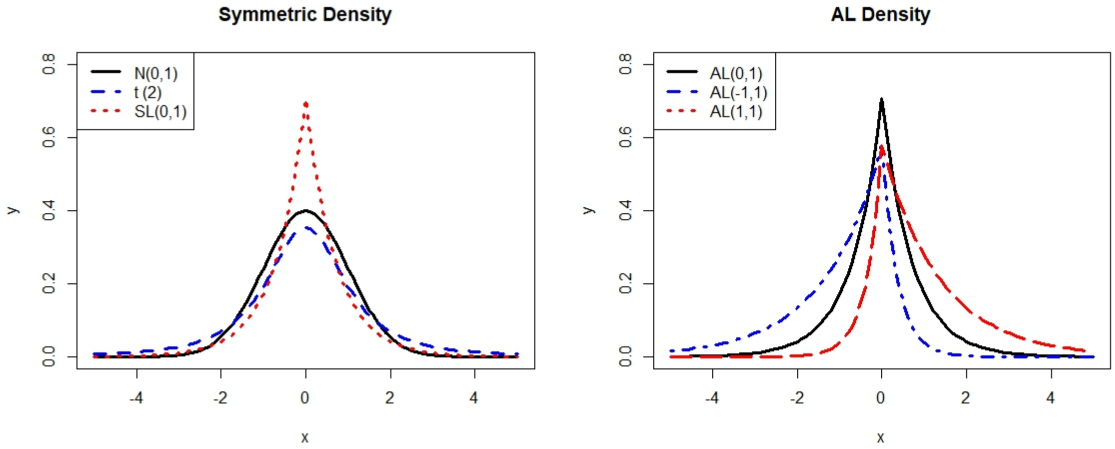

Figure 5.3: Univariate densities

Figure 5.3 displays plot of symmetric densities and AL densities. Symmetric densities including standard normal distribution, student t distribution with 2 degrees of freedom, and univariate symmetric Laplace distribution, denoted as $N\,(0,1), t\,(2)$, $SL_1\,(0,1)$. The student t distribution possesses heavier tail than normal distribution, whereas $SL_1\,(0,1)$ distribution imposes greater peakedness and heavier tail than normal case. As for plots of AL densities, when $\mu > 0$, the density skews to the right. On the other hand, when $\mu < 0$, the density skews to the left.

Important results of univariate and multivariate asymptotic Laplace distributions that will be used later on are presented below.

Proposition 5.2.1. *(See Kotz [68])*

(1). If $\boldsymbol{X} = (X_1, \cdots, X_n)$ follows multivariate Asymmetric Laplace distribution, i.e., $\boldsymbol{X} \sim AL_n\,(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, n is the number of securities. The linear combination $\boldsymbol{w}'\boldsymbol{X} = w_1X_1 + \cdots + w_nX_n$ follows univariate Asymmetric Laplace distribution, i.e. $\boldsymbol{w}'\boldsymbol{X} \sim AL_1\,(\mu, \sigma)$, with $\mu = \boldsymbol{w}'\boldsymbol{\mu}\,, \sigma = \sqrt{\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}}\,, \boldsymbol{w} = (w_1, \cdots, w_n)'$.

*(2). Assume that univariate random variable $Y \sim AL_1\,(\mu, \sigma)$. To measure the asymmetry and peakedness of the distribution, define the skewness (*Skew[Y]*) and*

*kurtosis (*Kurt[Y]*) as the third and fourth standardized moment of a random variable Y . Then,*

$$\begin{aligned}\text{Skew[Y]} &= \frac{\mathbb{E}(Y-\mathbb{E}Y)^3}{\left[\mathbb{E}(Y-\mathbb{E}Y)^2\right]^{3/2}} = \frac{2\mu^3+3\mu\sigma^2}{(\mu^2+\sigma^2)^{3/2}},\\ \text{Kurt[Y]} &= \frac{\mathbb{E}(Y-\mathbb{E}Y)^4}{\left[Var(Y)\right]^2} = \frac{9\mu^4+6\sigma^4+18\mu^2\sigma^2}{(\mu^2+\sigma^2)^2}.\end{aligned}$$

(3). Let $\boldsymbol{X} = (X_1, X_2, \cdots, X_n) \sim AL_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the first and second order moments of $\boldsymbol{X}$ are

$$\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu} \quad and \quad Cov(\boldsymbol{X}) = \boldsymbol{\Sigma} + \boldsymbol{\mu}'\boldsymbol{\mu}.$$

(4). The Asymmetric Laplace distribution can be represented as a mixture of normal vector and a standard exponential variable, i.e., $\boldsymbol{X} \sim AL_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be represented as

$$\boldsymbol{X} = \boldsymbol{\mu}Z + Z^{1/2}\boldsymbol{Y},$$

where $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$, $Z \sim Exp(1)$. This indicate that we can simulate multivariate Asymmetric Laplace random vector $\boldsymbol{X} \sim AL_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as follows:

1. *Generate a multivariate normal variable $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$;*
2. *Generate a standard exponential variable $Z \sim Exp(1)$;*
3. *Construct Asymmetric Laplace random vector as $\boldsymbol{X} = \boldsymbol{\mu}Z + Z^{1/2}\boldsymbol{Y}$.*

Figure 5.4 displays several realizations of bivariate Asymmetric Laplace distribution with different level of asymmetry and peakedness.

Risk measures

Since mean and covariance matrix cannot be used to characterize non-Gaussian distribution, alternative risk measures are necessary for portfolio selection problems.

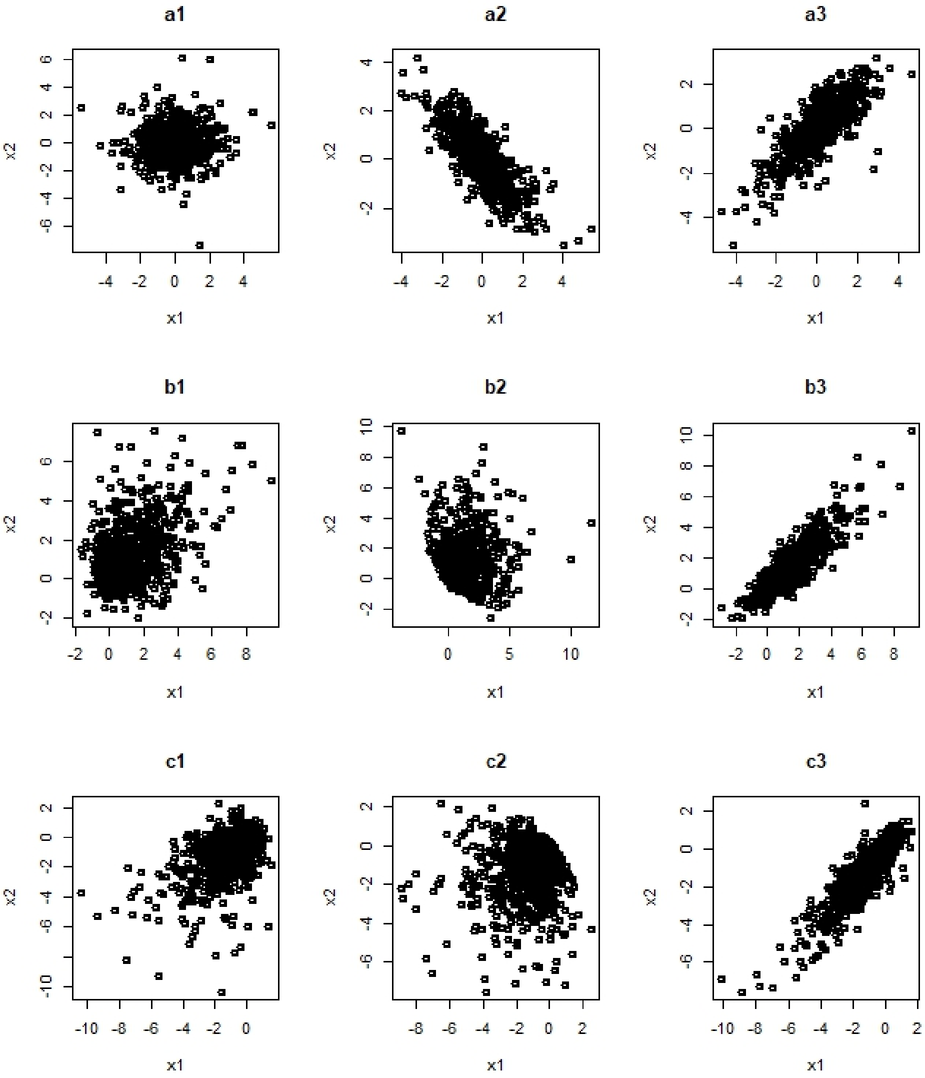

Figure 5.4: Bivariate Asymmetric Laplace data with μ cases: (a1, a2, a3): $\mu = (0, 0)$; (b1, b2, b3): $\mu = (1, 1)$; (c1, c2, c3): $\mu = (-1, -1)$. Covariance matrix $\mathbf{\Sigma}$ cases. (a1, b1, c1): $\sigma_{11} = \sigma_{22} = 1, \sigma_{12} = \sigma_{21} = 0$; (a2, b2, c2): $\sigma_{11} = \sigma_{22} = 1\,, \sigma_{12} = \sigma_{21} = 0.8$; (a3, b3, c3): $\sigma_{11} = \sigma_{22} = 1\,, \sigma_{12} = \sigma_{21} = -0.8$.

Artzner et al. [6] suggests that a desirable risk measure should be defined fulfilling certain properties and such a risk measure is said to be coherent.

A risk measure ϕ that maps a random variable to a real number is coherent if it satisfies the following conditions:

1). Translation invariance: $\phi(l + h) = \phi(l) + h$, for all random losses l and all

$h \in \mathbb{R}$;

2). Subadditivity: $\phi(l + h) \leqslant \phi(l) + \phi(h)$, for all random losses l, h;

3). Positive homogeneity: $\phi(\lambda l) = \lambda\phi(l)$ for all random losses l and all $\lambda > 0$;

4). Monotonicity: $\phi(l_1) \leqslant \phi(l_2)$ for all random losses l_1 , l_2 with $l_1 \leqslant l_2$ almost surely.

Standard deviation is not coherent in general excepting the Gaussian cases. VaR is coherent when the underlying distribution is elliptically distributed. Expected Shortfall, or the so-called conditional value at risk (CVaR) is a coherent risk measure since it always satisfies subadditivity, monotonicity, positive homogeneity, and convexity. For any fixed $\alpha \in (0, 1)$, α-VaR is the α-quantile loss while α-ES is the average of all β-VaR for $\beta \in (\alpha, 1)$. Both VaR and CVaR measure the potential maximal loss. VaR and ES can be written as

$$\mathrm{VaR}_\alpha = F^{-1}(\alpha) \quad \text{and} \quad \mathrm{ES}_\alpha = E[L|L \leqslant \text{-}\mathrm{VaR}_\alpha] = -\frac{1}{\alpha}\int_{-\infty}^{-\mathrm{VaR}_\alpha} \mathrm{VaR}_\beta d\beta,$$

where $F(\cdot)$ is the cumulative distribution function of loss L and ES_α is the expected loss above VaR_α. Thus, the estimation process are

$$\int_{-\infty}^{-\mathrm{VaR}_\alpha} f_X(x)dx = \alpha \quad \text{and} \quad \mathrm{ES}_\alpha = -\frac{1}{\alpha}\int_{-\infty}^{-\mathrm{VaR}_\alpha} xf(x)dx. \tag{5.42}$$

Under normality assumption, VaR_α and ES_α are

$$\mathrm{VaR}_\alpha = \mu + \sigma\Phi^{-1}(1 - \alpha),$$
$$\mathrm{ES}_\alpha = \mu + \sigma\frac{\psi(\Phi^{-1}(1 - \alpha))}{\alpha},$$

where $\psi(\cdot)$ as the normal density distribution, and $\Phi^{-1}(\cdot)$ is the quantile distribution. As shown in Hu et al. [49], portfolio selected by minimizing standard deviation, VaR_α, and ES_α are the equivalent under elliptical distribution assumption.

It is well-documented that asset securities are not normally distributed. As an alternative to Gaussian distribution, Asymmetric Laplace distribution exhibits tail-heaviness, skewness, and peakedness.

5.2.1 Portfolio selection under ALD framework

Let $\boldsymbol{X} = (X_1, X_2, \cdots, X_n) \sim AL_n(\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})$ be the return vectors of n securities, and $\boldsymbol{w} = (w_1, w_2, \cdots, w_n)'$ is the allocation weight vector. Then, the portfolio is

$$\mathcal{P}(\boldsymbol{w}) = \boldsymbol{w}'\boldsymbol{X} = \sum_{i=1}^{n} w_i X_i.$$

According to Proposition 5.2.1 (2), $\mathcal{P}(\boldsymbol{w}) \sim AL_1(\mu, \sigma)$ with $\mu = \boldsymbol{w}'\boldsymbol{\mu}\,, \sigma = \sqrt{\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}}$.

From Theorem 5.1–5.2 below, in order to select a portfolio under Asymmetric Laplace distribution, it suffices to obtain the unknown parameters $\boldsymbol{\mu}_{\mathcal{P}}$ and $\boldsymbol{\Sigma}_{\mathcal{P}}$. Thus portfolio selection models under Asymmetric Laplace distribution lead to parameter estimation for $AL_n(\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})$. Zhao [132] proposed the multi-objective portfolio selection model under Asymmetric Laplace framework and derived the simplified model that can be reformulated as quadratic programming problem. However, to estimate the unknown parameters, the authors adopt a moment estimation method that is less efficient compared to maximum likelihood method. Since Asymmetric Laplace distribution can be represented as a mixture of exponential distribution and multivariate normal distribution, we derived the Expectation-Maximization algorithm for parameter estimation of Asymmetric Laplace distribution. The algorithm for estimating these unknown parameters is discussed in Section 5.2.1.

Portfolio selection theorems

Theorem 5.1. *Let $\boldsymbol{X} = (X_1, \cdots, X_n) \sim AL_n(\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})$ be a n-dimensional random vector that follow multivariate Asymmetric Laplace distribution, each element $(X_i, i = 1, 2, \cdots, n)$ represent a stock. Let $\boldsymbol{w}$ be the weight vector and $\mathcal{P}(\boldsymbol{w}) =$*

$\boldsymbol{w}'\boldsymbol{X} = \sum_{i=1}^{n} w_i X_i$ *be the portfolio. Then, under Asymmetric Laplace framework, risk measures of StD, VaR$_\alpha$, and ES$_\alpha$ at $\alpha \in (0,1)$ level formulated as*

$$\text{Standard Deviation:} \quad \text{StD}\big(\mathcal{P}(\boldsymbol{w})\big) = \frac{\sigma}{\sqrt{2}};$$

$$\text{Value at Risk:} \quad \text{VaR}_\alpha\big(\mathcal{P}(\boldsymbol{w})\big) = -\frac{\sigma^2}{\gamma+\mu}\ln\frac{\alpha\gamma(\gamma+\mu)}{\sigma^2};$$

$$\text{Expected Shortfall:} \quad \text{ES}_\alpha\big(\mathcal{P}(\boldsymbol{w})\big) = \frac{\sigma^2}{\gamma+\mu} - \frac{\sigma^2}{\gamma+\mu}\ln\frac{\alpha\gamma(\gamma+\mu)}{\sigma^2}.$$

Here, $\mu = \boldsymbol{w}'\boldsymbol{\mu}_{\mathcal{P}} = mean\big(\mathcal{P}(\boldsymbol{w})\big)$, $\sigma = \sqrt{\boldsymbol{w}'\boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{w}} = std\big(\mathcal{P}(\boldsymbol{w})\big)$ *and* $\gamma = \sqrt{\mu^2 + 2\sigma^2}$.

Proof. Let $\boldsymbol{\mu}_{\mathcal{P}} = (\mu_1, \cdots, \mu_n)$ be the mean return vector of the securities $(X_1, \cdots, X_n)$ and $\Sigma_{\mathcal{P}} = \big(\sigma_{\mathcal{P}}\big)_{i,j=1}^{p}$ be the scale matrix of $(X_1, \cdots, X_n)$. Denote the allocation vector by $\boldsymbol{w} = (w_1, \cdots, w_n)'$. Then, the portfolio $\mathcal{P}(\boldsymbol{w}) = \sum_{i=1}^{n} w_i X_i$ follows univariate Asymmetric Laplace distribution with

$$\mathcal{P}(\boldsymbol{w}) = \sum_{i=1}^{n} w_i X_i \sim AL_1(\mu, \sigma) \quad \text{with} \quad \mu = \sum_{i=1}^{n} \mu_i w_i\,, \sigma = \Big(\sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_{\mathcal{P}_{ij}} w_i w_j\Big)^{1/2}.$$

If $\boldsymbol{\mu} = \boldsymbol{0}_n$, the univariate symmetric Asymmetric Laplace distribution becomes $AL_1(0, \sigma)$ with density

$$g(x) = \frac{1}{\gamma}\exp\Big\{-\frac{|x|}{\sigma^2}\gamma\Big\} \quad \text{with} \quad \gamma = \sqrt{2}\sigma.$$

Thus, standard deviation (StD) of portfolio $\mathcal{P}(\boldsymbol{w}) = \boldsymbol{w}'\boldsymbol{X}$ is

$$\text{StD}\left(\mathcal{P}(\boldsymbol{w})\right) = \int_{-\infty}^{+\infty} \frac{1}{\gamma}|x|\exp\Big\{-\frac{\gamma}{\sigma^2}|x|\Big\}dx = 2\int_{0}^{+\infty} \frac{x}{\gamma}\exp\Big\{-\frac{\gamma}{\sigma^2}x\Big\}dx = \frac{2\sigma^4}{\gamma^3} = \frac{\sigma}{\sqrt{2}}.$$

According to the definition of VaR$_\alpha$ and ES$_\alpha$ as defined in (5.42) and univariate

Asymmetric Laplace density (5.41), we have

$$
\begin{aligned}
&-\int_{-\infty}^{-\mathrm{VaR}_\alpha} \frac{1}{\gamma}\exp\Big\{-\frac{|x|}{\sigma^2}\big[\gamma-\mu\cdot\mathrm{sgn}(x)\big]\Big\}dx=\alpha,\\
&\frac{\sigma^2}{\gamma(\gamma+\mu)}\exp\Big\{-\frac{\gamma+\mu}{\sigma^2}\mathrm{VaR}_\alpha\Big\}=\alpha.
\end{aligned}
$$

Thus, VaR_α and ES_α are

$$
\begin{aligned}
\mathrm{VaR}_\alpha\big(\mathcal{P}(\boldsymbol{w})\big) &= -\frac{\sigma^2}{\sqrt{\mu^2+2\sigma^2}+\mu}\ln\frac{\alpha(\mu^2+2\sigma^2+\mu\sqrt{\mu^2+2\sigma^2})}{\sigma^2}\\
&= -\frac{\sigma^2}{\gamma+\mu}\ln\frac{\alpha\gamma(\gamma+\mu)}{\sigma^2};\\
\mathrm{ES}_\alpha\big(\mathcal{P}(\boldsymbol{w})\big) &= -\frac{1}{\alpha}\int_{-\infty}^{-\mathrm{VaR}_\alpha} x f_X(x)dx\\
&= -\frac{1}{\alpha}\int_{-\infty}^{-\mathrm{VaR}_\alpha} x\frac{1}{\gamma}\exp\Big\{-\frac{|x|}{\sigma^2}\big[\gamma-\mu\cdot\mathrm{sgn}(x)\big]\Big\}dx\\
&= \frac{\sigma^2}{\mu+\sqrt{\mu^2+2\sigma^2}}-\frac{\sigma^2}{\mu+\sqrt{\mu^2+2\sigma^2}}\ln\Big\{2\alpha+\frac{\alpha(\mu^2+\mu\sqrt{\mu^2+2\sigma^2})}{\sigma^2}\Big\}\\
&= \frac{\sigma^2}{\gamma+\mu}-\frac{\sigma^2}{\gamma+\mu}\ln\frac{\alpha\gamma(\gamma+\mu)}{\sigma^2}.
\end{aligned}
$$

□

Then we have the following theorem.

Theorem 5.2. *Let $\boldsymbol{X}\sim AL_n(\boldsymbol{\mu}_{\mathcal{P}},\boldsymbol{\Sigma}_{\mathcal{P}})$. Then, portfolio $\mathcal{P}(\boldsymbol{w})=\boldsymbol{w}'\boldsymbol{X}$ with following* *models based on ES_α, VaR_α, and StD (as defined in Theorem 5.1)*

$$
\begin{aligned}
&\min_{\boldsymbol{w}} \mathrm{ES}_\alpha\big(\mathcal{P}(\boldsymbol{w})\big) \ \text{ or } \ \min_{\boldsymbol{w}} \mathrm{VaR}_\alpha\big(\mathcal{P}(\boldsymbol{w})\big) \ \text{ or } \ \min_{\boldsymbol{w}} \mathrm{StD}\big(\mathcal{P}(\boldsymbol{w})\big)\\
&\max_{\boldsymbol{w}} \quad \mathrm{Skew}[\mathcal{P}(\boldsymbol{w})] = \frac{2\mu^3 + 3\mu\sigma^2}{(\mu^2+\sigma^2)^{3/2}}\\
&\max_{\boldsymbol{w}} \quad \mathrm{Kurt}[\mathcal{P}(\boldsymbol{w})] = \frac{9\mu^4 + 6\sigma^4 + 18\mu^2\sigma^2}{(\mu^2+\sigma^2)^2}\\
&\text{s.t.} \qquad \boldsymbol{w}'\boldsymbol{\mu} = r_0\,, \boldsymbol{w}'\boldsymbol{1} = 1,
\end{aligned}
$$

are equivalent. Here, $\mu = \boldsymbol{w}'\boldsymbol{\mu}_{\mathcal{P}} = mean\big[\mathcal{P}(\boldsymbol{w})\big]\,, \sigma = \sqrt{\boldsymbol{w}'\boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{w}} = std\big[\mathcal{P}(\boldsymbol{w})\big]\,, \boldsymbol{w} = (w_1, w_2, \cdots, w_n)'$.

Proof. Let $g(\mu,\sigma) = \frac{\sigma^2}{\mu+\sqrt{\mu^2+2\sigma^2}}$. Then, $\mathrm{ES}_\alpha[\mathcal{P}(\boldsymbol{w})]$ and $\mathrm{VaR}_\alpha[\mathcal{P}(\boldsymbol{w})]$ are

$$
\begin{aligned}
\mathrm{VaR}_\alpha[\mathcal{P}(\boldsymbol{w})] &= -g(\mu,\sigma)\ln\big(2\alpha + \frac{\alpha\mu}{g(\mu,\sigma)}\big) = -g(\mu,\sigma)\big[\ln\alpha + \ln\big(2 + \frac{\mu}{g(\mu,\sigma)}\big)\big],\\
\mathrm{ES}_\alpha[\mathcal{P}(\boldsymbol{w})] &= g(\mu,\sigma) - g(\mu,\sigma)\ln\big(2\alpha + \frac{\alpha\mu}{g(\mu,\sigma)}\big)\\
&= (1-\ln\alpha)g(\mu,\sigma) - g(\mu,\sigma)\ln(2 + \frac{\mu}{g(\mu,\sigma)}).
\end{aligned}
$$

Differentiating the above expressions with respect to σ, we have

$$
\begin{aligned}
\frac{\partial \mathrm{VaR}_\alpha[\mathcal{P}(\boldsymbol{w})]}{\partial\sigma} &= \frac{\partial g(\mu,\sigma)}{\partial\sigma}\left[-\ln\alpha - \ln\big(2 + \frac{\mu}{g(\mu,\sigma)}\big) + \frac{\frac{\mu}{g(\mu,\sigma)}}{2 + \frac{\mu}{g(\mu,\sigma)}}\right] > 0,\\
\frac{\partial \mathrm{ES}_\alpha[\mathcal{P}(\boldsymbol{w})]}{\partial\sigma} &= \frac{\partial g(\mu,\sigma)}{\partial\sigma}\left[1 - \ln\alpha - \ln\big(2 + \frac{\mu}{g(\mu,\sigma)}\big) + \frac{\frac{\mu}{g(\mu,\sigma)}}{2 + \frac{\mu}{2+g(\mu,\sigma)}}\right] > 0,
\end{aligned}
$$

where

$$
\frac{\partial g(\mu,\sigma)}{\partial\sigma} = \frac{\partial\Big[\frac{\sigma^2}{\mu+\sqrt{\mu^2+2\sigma^2}}\Big]}{\partial\sigma} = 2\sigma\frac{\mu + \frac{\mu^2+\sigma^2}{\sqrt{\mu^2+2\sigma^2}}}{(\mu^2 + \sqrt{\mu^2+2\sigma^2})^2} > 0.
$$

The derivative of skewness measure with respect to σ is

$$\frac{\partial \mathrm{Skew}[\mathcal{P}(\boldsymbol{w})]}{\partial \sigma} = \frac{\partial \left[\frac{2\mu^3+3\mu\sigma^2}{(\mu^2+\sigma^2)^{3/2}}\right]}{\partial \sigma} = \frac{-3\mu\sigma^3}{(\mu^2+\sigma^2)^{5/2}} < 0.$$

The derivative of kurtosis measure with respect to σ is

$$\frac{\partial \mathrm{Kurt}[\mathcal{P}(\boldsymbol{w})]}{\partial \sigma} = \frac{\frac{9\mu^4+6\sigma^4+18\mu^2\sigma^2}{(\mu^2+\sigma^2)^2}}{\partial \sigma} = \frac{-12\mu^4\sigma^3 - 12\mu^2\sigma^5}{(\mu^2+\sigma^2)^4} < 0.$$

The monotonicity of $\mathrm{VaR}_\alpha[\mathcal{P}(\boldsymbol{w})]$, $\mathrm{ES}_\alpha[\mathcal{P}(\boldsymbol{w})]$, $\mathrm{Skew}[\mathcal{P}(\boldsymbol{w})]$, and $\mathrm{Kurt}[\mathcal{P}(\boldsymbol{w})]$ with respect to σ indicate that the portfolio selection problems based on these risk measures are equivalent. This means that minimizing $\mathrm{VaR}_\alpha[\mathcal{P}(\boldsymbol{w})]$, $\mathrm{ES}_\alpha[\mathcal{P}(\boldsymbol{w})]$, $\mathrm{StD}[\mathcal{P}(\boldsymbol{w})]$ are equivalent to minimizing $\boldsymbol{w}'\boldsymbol{\Sigma}_{\mathcal{P}}\boldsymbol{w}$. □

Parameter estimation of Asymmetric Laplace Distribution

Assume $\boldsymbol{X} = (X_1, X_2, \cdots, X_n) \sim AL_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T \in \mathbb{R}^n$ be the T observations. We aim at fitting a multivariate Asymmetric Laplace distribution $AL_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with unknown parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$.

Hrlimann [55], Kollo and Srivastava [66], Visk [112] consider moment matching methods that is less efficient than maximum likelihood estimation. Kotz [69] and Kotz [68] presented the maximum likelihood estimators for parameter estimation of Asymmetric Laplace distributions. However, maximum likelihood estimation require computation of complicated Bessel function. Thus we derived the expectation-maximization algorithm for parameter estimation of Asymmetric Laplace distribution.

Moment estimation

As Zhao [132] pointed out, according to Proposition 5.2.1 (3), Asymmetric Laplace distribution can be estimated via moment method (Moment-AL) with

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \mathrm{cov}(\boldsymbol{X}) - \hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}},$$

where $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\,, \mathrm{Cov}(\boldsymbol{X}) = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$.

Maximum likelihood estimation

Consider sample points $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ and density function of Asymmetric Laplace distribution as defined in (5.38). Taken logarithm with respect to likelihood function, the log-likelihood is

$$\begin{aligned}
\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{t=1}^{T} \ln f(\boldsymbol{x}_t; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \sum_{t=1}^{T} \boldsymbol{x}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + T\ln 2 - \frac{Tn}{2}\ln(2\pi) - \frac{T}{2}\ln\big(|\boldsymbol{\Sigma}|\big) + \frac{\nu}{2}\sum_{t=1}^{T}\ln\big(\boldsymbol{x}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_t\big) - \\
&\quad \frac{\nu T}{2}\ln\big(2 + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\big) + \sum_{t=1}^{T}\ln K_v\Big\{\sqrt{(2 + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})(\boldsymbol{x}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_t)}\Big\}.
\end{aligned}$$

Generally, we can directly maximize the log-likelihood function $\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ and thus obtain the maximum likelihood estimator. Unfortunately, the density function involves modified Bessel function of the third kind with density (5.39), (5.40) that are too complex and complicated for numerical maximization. However, Gaussian-Exponential mixture representation of the Asymmetric Laplace distribution allows us to employ the expectation-maximization algorithm without involving modified Bessel functions.

Expectation-Maximization algorithm

Then we derive the Expectation-Maximization algorithm for parameter estimation of multivariate Asymmetric Laplace Distribution (mALD), we follow the EM derivation for Multivariate Skew Laplace distribution in [4].

Let $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ be Asymmetric Laplace distributed random vector. Proposition 5.2.1 suggests that $\boldsymbol{X}$ can be generated from a latent random variable

$Z = z$ through multivariate Gaussian distribution with $z\boldsymbol{\mu}, z\boldsymbol{\Sigma}$, i.e. $X|Z = z \sim N_n(z\boldsymbol{\mu}, z\boldsymbol{\Sigma}), Z \sim \text{Exp}(1)$ with density

$$\begin{aligned} f_{\boldsymbol{X}|Z}(\boldsymbol{x}, z) &= \frac{1}{(2\pi)^{n/2}|z\boldsymbol{\Sigma}|^{1/2}} \exp\Big\{ -\frac{1}{2}(\boldsymbol{x} - z\boldsymbol{\mu})'(z\boldsymbol{\Sigma})^{-1}(\boldsymbol{x} - z\boldsymbol{\mu})\Big\}, \\ f_Z(z) &= e^{-z}1_{\{z \geqslant 0\}}. \end{aligned}$$

Thus the joint density function of $\boldsymbol{X}$ and Z is

$$\begin{aligned} & f_{\boldsymbol{X},Z}(\boldsymbol{x}, z) = f_{\boldsymbol{X}|Z}(\boldsymbol{x}, z) f_Z(z) \\ = & \frac{1}{(2\pi)^{\frac{n}{2}} z^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\Big\{ -\frac{1}{2z}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{z}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - z1_{\{z \geqslant 0\}}\Big\}. \end{aligned}$$

Suppose that there are T observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_T$ generated from the latent random variables $z_1, z_2, \cdots, z_T$ respectively. The complete data is defined as $\{(\boldsymbol{x}_t, z_t)\}, t = 1, 2, \cdots, T$. In the EM algorithm, $\boldsymbol{x}_t$ and z_t are the observed and missing data respectively. The log-likelihood up to an additive constant can be written as

$$\begin{aligned} & \tilde{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{t=1}^{T} \ln f_{\boldsymbol{X},Z}(\boldsymbol{x}_t, z_t) \\ = & -\frac{T}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{z_t}\boldsymbol{x}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_t + \sum_{t=1}^{T}\boldsymbol{x}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\sum_{t=1}^{T} z_t - \sum_{t=1}^{T} z_t 1_{\{z_t \geqslant 0\}}. \end{aligned}$$

Note that the last term of the above equation does not contain any unknown parameters and thus is negligible. Then, the E-step becomes

$$\begin{aligned} & E\Big(\tilde{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma})|\boldsymbol{x}_t, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\Big) \propto \\ & \frac{1}{2}\sum_{t=1}^{T} E(z_t^{-1}|\boldsymbol{x}_t, \hat{\boldsymbol{\mu}}, \hat{\Sigma})\boldsymbol{x}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_t - \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\sum_{t=1}^{T} E(z_t|\boldsymbol{x}_t, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}). \end{aligned}$$

where $E(z_t|\boldsymbol{x}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ and $E(z_t^{-1}|\boldsymbol{x}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ are the conditional expectations of z_t and z_t^{-1} given $\boldsymbol{x}_t$ and the current estimates $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$.

To evaluate conditional expectations $E(z_t^{-1}|\boldsymbol{x}_t, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ and $E(z_t|\boldsymbol{x}_t, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, we need the conditional density of Z given $\boldsymbol{X}$, $f_{Z|\boldsymbol{X}}$. After some straightforward algebra, the conditional distribution of Z given $\boldsymbol{X}$ is an inverse Gaussian distribution with density function

$$
\begin{aligned}
f_{Z|\boldsymbol{X}}(z|\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{f_{\boldsymbol{X},Z}(\boldsymbol{x}, z)}{f_{\boldsymbol{X}}(\boldsymbol{x})} \qquad (5.43) \\
&= \frac{\frac{1}{(2\pi)^{\frac{n}{2}} z^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\big\{ -\frac{1}{2z}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{z}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - z1_{\{z\geqslant 0\}}\big\}}{\frac{2e^{\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}\big(\frac{\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}{2+\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}\big)^{v/2} K_v\big(\sqrt{(2+\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})(\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x})}\big)} \\
&= \Big(\frac{2+\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}\Big)^{v/2} z^{-n/2} \frac{\exp\big\{ -\frac{1}{2}\big[z^{-1}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + z(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + z1_{\{z\geqslant 0\}})\big]\big\}}{2K_v\big(\sqrt{(2+\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})(\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x})}\big)}.
\end{aligned}
$$

Lemma 5.1. *(GIG [104]) A random variable X follows Generalized Inverse Gaussian distribution(denoted as $X \sim N^{-}(\lambda, \chi, \psi)$) if its density function could be represented as*

$$
f(x) = \frac{\chi^{-\lambda}(\sqrt{\chi\psi})^{\lambda}}{2K_{\lambda}(\sqrt{\chi\psi})} x^{\lambda-1} \exp\Big\{ -\frac{1}{2}(\chi x^{-1} + \psi x)\Big\}, \quad x > 0.
$$

where K_λ denotes the third kind modified Bessel function, and the parameters satisfy

$$
\begin{cases}
\chi > 0, \psi \geqslant 0, & \textit{if } \lambda < 0; \\
\chi > 0, \psi > 0, & \textit{if } \lambda = 0; \\
\chi \geqslant 0, \psi > 0, & \textit{if } \lambda > 0.
\end{cases}
$$

After some algebraic manipulations, it is easy to show that $Z|\boldsymbol{X}$ follows Generalized Inverse Gaussian distribution:

$$
Z|\boldsymbol{X} \sim N^{-}\Big(\frac{2-n}{2}, \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\,, 2 + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\Big).
$$

If $\chi > 0, \psi > 0$, the moments could be calculated through the following formulas:

$$
E(X^{\alpha}) = \Big(\frac{\chi}{\psi}\Big)^{\alpha/2} \frac{K_{\lambda+\alpha}(\sqrt{\chi\psi})}{K_{\lambda}(\sqrt{\chi\psi})}, \quad \alpha \in \mathbb{R},
$$

$$
E(\ln X) = \frac{dE(X^{\alpha})}{d\alpha}\Big|_{\alpha=0}.
$$

Denote $\chi = \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}, \psi = 2 + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Then, $Z|\boldsymbol{X} \sim N^{-}(\frac{2-n}{2}, \chi, \psi)$. From the conditional density function of (5.43), we can obtain the conditional expectations with the following moment properties:

$$\begin{aligned} a_t &= E(z_t|\boldsymbol{x}_t, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \sqrt{\frac{\chi_t}{\psi} \frac{K_{\frac{n}{2}-2}(\sqrt{\chi_t \psi})}{K_{\frac{n}{2}-1}(\sqrt{\chi_t \psi})}}, \quad t = 1, 2, \cdots, T, \\ b_t &= E(z_t^{-1}|\boldsymbol{x}_t, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \sqrt{\frac{\psi}{\chi_t} \frac{K_{\frac{n}{2}}(\sqrt{\chi_t \psi})}{K_{\frac{n}{2}-1}(\sqrt{\chi_t \psi})}}, \quad t = 1, 2, \cdots, T, \end{aligned}$$

where $\chi_t = \boldsymbol{x}_t'\Sigma^{-1}\boldsymbol{x}_t$, $R(\lambda) = \frac{K_{\lambda+1}(x)}{K_{\lambda}(x)}$ is strictly decreasing in x with $\lim_{x\to\infty} R_\lambda(x) = 1$ and $\lim_{x\to 0^+} R_\lambda(x) = \infty$. Thus, $a_t > 0, b_t > 0\,, t = 1, 2, \cdots, T$.

Finally, if the conditional expectation $E(z_t|\boldsymbol{x}_t, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ and $E(z_t^{-1}|\boldsymbol{x}_t, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ are replaced by a_t and b_t respectively, the objective function becomes

$$Q(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{x}_t, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = -\frac{T}{2}\ln|\boldsymbol{\Sigma}| + \sum_{t=1}^{T} \boldsymbol{x}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\sum_{t=1}^{T} b_t \boldsymbol{x}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_t - \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\sum_{t=1}^{T} a_t. \tag{5.44}$$

Denote $\boldsymbol{S} = \boldsymbol{\Sigma}^{-1}$. The objective function (5.44) becomes

$$Q(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{x}_t, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{S}}) = \frac{T}{2}\ln|\boldsymbol{S}| + \sum_{t=1}^{T} \boldsymbol{x}_t'\boldsymbol{S}\boldsymbol{\mu} - \frac{1}{2}\sum_{t=1}^{T} b_t \boldsymbol{x}_t'\boldsymbol{S}\boldsymbol{x}_t - \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{S}\boldsymbol{\mu}\sum_{t=1}^{T} a_t. \tag{5.45}$$

Taking derivative of objective function (5.45) with respect to $\boldsymbol{\mu}, \boldsymbol{S}$, we obtain

$$\begin{aligned} \frac{\partial Q}{\partial \boldsymbol{\mu}} &= \sum_{t=1}^{T} \boldsymbol{x}_t'\boldsymbol{S} - \sum_{t=1}^{T} a_t \boldsymbol{\mu}'\boldsymbol{S} = 0, \\ \frac{\partial Q}{\partial \boldsymbol{S}} &= \frac{T}{2}\boldsymbol{S}^{-1} - \frac{1}{2}\sum_{t=1}^{T} b_t \boldsymbol{x}_t'\boldsymbol{x}_t + \sum_{t=1}^{T} \boldsymbol{x}_t'\boldsymbol{\mu} - \frac{1}{2}\sum_{t=1}^{T} a_t \boldsymbol{\mu}'\boldsymbol{\mu} = 0. \end{aligned}$$

Substituting $\boldsymbol{S}$ by $\boldsymbol{\Sigma}$ and setting these derivatives to zero yield

$$\sum_{t=1}^{T} \boldsymbol{x}_t' \boldsymbol{\Sigma}^{-1} - \sum_{t=1}^{T} a_t \boldsymbol{\mu}' \boldsymbol{\Sigma}^{-1} = \mathbf{0},$$

$$\frac{T}{2}\boldsymbol{\Sigma} - \frac{1}{2}\sum_{t=1}^{T} b_t \boldsymbol{x}_t' \boldsymbol{x}_t + \sum_{t=1}^{T} \boldsymbol{x}_t' \boldsymbol{\mu} - \frac{1}{2}\sum_{t=1}^{T} a_t \boldsymbol{\mu}' \boldsymbol{\mu} = \mathbf{0}.$$

Thus, maximization of the objective function $Q(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{x}_t, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ can be achieved by the following iterative updating formulas:

$$\hat{\boldsymbol{\mu}} = \frac{\bar{\boldsymbol{x}}}{\bar{a}}; \quad \hat{\boldsymbol{\Sigma}} = \overline{b_t \boldsymbol{x}_t' \boldsymbol{x}_t} - \frac{\bar{\boldsymbol{x}}' \bar{\boldsymbol{x}}}{\bar{a}}.$$

where $\bar{a}, \bar{b}$ stand for the average of $\{a_t\}_{t=1}^{T}$ and $\{b_t\}_{t=1}^{T}$ respectively and $\bar{\boldsymbol{x}}$ is the average of $\{\boldsymbol{x}_t\}_{t=1}^{T}$. In what follows, we present the iterative reweighted Expectation-Maximization algorithm for parameter estimation of Asymmetric Laplace distribution.

5.2.2 Numerical experiments

To evaluate the performance of portfolio selection models and parameter estimation methods in Section 5.2.1, we generate 100 datasets from Gaussian distribution and Asymmetric Laplace distribution respectively. Each dataset consists of $T = 200$ observations with the following parameter settings: Case (1): $n = 3\,, \boldsymbol{\mu} = (0.03\,, 0.06\,, 0.09)$; Case (2): $n = 5\,, \boldsymbol{\mu} = (0.01\,, 0.02\,, 0.06\,, 0.08\,, 0.09)$; Case (3): $n = 10\,, \boldsymbol{\mu} = (0.01\,, 0.02\,, 0.03\,, \cdots, 0.10)$. For each case, we set $\boldsymbol{\Sigma} = \mathrm{diag}\,(\boldsymbol{\mu}/10)$. All the simulation studies are carried out on a PC with Intel Core i7 3.6 GHz processor under `R` platform.

Each dataset are estimated under both multivariate Gaussian and Asymmetric Laplace distribution. ALD (EM-AL) is estimated using the EM algorithm described in Section 5.2.1. We evaluate the estimation performance using Bias measure, defined

Algorithm 5 Iterative reweighting algorithm

1. Set iteration number $k = 1$ and select initial estimates for parameters $\boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}$.

2. (E-Step) At k-th iteration with current estimates $\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}$, define the corresponding log-likelihood as

   $$l^{(k)} = \log \sum_{t=1}^{T} f(\boldsymbol{x}_t | \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}), \quad k = 1, 2, \cdots .$$

   With notations $\chi_t = \boldsymbol{x}_t' \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_t, \psi = 2 + \boldsymbol{\mu}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$, we can obtain iterative weights

   $$\begin{aligned} a_t &= E(z_t | \boldsymbol{x}_t, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \sqrt{\frac{\chi_t}{\psi}} \frac{K_{2-\frac{n}{2}}(\sqrt{\chi_t \psi})}{K_{1-\frac{n}{2}}(\sqrt{\chi_t \psi})}, \quad t = 1, 2, \cdots, T; \\ b_t &= E(z_t^{-1} | \boldsymbol{x}_t, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \sqrt{\frac{\psi}{\chi_t}} \frac{K_{-\frac{n}{2}}(\sqrt{\chi_t \psi})}{K_{1-\frac{n}{2}}(\sqrt{\chi_t \psi})}, \quad t = 1, 2, \cdots, T. \end{aligned}$$

3. (M-Step) Employ the following iteration formulas to calculate the new estimates $\boldsymbol{\mu}^{(k+1)}, \boldsymbol{\Sigma}^{(k+1)}$ at $(k+1)$-th iteration:

   $$\boldsymbol{\mu}^{(k+1)} = \frac{\bar{\boldsymbol{x}}}{\bar{a}}, \quad \boldsymbol{\Sigma}^{(k+1)} = \overline{b_t \boldsymbol{x}_t' \boldsymbol{x}_t} - \frac{\bar{\boldsymbol{x}}' \bar{\boldsymbol{x}}}{\bar{a}}. \tag{5.46}$$

   The log-likelihood at $(k+1)$-th iteration becomes

   $$l^{(k+1)} = \log \sum_{t=1}^{T} f(\boldsymbol{x}_t | \boldsymbol{\mu}^{(k+1)}, \boldsymbol{\Sigma}^{(k+1)}).$$

4. Repeat these iteration steps until convergence with criterion $l^{(k+1)} - l^{(k)} < \varepsilon$, where $\varepsilon > 0$ is a small number that control the convergence precision, for convenience, we take $\varepsilon = 1e^{-16}$.

as Bias $= \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1 + \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_1$. The mean log-likelihood and mean bias of the simulated 200 datasets are reported in Table 5.5.

Table 5.5 indicate that if the model is correctly specified, the estimation performance is always the best in terms of bias. If the data is generated from Gaussian distribution, the estimation from Gaussian model is the best, so does Asymmetric Laplace distribution. If data is generated from Gaussian distribution, then the estimation log-likelihood of Gaussian model is larger than Asymmetric Laplace distribution, this is true for Asymmetric Laplace data as well.

Gaussian Data					
	Log-Likelihood		Bias		
	Gauss	EM-AL	Gauss	Moment-AL	EM-AL
Case (1)	709.6159	641.2472	0.0153	0.0451	0.0183
Case (2)	1363.0231	1260.3676	0.0280	0.0886	0.0319
Case (3)	2593.1151	2432.2192	0.0672	0.3309	0.0766

Asymmetric Laplace Data					
	Log-Likelihood		Bias		
	Gauss	EM-AL	Gauss	Moment-AL	EM-AL
Case (1)	619.1798	735.0847	0.0542	0.0252	0.0226
Case (2)	1250.7110	1463.9149	0.0870	0.0376	0.0302
Case (3)	2432.3401	2919.9878	0.3750	0.1304	0.0832

Table 5.5: Model fitting results of Gaussian data and Asymmetric Laplace data using Gauss model and EM-AL model.

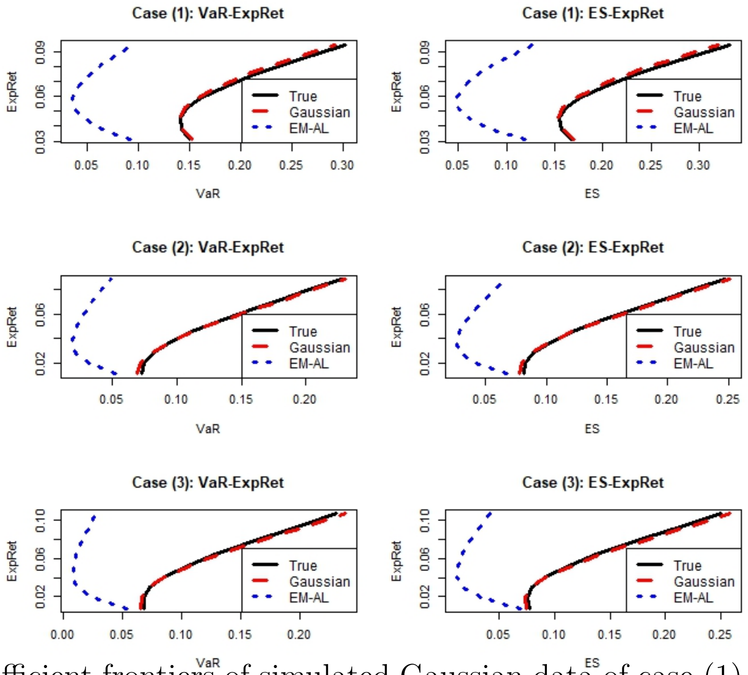

Figure 5.5: Efficient frontiers of simulated Gaussian data of case (1)-(3) using Gaussian and EM-AL model.

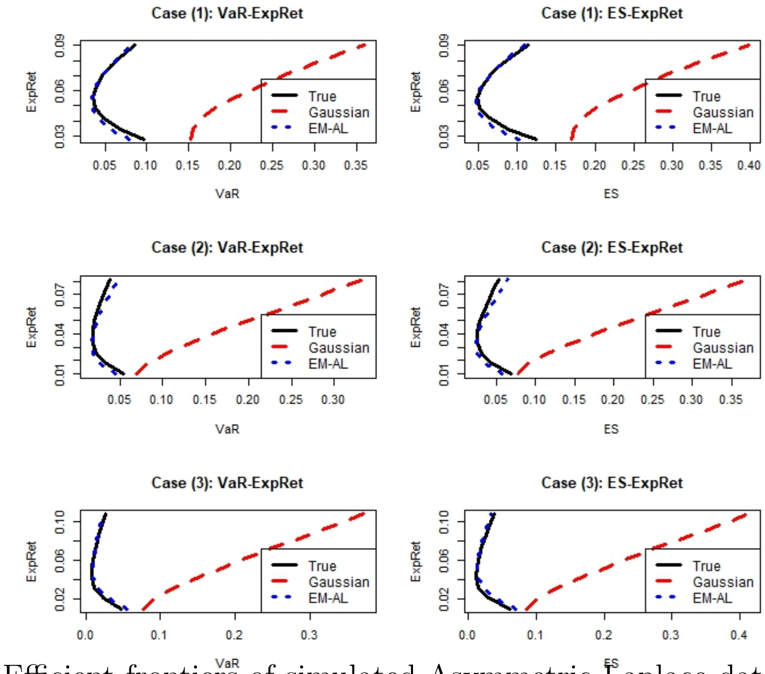

Figure 5.6: Efficient frontiers of simulated Asymmetric Laplace data of case (1)-(3) using Gaussian and EM-AL model.

Figure 5.5 show that for Gaussian data, since Gaussian data fit the model better, efficient frontiers under Gaussian data are more close to Gaussian models; Figure 5.6 indicate that for generated Asymmetric Laplace data, efficient frontiers nearly equivalent to true Asymmetric Laplace data. Figure 5.5-5.6 suggest that we can first modeling data using Gaussian and Asymmetric Laplace distribution, and use the fitted log-likelihood to determine the distribution, then we evaluate the performance with the corresponding efficient frontier analysis.

Real data analysis

Then we apply our proposed methodology to two real financial datasets, Hang Seng Index and Nasdaq Index, both datasets are downloaded from Bloomberg, with daily data range from January 4, 2011, to December 29, 2017. The variable of interest is the rate of returns multiplied by the annualized ratio $\sqrt{252}$, formulated as

$$\text{LogRet}\,(t) = \sqrt{252}\Big\{\log\big(\text{price}[t+1]\big) - \log\big(\text{price}[t]\big)\Big\}, \quad t = 1, 2, \cdots, 1721.$$

These two datasets are analyzed through efficient frontier analysis under ALD framework on `R` platform. Theorem 5.1-5.2 indicate that portfolio selection models under ALD framework can be reduced to the following quadratic programming problem:

$$\min_{\boldsymbol{w}} \sigma^2 = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} \quad \text{s.t.} \quad \boldsymbol{w}'\boldsymbol{\mu} = r_0\,, \boldsymbol{w}'\mathbf{1} = 1.$$

The explicit solution (see [74]) is as follows:

$$\hat{\boldsymbol{w}} = \frac{D - r_0 B}{AD - B^2}\boldsymbol{\Sigma}^{-1}\mathbf{1} + \frac{r_0 A - B}{AD - B^2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}. \tag{5.47}$$

Here, $A = \mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}, B = \mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $D = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.

Example 1: Hang Seng Index

In the first example, we construct a portfolio consisting of 8 Hang Seng indexes: HK1, HK175, HK2007, HK2318, HK4, HK6, HK66. The summary descriptive statistics are reported in Table 5.6. It is clear that the all stock returns exhibit larger skewness and kurtosis. The median of these stocks are close to zero, the log-likelihood of Asymmetric Laplace distribution is larger than gaussian distribution, indicating that Asymmetric Laplace distribution would be a good fit than gaussian distribution. Then we fit the data to Asymmetric Laplace distribution through EM algorithm as described in Section 5.2.1. Parameter estimation results are displayed in Table 5.6, we construct portfolios under Asymmetric Laplace framework at different levels of expected return. Consider increasing target expected return values

$$r_0 = 0.040\,, 0.0745\,, 0.1081\,, 0.1417\,, 0.1753\,, 0.2088\,, 0.2424\,, 0.2760\,, 0.3096\,, 0.3431\,.$$

Portfolio selection results are summarized in Table 5.7, with kurtosis, skewness, sharpe ratio, VaR and ES results at $\alpha = 0.01, 0.05, 0.10$ levels.

Descriptive Statistics							
	StD	Mean	Median	Skewness	Kurtosis	Jarq.Test	Jarq.Prob
HK1	19.7819	0.2370	0.0000	3.9601	71.8163	375244.7434	0.0000
HK175	3.7101	0.2183	0.0000	1.1510	28.6192	59264.8528	0.0000
HK2007	2.1968	0.1115	0.0000	0.5928	16.5623	19825.4392	0.0000
HK2318	12.6007	0.3431	0.0000	0.6174	6.2371	2908.6947	0.0000
HK4	5.8690	0.0409	0.0000	0.4894	7.6362	4263.7750	0.0000
HK6	13.0712	0.1517	0.0000	-1.1430	20.3112	30037.3385	0.0000
HK66	5.8708	0.1545	0.0000	-1.7941	19.9392	29510.3684	0.0000

	Gaussian	EM-AL
Log-likelihood	-39707.45	-37865.83

Parameter Estimation							
	HK1	HK175	HK2007	HK2318	HK4	HK6	HK66
μ	0.2370	0.2183	0.1115	0.3431	0.0409	0.1517	0.1545
Σ	HK1	HK175	HK2007	HK2318	HK4	HK6	HK66
HK1	406.4426	12.8089	11.4547	130.0934	68.3857	94.4069	60.3087
HK175	12.8089	7.3656	1.6330	12.0356	4.7467	4.8725	3.6070
HK2007	11.4547	1.6330	3.8506	9.9261	4.6830	3.9055	2.4395
HK2318	130.0934	12.0356	9.9261	174.0423	42.1879	48.2877	35.0499
HK4	68.3857	4.7467	4.6830	42.1879	44.6336	26.5314	17.6232
HK6	94.4069	4.8725	3.9055	48.2877	26.5314	205.1471	32.5570
HK66	60.3087	3.6070	2.4395	35.0499	17.6232	32.5570	41.6153

Table 5.6: Hang Seng data statistics

	r_0	μ	σ	Skew	Kurt	Sharpe
1	0.0409	0.0409	2.5461	0.0482	6.0016	0.0161
2	0.0745	0.0745	2.0565	0.1086	6.0079	0.0362
3	0.1081	0.1081	1.7299	0.1869	6.0233	0.0625
4	0.1417	0.1417	1.6650	0.2537	6.0430	0.0851
5	0.1753	0.1753	1.8891	0.2763	6.0510	0.0928
6	0.2088	0.2088	2.3199	0.2682	6.0480	0.0900
7	0.2424	0.2424	2.8656	0.2523	6.0425	0.0846
8	0.2760	0.2760	3.4724	0.2372	6.0375	0.0795
9	0.3096	0.3096	4.1134	0.2247	6.0337	0.0753
10	0.3431	0.3431	4.7749	0.2147	6.0307	0.0719

	$\text{VaR}_{0.01}$	$\text{ES}_{0.01}$	$\text{VaR}_{0.05}$	$\text{ES}_{0.05}$	$\text{VaR}_{0.10}$	$\text{ES}_{0.10}$
1	6.9429	8.7229	4.0782	5.8582	2.8444	4.6244
2	5.5080	6.9254	3.2268	4.6442	2.2444	3.6618
3	4.5256	5.6960	2.6420	3.8124	1.8308	3.0011
4	4.2684	5.3771	2.4841	3.5928	1.7156	2.8243
5	4.8095	6.0606	2.7960	4.0471	1.9288	3.1799
6	5.9208	7.4601	3.4434	4.9827	2.3764	3.9157
7	7.3493	9.2579	4.2774	6.1861	2.9544	4.8631
8	8.9467	11.2679	5.2108	7.5321	3.6018	5.9231
9	10.6386	13.3966	6.1999	8.9578	4.2882	7.0462
10	12.3871	15.5962	7.2222	10.4313	4.9978	8.2069

Table 5.7: Efficient frontier results of Hang Seng data

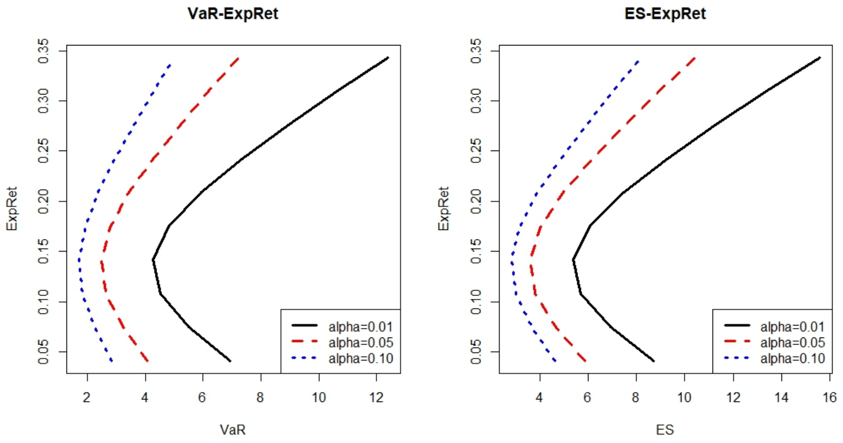

Figure 5.7: ALD efficient frontier of Hang Seng data

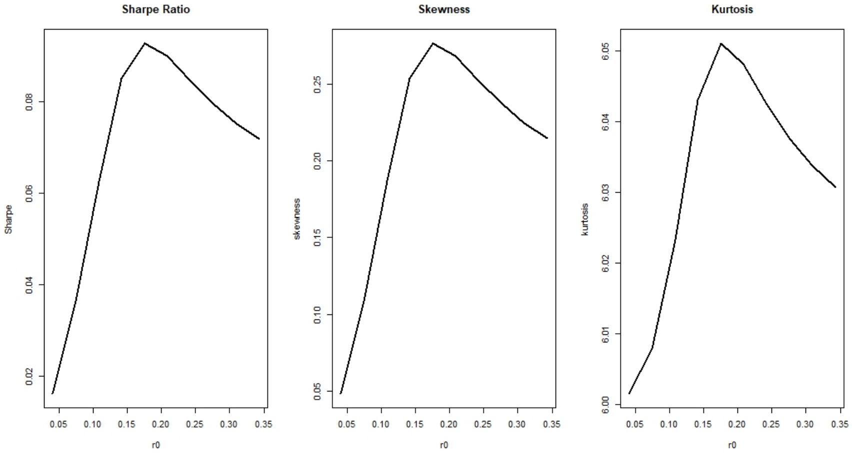

Figure 5.8: Skewness, Kurtosis and Sharpe ratio tendency of Hang Seng data

The efficient frontier tendencies are displayed in Figure 5.7. It is suggested that aggressive investors should impose higher confidence levels and conservative investors may choose smaller confidence levels. Figure 5.8 depicts the kurtosis, skewness, and sharpe ratio tendency of portfolio selection models. Results show that Sharpe Ratio, Skewness and Kurtosis increase fast and decreases slowly down as the target expected returns increases.

Example 2: Nasdaq Index

In the second example, we consider Nasdaq index, including CTRP, MNST, NFLX, NTES, NVDA, TTWO, and report the descriptive statistics in Table 5.8. All the indexes exhibit significant skewness and kurtosis. Jarq.Test results indicate that this dataset deviates from normality significantly. We fit the log returns data to Gaussian and Asymmetric Laplace distributions. Since Asymmetric Laplace model achieve higher log-likelihood results compared to Gaussian model, we choose EM-AL model for data fitting. Parameter estimation results are displayed in Table 5.8.

Descriptive Statistics							
	StD	Mean	Median	Skewness	Kurtosis	Jarq.Test	Jarq.Prob
CTRP	12.5853	0.2100	0.0000	1.5907	16.5100	20786.4126	0.0000
MNST	10.0679	0.4904	0.1587	1.9632	25.7988	50065.6119	0.0000
NFLX	32.9311	1.5015	-0.0079	1.0314	20.0251	29796.6074	0.0000
NTES	58.0910	2.7817	1.2700	1.2457	26.0418	50315.0307	0.0000
NVDA	25.2360	1.6024	0.3175	1.7413	40.3981	120864.0386	0.0000
TTWO	12.2826	0.8786	0.1587	2.1392	52.3926	203127.9368	0.0000

	Gaussian	EM-AL
Log-Likelihood	-46400.25	-42798.03

Parameter Estimation						
	CTRP	MNST	NFLX	NTES	NVDA	TTWO
μ	0.2100	0.4904	1.5015	2.7817	1.6024	0.8786
Σ	CTRP	MNST	NFLX	NTES	NVDA	TTWO
CTRP	187.1236	25.4281	129.9104	246.6268	46.1539	40.0369
MNST	25.4281	116.7107	51.2572	82.0634	26.2438	23.5405
NFLX	129.9104	51.2572	900.1261	327.6980	122.5361	83.8650
NTES	246.6268	82.0634	327.6980	2380.8894	213.1414	121.1493
NVDA	46.1539	26.2438	122.5361	213.1414	259.7019	53.8530
TTWO	40.0369	23.5405	83.8650	121.1493	53.8530	111.0020

Table 5.8: Nasdaq data statistics

Then we consider increasing target expected returns

$$r_0 = 0.2100, 0.4958, 0.7815, 1.0673, 1.3530, 1.6388, 1.9245, 2.2102, 2.4960, 2.7817.$$

Results of skewness, kurtosis, sharpe ratio and VaR, ES results are summarized in Table 5.9. Figure 5.9 displays the efficient frontiers at confidence level $\alpha =$

$0.01, 0.05, 0.10$. These results show that the portfolio capture higher risk at higher α levels. Figure 5.10 displays the skewness, kurtosis, and Sharpe ratio tendency. The optimal portfolios can be obtained from Eqn. (5.47) with the corresponding VaR, ES, skewness, kurtosis and Sharpe ratio.

Table 5.9 and Figure 5.9 suggests that as r_0 increases, all ES ($\mathrm{ES}_{0.01}$, $\mathrm{ES}_{0.05}$, $\mathrm{ES}_{0.10}$) increases, indicating that higher return is derived from higher risk. It is interesting that under the ALD assumption, as r_0 increases, Sharpe ratio and skewness first decreases then increases accordingly. As α increases, VaR and ES measures decreases. Thus, conservative investors can choose larger α levels and aggressive investors would select smaller α levels.

	r	μ	σ	Skew	Kurt	Sharpe
1	0.2100	0.2100	8.5237	0.0739	6.0036	0.0246
2	0.4958	0.4958	7.5957	0.1951	6.0254	0.0653
3	0.7815	0.7815	7.8219	0.2973	6.0590	0.0999
4	1.0673	1.0673	9.1167	0.3472	6.0806	0.1171
5	1.3530	1.3530	11.1127	0.3608	6.0870	0.1218
6	1.6388	1.6388	13.5025	0.3597	6.0865	0.1214
7	1.9245	1.9245	16.1117	0.3541	6.0838	0.1194
8	2.2102	2.2102	18.8494	0.3478	6.0808	0.1173
9	2.4960	2.4960	21.6671	0.3418	6.0781	0.1152
10	2.7817	2.7817	24.5371	0.3365	6.0757	0.1134

	$\mathrm{VaR}_{0.01}$	$\mathrm{ES}_{0.01}$	$\mathrm{VaR}_{0.05}$	$\mathrm{ES}_{0.05}$	$\mathrm{VaR}_{0.10}$	$\mathrm{ES}_{0.10}$
1	23.0672	28.9903	13.5344	19.4575	9.4288	15.3519
2	19.8220	24.9509	11.5675	16.6963	8.0125	13.1413
3	19.7857	24.9397	11.4907	16.6447	7.9183	13.0723
4	22.7065	28.6414	13.1547	19.0896	9.0409	14.9758
5	27.5608	34.7713	15.9560	23.1665	10.9581	18.1686
6	33.4993	42.2627	19.3952	28.1586	13.3208	22.0843
7	40.0422	50.5132	23.1898	33.6608	15.9318	26.4028
8	46.9392	59.2083	27.1927	39.4619	18.6883	30.9575
9	54.0563	68.1800	31.3251	45.4488	21.5353	35.6590
10	61.3178	77.3329	35.5424	51.5576	24.4416	40.4567

Table 5.9: Efficient frontier analysis of Nasdaq data

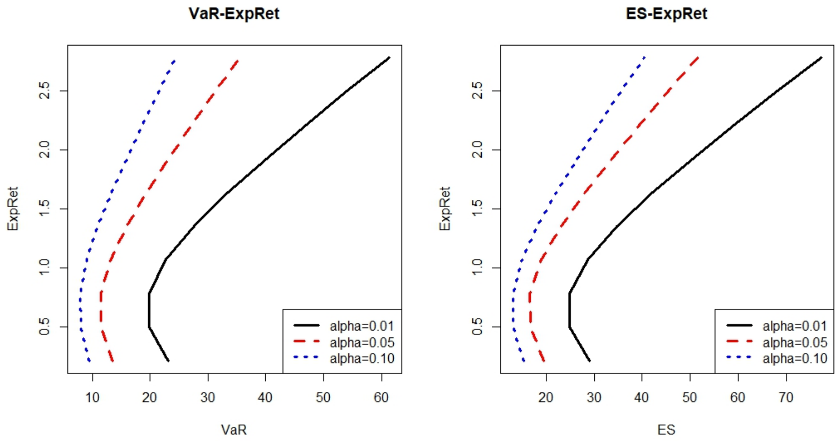

Figure 5.9: ALD efficient frontier of Hang Seng data

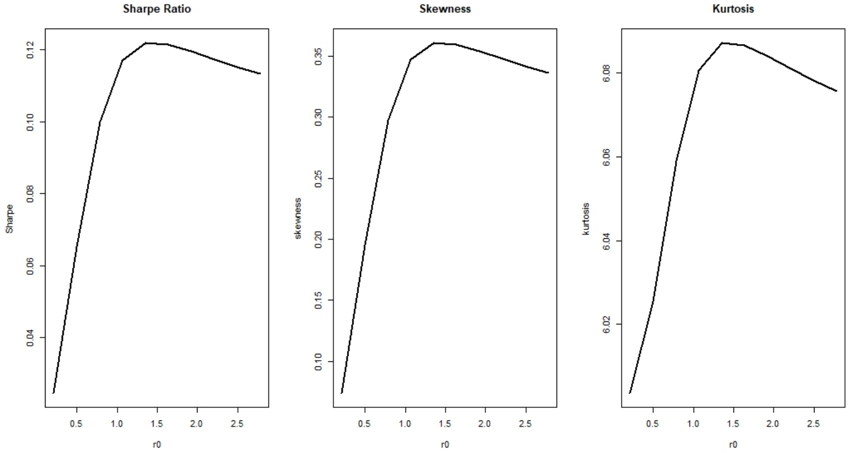

Figure 5.10: Skewness, Kurtosis and Sharpe ratio tendency of Nasdaq data

Chapter 6

Concluding Remarks

In this chapter, we summarize the main results in this thesis and provide some possible directions for further research.

6.1 Summary

This thesis consider LAD Generalized Lasso models, Constrained LAD Lasso models, selection of penalty parameter for compressive sensing and two models under Asymmetric Laplace Distributions.

We first studied the LAD-Lasso problem, and derived the optimality condition and a descent algorithm such that the nonsmooth optimization problem can be optimized directly. Numerical experiments with both simulated and real data have been employed to demonstrate that our proposed method is more efficient than the traditional interior point method and the state-of-the-art LP solver Gurobi. We also proposed a new Active Zero Set Descent (AZSD) algorithm for LAD Generalized Lasso problem. The main advantage of this algorithm is that the zero set can be obtained without any user-chosen threshold values. Moreover, the algorithm is proved to be convergent in finite steps where the stopping condition can be described through infinitely many basis directional set explicitly. Nested iteration as in the interior point algorithm is not needed. The estimation performances of the proposed method are investigated based on simulation studies.

Then we proposed the MAD-Lasso portfolio selection strategy that can be reformulated as Constrained LAD Lasso with linearly equality constraints. Based on nonsmooth optimality conditions, we derived a descent algorithm by updating descent directions from basis directional set and optimal step length iteratively, extensive simulation studies and real data analysis show that our methodology is much more time efficient than interior point method. For portfolio selection results, the MAD-Lasso model can robustify portfolio selection models and encourages sparsity.

Next we present a two-level optimization approach to incorporate quality measures in a speech application such as compressive speech enhancement. The results show that quality measures can be factored in the solutions by hyperparameterizing the tuning parameter in the sparse reconstruction. By doing so, the solutions are effectively tailored to the desired design attributes by a single parameter. The two-level approach first compresses the big data and subsequently optimizes the sparse the solution via the AIC, BIC model selection and the Gini performance index. The set of solutions is then measured against the quality measures for the desired solution. Comprehensive numerical experiments in a range of real-world noise with varying S-NRs show that proper tuning of the hyperparameter can effectively trade-off between speech distortion and noise suppression.

Finally, we consider two models under Asymmetric Laplace Distributions. We first conduct a new robust procedure for mixture of regression with the assumption of mixture asymmetric Laplace outliers under different skewness levels. An EM algorithm is derived to estimate parameter upon the fact that the Asymmetric Laplace distribution is a mixture of exponential and normal distribution. Extensive simulations and real data analysis confirmed the efficiency of our algorithm. Then we derive several equivalent portfolio selection methods under Asymmetric Laplace Distribution framework that can be transformed to quadratic programming problem with explicit solutions. The Expectation-Maximization algorithm for parameter estimation of Asymmetric Laplace distribution is obtained and outperforms moment estimation. There are several advantages of ALD models. First, the equivalence of risk measures such as VaR, ES and StD faciliate the portfolio selection process significantly. Second, the confidence levels of these models offer investors various portfolio selection choices.

6.2 Further research

The investigation of LAD Generalized Lasso models, Constrained LAD Lasso models, selection of penalty parameter for compressive sensing and two models under Asymmetric Laplace distributions in this thesis is a start for exploration of Lasso regression and Asymmetric Laplace distributions. These topics are promising and inspiring, some possible future works are as follows.

- **LAD Generalized Lasso models**.

  The descent algorithm for LAD Generalized Lasso Models performs well under nonsmooth optimization conditions. It is interesting to study the non-full-rank cases in the future so that even $p \gg n$ cases can be handled. Another possible research direction is to develop new algorithm for change-point detection problem.

- **Constrained LAD Lasso models**.

  We derive a descent algorithm for Constrained LAD Lasso with linearly equality constraints, with applications in MAD-Lasso models. Further possible directions include partial index tracking, portfolio hedging and portfolio adjustment under MAD-Lasso framework.

- **Selection of penalty parameter for compressive sensing**.

Speech enhancement with compressive sensing is promising and intriging, we may further investigate other signal performance measures, such as intelligibility.

- **Two models under Asymmetric Laplace Distributions**.

  For mixture linear regression models, it is promising to extend the model to high dimensional case; For portfolio selection under Asymmetric Laplace distribution, we can further consider Bayesian models.

Bibliography

[1] M. Abramowitz, I.A. Stegun, et al. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Dover publications New York, 1972.

[2] H Akaike. Information theory as an extension of the maximum likelihood. Proceeding if IEEE international symposium on information theory, 1973.

[3] P. Alquier. An algorithm for iterative selection of blocks of features. In *Algorithmic Learning Theory*, pages 35–49. Springer, 2010.

[4] O. Arslan. An alternative multivariate skew laplace distribution: properties and estimation. *Statistical Papers*, 51(4):865–887, 2010.

[5] O. Arslan. Weighted lad-lasso method for robust parameter estimation and variable selection in regression. *Computational Statistics & Data Analysis*, 56(6):1952–1965, 2012.

[6] P. Artzner, F. Delbaen, J.M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

[7] A. Ayebo and T.J. Kozubowski. An asymmetric generalization of gaussian and laplace laws. *Journal of Probability and Statistical Science*, 1(2):187–210, 2003.

[8] O.E. Barndorff-Nielsen. Normal inverse gaussian distributions and the modeling of stock returns. 1995.

[9] C. Barnes and S.P. Lillford. Decision support for the design of affective products. *Journal of Engineering Design*, 20(5):477–492, 2009.

[10] C. Beaulieu, J. Chen, and J.L. Sarmiento. Change-point analysis as a tool to detect abrupt climate variations. *Phil. Trans. R. Soc. A*, 370(1962):1228–1249, 2012.

[11] A. Behr and U. Pötter. Alternatives to the normal model of stock returns: Gaussian mixture, generalised logf and generalised hyperbolic models. *Annals of Finance*, 5(1):49–68, 2009.

[12] J. Benesty, S. Makino, and J. Chen. Speech enhancement, ser. signals and communication technology, 2005.

[13] N.H. Bingham and R. Kiesel. Modelling asset returns with hyperbolic distributions. In *Return Distributions in Finance*, pages 1–20. Elsevier, 2001.

[14] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[15] M. Broadie. Computing efficient frontiers using estimated parameters. *Annals of Operations Research*, 45(1):21–58, 1993.

[16] J. Brodie, I. Daubechies, C. De Mol, D. Giannone, and I. Loris. Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

[17] E.J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[18] E.J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

[19] E.J. Candès and M.B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[20] M. Carrasco and N. Noumon. Optimal portfolio selection using regularization. Technical report, Citeseer, 2011.

[21] S. Chen, D. Donoho, and M. Suanders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[22] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[23] E.A. Cohen. Some effects of inharmonic partials on interval perception. *Music Perception: An Interdisciplinary Journal*, 1(3):323–349, 1984.

[24] J.J. Dahlgaard, S. Schütte, E. Ayas, and S. Mi Dahlgaard-Park. Kansei/affective engineering design: a methodology for profound affection and attractive quality creation. *The TQM Journal*, 20(4):299–311, 2008.

[25] A. Dalalyan and Y. Chen. Fused sparsity and robust estimation for linear models with unknown variance. In *Advances in Neural Information Processing Systems*, pages 1259–1267, 2012.

[26] M. Davidian and D.M. Giltinan. *Nonlinear models for repeated measurement data*, volume 62. CRC press, 1995.

[27] V. DeMiguel, L. Garlappi, and R. Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2007.

[28] L. Dicker, B. Huang, and X. Lin. Variable selection and estimation with the seamless-l_0 penalty. *Statistica Sinica*, pages 929–962, 2013.

[29] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[30] E. Eberlein. Application of generalized hyperbolic lévy motions to finance. In *Lévy Processes*, pages 319–336. Springer, 2001.

[31] E. Eberlein, U. Keller, et al. Hyperbolic distributions in finance. *Bernoulli*, 1(3):281–299, 1995.

[32] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[33] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[34] B. Fastrich, S. Paterlini, and P. Winker. Penalized least squares for optimal sparse portfolio selection. In *COMPSTAT 2014 Conference Proceedings, International Conference on Computational Statistics*, 2014.

[35] M. Fernandes, G. Rocha, and T. Souza. Regularized minimum-variance portfolios using asset group information, 2012.

[36] K. Frick, A. Munk, and H. Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580, 2014.

[37] P.A. Frost and J.E. Savarino. For better performance: Constrain portfolio weights. *The Journal of Portfolio Management*, 15(1):29–34, 1988.

[38] B.R. Gaines and H. Zhou. Algorithms for fitting the constrained lasso. *arXiv preprint arXiv:1611.01511*, 2016.

[39] X. Gao and J. Huang. Asymptotic analysis of high-dimensional lad regression with lasso. *Statistica Sinica*, pages 1485–1506, 2010.

[40] X. Gao and J. Huang. A robust penalized method for the analysis of noisy dna copy number data. *BMC Genomics*, 11(1):517, 2010.

[41] T.J. Gardner and M.O. Magnasco. Sparse time-frequency representations. *Proceedings of the National Academy of Sciences*, 103(16):6094–6099, 2006.

[42] M. Geraci and M. Bottai. Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8(1):140–154, 2007.

[43] P.K. Ghosh, A. Tsiartas, and S. Narayanan. Robust voice activity detection using long-term signal variability. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):600–613, 2011.

[44] P.R. Gill, A. Wang, and A. Molnar. The in-crowd algorithm for fast basis pursuit denoising. *IEEE Transactions on Signal Processing*, 59(10):4595–4605, 2011.

[45] D. Harrison and D.L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.

[46] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[47] M. Hellmich and S. Kassberger. Efficient and robust portfolio optimization in the multivariate generalized hyperbolic framework. *Quantitative Finance*, 11(10):1503–1516, 2011.

[48] W. Hu and A. Kercheval. Risk management with generalized hyperbolic distributions. In *Proceedings of the Fourth IASTED International Conference on Financial Engineering and Applications*, pages 19–24. ACTA Press, 2007.

[49] W. Hu and A.N. Kercheval. Portfolio optimization for student t and skewed t returns. *Quantitative Finance*, 10(1):91–105, 2010.

[50] Y. Hu and P.C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238, 2008.

[51] T. Huang, B. Wu, P. Lizardi, and H. Zhao. Detection of dna copy number alterations using penalized least squares regression. *Bioinformatics*, 21(20):3811–3817, 2005.

[52] D.R. Hunter and K. Lange. Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77, 2000.

[53] D.R. Hunter and R. Li. Variable selection using mm algorithms. *Annals of Statistics*, 33(4):1617, 2005.

[54] N. Hurley and S. Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

[55] W. Hürlimann. A moment method for the multivariate asymmetric laplace distribution. *Statistics & Probability Letters*, 83(4):1247–1253, 2013.

[56] G.M. James, C. Paulson, and P. Rusmevichientong. Penalized and constrained regression. Technical report, mimeo, Marshall School of Business, University of Southern California, 2013.

[57] H. Jiang, C.K. Kwong, Y. Liu, and W.H. Ip. A methodology of integrating affective design with defining engineering specifications for product design. *International Journal of Production Research*, 53(8):2472–2488, 2015.

[58] J.D. Jobson and B. Korkie. Estimation for markowitz efficient portfolios. *Journal of the American Statistical Association*, 75(371):544–554, 1980.

[59] D. Karlis. An em type algorithm for maximum likelihood estimation of the normal-inverse gaussian distribution. *Statistics* & *Probability letters*, 57(1):43–52, 2002.

[60] R. Killick, P. Fearnhead, and I.A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[61] S.J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior point method for large scale ℓ_1 regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):606–617, 2007.

[62] R. Koenker. *Quantile regression.* Cambridge university press, 2005.

[63] R. Koenker and Bassett J.G. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.

[64] R. Koenker and J.A.F. Machado. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310, 1999.

[65] R. Koenker and B.J. Park. An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71(1):265–283, 1996.

[66] T. Kollo and M.S. Srivastava. Estimation and testing of parameters in multivariate laplace distribution. *Communications in Statistics-Theory and Methods*, 33(10):2363–2387, 2005.

[67] H. Konno and H. Yamazaki. Mean-absolute deviation portfolio optimization model and its applications to tokyo stock market. *Management Science*, 37(5):519–531, 1991.

[68] S. Kotz, T.J. Kozubowski, and K. Podgórski. Asymmetric multivariate laplace distribution. In *The Laplace Distribution and Generalizations*, pages 239–272. Springer, 2001.

[69] S. Kotz, T.J. Kozubowski, and K. Podgórski. Maximum likelihood estimation of asymmetric laplace parameters. *Annals of the Institute of Statistical Mathematics*, 54(4):816–826, 2002.

[70] T.J. Kozubowski and K. Podgórski. A class of asymmetric distributions. *Actuarial Research Clearing House*, 1:113–134, 1999.

[71] T.J. Kozubowski and K. Podgórski. Asymmetric laplace laws and modeling financial data. *Mathematical and Computer Modelling*, 34(9-11):1003–1021, 2001.

[72] H. Kozumi and G. Kobayashi. Gibbs sampling methods for bayesian quantile regression. *Journal of Statistical Computation and Simulation*, 81(11):1565–1578, 2011.

[73] C.K. Kwong, H. Jiang, and X.G. Luo. Ai-based methodology of integrating affective design, engineering, and marketing for defining design specifications of new products. *Engineering Applications of Artificial Intelligence*, 47:49–60, 2016.

[74] T.L. Lai and H. Xing. *Statistical models and methods for financial markets.* Springer, 2008.

[75] H. Li and H. Sieling. Fdrseg: Fdr-control in multiscale change-point segmentation. *R package version*, pages 1–0, 2015.

[76] Q. Li and L. Wang. Robust change point detection method via adaptive ladlasso. *Statistical Papers*, pages 1–13, 2017.

[77] J. Liu, S. Ji, J. Ye, et al. Slep: Sparse learning with efficient projections. *Arizona State University*, 6(491):7, 2009.

[78] W.M. Liu, K.A. Jellyman, J.S.D. Mason, and N.W.D. Evans. Assessment of objective quality measures for speech intelligibility estimation. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006.

[79] P. Loizou. Speech quality assessment. *Multimedia Analysis, Processing and Communications*, pages 623–654, 2011.

[80] P.C. Loizou. *Speech enhancement: theory and practice.* CRC press, 2013.

[81] A.M. Lokman. Ke as affective design methodology. In *Computer, Control, Informatics and Its Applications (IC3INA), 2013 International Conference on*, pages 7–13. IEEE, 2013.

[82] S.Y. Low, N. Grbic, and S. Nordholm. Robust microphone array using subband adaptive beamformer and spectral subtraction. In *Communication Systems, 2002. ICCS 2002. The 8th International Conference on*, volume 2, pages 1020–1024. IEEE, 2002.

[83] S.Y. Low, S. Nordholm, and K.L. Teo. Use of efficient frontier in microphone arrays. *Electronics Letters*, 42(20):1186–1187, 2006.

[84] S.Y. Low, D.S. Pham, and S. Venkatesh. Compressive speech enhancement. *Speech Communication*, 55(6):757–768, 2013.

[85] J. Ma, Y. Hu, and P.C. Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America*, 125(5):3387–3405, 2009.

[86] A. Mariani, A. Giorgetti, and M. Chiani. Model order selection based on information theoretic criteria: design of the penalty. *IEEE Trans. Signal Processing*, 63(11):2779–2789, 2015.

[87] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[88] C.T. Ng, W. Lee, Y. Lee, et al. Change-point estimators with true identification property. *Bernoulli*, 24(1):616–660, 2018.

[89] D. Peel and G.J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348, 2000.

[90] P. Perron et al. Dealing with structural breaks. *Palgrave Handbook of Econometrics*, 1(2):278–352, 2006.

[91] D.T. Pham, Z. El-Chami, A. Guérin, and C. Serviere. Modeling the short time fourier transform ratio and application to underdetermined audio source separation. In *ICA*, pages 98–105. Springer, 2009.

[92] R.F. Phillips. Least absolute deviations estimation via the em algorithm. *S-tatistics and Computing*, 12(3):281–285, 2002.

[93] S. Portnoy, R. Koenker, et al. The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300, 1997.

[94] B. Punathumparambath. The multivariate asymmetric slash laplace distribution and its applications. *Statistica*, 72(2):235, 2012.

[95] G. Rigaill. Pruned dynamic programming for optimal multiple change-point detection. *arXiv preprint arXiv:1004.0887*, 2010.

[96] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 2, pages 749–752. IEEE, 2001.

[97] T.E. Scheetz, K.Y.A. Kim, R.E. Swiderski, A.R. Philp, T.A. Braun, K.L. Knudtson, A.M. Dorrance, G.F. DiBona, J. Huang, T.L. Casavant, et al. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434, 2006.

[98] E.J. Schlossmacher. An iterative technique for absolute deviations curve fitting. *Journal of the American Statistical Association*, 68(344):857–859, 1973.

[99] G. Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[100] E. Seneta. Fitting the variance-gamma model to financial data. *Journal of Applied Probability*, 41(A):177–187, 2004.

[101] Y. Shi, Z. Feng, and K.F.C. Yiu. A descent method for least absolute deviation lasso problems. *Optimization Letters*, pages 1–17, 2017.

[102] W. Song, W. Yao, and Y. Xing. Robust mixture regression model fitting by Laplace distribution. *Computational Statistics & Data Analysis*, 71:128–137, 2014.

[103] T.V. Sreenivas and W.B. Kleijn. Compressive sensing for sparsely excited speech signals. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4125–4128. IEEE, 2009.

[104] E.W. Stacy. A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, pages 1187–1192, 1962.

[105] T.A. Stamey, J.N. Kabalin, J.E. McNeal, I.M. Johnstone, F. Freiha, E.A. Redwine, and N. Yang. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. *The Journal of Urology*, 141(5):1076–1083, 1989.

[106] B.A. Surya and R. Kurniawan. Optimal portfolio selection based on expected shortfall under generalized hyperbolic distribution. *Asia-Pacific Financial Markets*, 21(3):193–236, 2014.

[107] Konlack S.V. and D. Wilcox. A comparison of generalized hyperbolic distribution models for equity returns. *Journal of Applied Mathematics*, 2014, 2014.

[108] L. Tang, Z. Zhou, and C. Wu. The lad estimation of the change-point linear model with randomly censored data. *Communications in Statistics-Theory and Methods*, 45(2):479–491, 2016.

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

[110] R. Tibshirani and P. Wang. Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, 9(1):18–29, 2007.

[111] R.J. Tibshirani. *The solution path of the generalized lasso.* Stanford University, 2011.

[112] H. Visk. On the parameter estimation of the asymmetric multivariate laplace distribution. *Communications in Statistics– Theory and Methods*, 38(4):461–470, 2009.

[113] H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25(3):347–355, 2007.

[114] L. Wang. The l_1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151, 2013.

[115] L. Wang, M.D. Gordon, and J. Zhu. Regularized least absolute deviations regression and an efficient algorithm for parameter tuning. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 690–700. IEEE, 2006.

[116] L. Wang, Y. You, and H. Lian. A simple and efficient algorithm for fused lasso signal approximator with convex loss function. *Computational Statistics*, 28(4):1699–1714, 2013.

[117] S. Wang and L. Xiang. Two-layer em algorithm for ald mixture regression models: a new solution to composite quantile regression. *Computational Statistics & Data Analysis*, 115:136–154, 2017.

[118] G.A. Watson and K.F.C. Yiu. On the solution of the errors in variables problem using the l_1 norm. *BIT Numerical Mathematics*, 31(4):697–710, 1991.

[119] D. Wu, W.P. Zhu, and MNS Swamy. A compressive sensing method for noise reduction of speech and audio signals. In *Circuits and Systems (MWSCAS), 2011 IEEE 54th International Midwest Symposium on*, pages 1–4. IEEE, 2011.

[120] J. Xu and Z. Ying. Simultaneous estimation and variable selection in median regression using lasso-type penalty. *Annals of the Institute of Statistical Mathematics*, 62(3):487–514, 2010.

[121] W. Yao, Y. Wei, and C. Yu. Robust mixture regression using the t-distribution. *Computational Statistics & Data Analysis*, 71:116–127, 2014.

[122] Y.C. Yao. Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables. *The Annals of Statistics*, pages 1321–1328, 1987.

[123] Y.M. Yen et al. A note on sparse minimum variance portfolios and coordinatewise descent algorithms. *Available at SSRN*, 2010.

[124] Y.M. Yen and T.J. Yen. Solving norm constrained portfolio optimization via coordinate-wise descent algorithms. *Computational Statistics & Data Analysis*, 76:737–759, 2014.

[125] K.F.C. Yiu. Optimal portfolios under a value-at-risk constraint. *Journal of Economic Dynamics and Control*, 28(7):1317–1334, 2004.

[126] K.F.C. Yiu, K.Y. Chan, S.Y. Low, and S. Nordholm. A multi-filter system for speech enhancement under low signal-to-noise ratios. *Journal of Industrial and Management Optimization*, 2009.

[127] K.F.C. Yiu, X. Yang, S. Nordholm, and K.L. Teo. Near-field broadband beamformer design via multidimensional semi-infinite-linear programming techniques. *IEEE Transactions on Speech and Audio Processing*, 11(6):725–732, 2003.

[128] K. Yu and R.A. Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447, 2001.

[129] K. Yu and J. Zhang. A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics-Theory and Methods*, 34(9-10):1867–1879, 2005.

[130] Z. Zeebari. A simulation study on the least absolute deviations method for ridge regression. *Forthcoming in Communications in Statistics–Theory and Methods*, 2012.

[131] C.H. Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[132] S. Zhao, Q. Lu, L. Han, Y. Liu, and F. Hu. A mean-cvar-skewness portfolio optimization model based on asymmetric laplace distribution. *Annals of Operations Research*, 226(1):727–739, 2015.

[133] Y.H. Zhou, Z.X. Ni, and Y. Li. Quantile regression via the EM algorithm. *Communications in Statistics-Simulation and Computation*, 43(10):2162–2172, 2014.

[134] Y. Zhu. *Application of Asymmetric Laplace laws in financial risk measures and time series analysis.* PhD thesis, University of Florida, 2007.

[135] D. Zonoobi, A.A. Kassim, and Y.V. Venkatesh. Gini index as sparsity measure for signal reconstruction from compressive samples. *IEEE Journal of Selected Topics in Signal Processing*, 5(5):927–932, 2011.

[136] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[137] H. Zou, T. Hastie, R. Tibshirani, et al. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.