SIMULTANEOUS VARIABLE SELECTION AND OUTLIER DETECTION USING LASSO WITH APPLICATIONS TO AIRCRAFT LANDING DATA ANALYSIS

BY WEI LI

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Statistics

Written under the direction of

Regina Y. Liu, Minge Xie, and Cun-Hui Zhang

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2012

ABSTRACT OF THE DISSERTATION

Simultaneous Variable Selection and Outlier Detection Using LASSO with Applications to Aircraft Landing Data Analysis

by Wei Li

Dissertation Director: Regina Y. Liu, Minge Xie, and Cun-Hui Zhang

We propose a LASSO-type penalized regression method for simultaneous variable selection and outlier detection in high dimensional linear regression. We apply a mean-shift model to incorporate the coefficients associated with the potential outliers by expressing them as different intercept terms. The sparsity assumption is imposed on both X-covariates and the outlier indicator variables. With suitable penalty factors between X-covaraites and the outlier indicators, we show that the proposed method selects a model of the correct order of dimensionality, under the sparse Riesz condition on the correlation of design variables and a joint sparse Reisz condition on the augmented design matrix. We also show that the estimation/prediction of the selected model can be controlled at a level determined by the sizes of the true model, the outliers and the thresholding level. Moreover, the estimation has a positive breakdown point when both the dimension p and the sample size n tend to infinity, and $p >> n$. We also provide a generalized version for the estimator by adjusting the penalty weight factor. Finally, we apply the proposed method to analyze an aircraft landing performance data set, for identifying the precursors for undesirable landing performance and reducing the risk of runway overruns.

Acknowledgements

I would like to express my deepest thanks to all of the committee members for offering me insightful guidance, valuable comments and support. Without any of them, this dissertation would certainly have not been finished. I enjoyed every meeting, conversation, and discussion with all of them.

I am truly fortunate to have Dr. Regina Y. Liu being my Ph.D advisor, who is not only my best research mentor, but also a best friend, a best teacher, and a best role model in my life. I feel extremely proud when introducing myself as her Ph.D student to other people. Her enormous devotion to statistical research, her prodigious knowledge and expertise, her stunning intelligence, excellent humorousness, and incredible patience have greatly inspired me. She definitely has had the most formative influence on my life in U.S. Before saying thanks to other people, I wish I could list all of her virtues and the spirits that I am learning from her or trying to, with my poor writing skills but a truly grateful heart.

I would also like to extend my deepest gratitude to Dr. Minge Xie and Dr. Cun-Hui Zhang, who have offered me invaluable guidance, brilliant ideas, endless encouragement and overwhelming patience. Not only that they give me hands-on guidance on the topics of my thesis, but they also teach me the way to learn knowledge, to conduct research, and the most important, to think using the brain power of a statistician, or, sometimes, even like a mathematician. Their intelligence and expertise are boundless, and I am honored and deeply indebted to be working with them.

Without Dr. Andrew Cheng, this dissertation would not have existed. His energetic and hard working attitude and enthusiasm for research have inspired me a lot. I am indebted to his many valuable ideas, comments and encouragement. I am grateful for the research opportunity and support from him and the Federal Aviation Administration.

Last but not least, I want to thank the loving faculty and staff members and my fellow Ph.D students in the Rutgers Statistics Department. I will always cherish my days here. The friendship, the research environment and the wonderful lectures and seminars are the biggest gifts a graduate student could have ever had.

Dedication

This dissertation is dedicated to my mother, Minghua Wang and my father, Jianjiang Li, for giving me tremendous and unimaginable love and support, for making my life so joyful and special.

Table of Contents

Chapter 1

Model Configuration and Sparsity Assumptions

1.1 Introduction

There have been a significant amount of methodologies developed for outlier detections and robust estimation of linear regression models since decades ago. However, most of the estimators require full rank for the design matrix, thus they can not adapt to high dimensional data sets which are pervasive nowadays in various research areas. Even though some robust estimation procedures are deployed to analyze large data sets, the dimension of the data is generally restricted to be smaller than the sample size.

Meanwhile, despite the extensive development of model selection procedures, in particular, for high/ultra-high dimensional data sets, such as the class of penalization methods which possess many desirable statistical properties, those methods can be sensitive to outliers and are not robust. For example, if there are a group of outliers which are also leverage points, then the estimation and model selection could be problematic.

There also exist remedies for simultaneous model selection and outlier detection in literatures, see, for example, Hoeting et al [26], Müller-Welsh[50], Gannaz[20], A Khan, van Aelst and Zamar [30], She-Owen [66], and Maronna [44]. However, besides the underlying limitation on the dimension of the data set, there have been also the lacking of robustness properties such as the breakdown point and the loss bounds of the proposed estimators.

In this dissertation, we use the following general linear regression model :

$$\mathbf{y} = \sum_{i=1}^{p} \beta_j \boldsymbol{x_j} + \sqrt{n}\boldsymbol{\gamma} + \boldsymbol{\epsilon} = \boldsymbol{X\beta} + \sqrt{\boldsymbol{n}}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}). \tag{1.1}$$

Here the parameter γ_i is nonzero when the i-th observation is an outlier, and p is the dimension of the data and n is the sample size. This mean-shift model has been widely

adopted in literature, such as in McCann–Welsch [45], and She-Owen [66]. Assume that the column vectors of X are standardized to have equal length $\sqrt{n}$. Thus the factor $\sqrt{n}$ before $\boldsymbol{\gamma}$ is to make the outlier indicator covariates being in the same scales as the X-covariates. This setting follows the generalized setting of the mean shift outlier model, as discussed in [3] and [54].

1.2 Model Settings and Assumptions

Alternatively, we can write the model above into a shorter form:

$$\boldsymbol{y} = \sum_{i=1}^{p+n} \theta_j \boldsymbol{z_j} + \boldsymbol{\epsilon} = \boldsymbol{Z\theta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}), \tag{1.2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ and

$$\mathbf{Z} = \left(\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \sqrt{n} & 0 & \cdots & 0 \\ 0 & \sqrt{n} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sqrt{n} \end{pmatrix} \right). \tag{1.3}$$

Denote the size of $\boldsymbol{\beta}$ by p. The number of parameters in the new model includes $p+n$ parameters. The first p of θ_j's come from the original β_j's, $j = 1, \cdots, p$, and the other n coefficients represent the effects of the outliers. The model has $p+n$ parameters and n observations only. Assume the sparsity assumptions as follows:

$$\|\boldsymbol{\beta}\|_0 = \#\{j \leq p : \beta_j \neq 0\} = d^0, \quad \|\boldsymbol{\gamma}\|_0 = \#\{i \leq n, \gamma_i \neq 0\} = s^0. \tag{1.4}$$

The plan for variable selection is to use the regularization method. We propose a ℓ_1 penalty regression method for this mean shift model, with a tuning parameter λ. The estimator $\widehat{\boldsymbol{\theta}}$ is the minimizer of the loss function

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &\equiv (\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\gamma}}(\lambda)) \\ &\equiv \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\arg\min} \left\{ \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sqrt{n}\boldsymbol{\gamma}\|^2}{2n} + \lambda \sqrt{\frac{\log n}{n}} \|\boldsymbol{\beta}\|_1 + \lambda \sqrt{\frac{\log n}{n \log p}} \|\boldsymbol{\gamma}\|_1 \right\}. \end{aligned} \tag{1.5}$$

Thus the goal of simultaneous outlier detection and variable selection boils down to the identification of the non-zero set for θ's which is a minimizer of the above formula.

Denote the coefficients corresponding to X-covariates selected by LASSO by

$$\widehat{A} \equiv \widehat{A}(\lambda) \equiv \{j \leq p : \widehat{\beta}_j \neq 0\}. \tag{1.6}$$

The outliers identified by the nonzero outlying coefficients are in the set

$$\widehat{S} \equiv \widehat{S}(\lambda) \equiv \{i \leq n : \widehat{\gamma}_i \neq 0\}. \tag{1.7}$$

Thus the overall model selected is

$$\widehat{A} \cup \{\widehat{S} + p\} = \{l : \widehat{\theta}_l \neq 0, l = 1, \cdots p + n\} = \{j, i + p : \widehat{\beta}_j \neq 0,\ \widehat{\gamma}_i \neq 0\}. \tag{1.8}$$

We will write the selection set in the form of $\widehat{A} \cup \widehat{S}$ instead of $\widehat{A} \cup \{\widehat{S} + p\}$ for simplicity unless there is further specification.

1.3 Questions and Goals

Note that the model in (1.2) is simply a weighted LASSO estimator for $\boldsymbol{\theta} \in R^{p+n}$, by penalizing differently on β's and γ's, where the thresholding level on γ's is lower than that of β's, by dividing a factor of $\log p$. There has been extensive discussion in literatures regarding the good properties for LASSO typed estimators. Besides computational efficiency, it is known that LASSO also possess theoretical advantages, such as estimation accuracy, the normal convergence of the estimators, rate consistency, and selection consistency with regularity conditions. As long as the corresponding regularity conditions are satisfied, all the theoretical results from LASSO are inherited by $\widehat{\boldsymbol{\theta}}$. However, two critical issues need to be addressed due to the characteristics of "simultaneity".

Question 1. Since we may not have any information about the outliers, and by the model setting, the outlying coefficient can be unique. An outlier only occurs in one observation. This is different from X-covarites, where each non-zero coefficient contributes to all the y_i's. Mathematically, this "one coefficient in only one equation" fact leads to the special setting of our design matrix, Z. Is it likely that the augmented new design matrix Z will miss the good properties due to the difficulty to meet the regularity conditions?

Question 2. Sparsity constraints are assumed for both blocks of Z: the parsimony of the underlying relationship between y and X, and the contamination size of the data

set, i.e., the portion of the outliers, which could be a fixed fraction no matter how large the size of n or p is. Thus the model dimension (greater than the size of outliers) could have order up to $O(n)$. Thus a natural question is, will this large model size breaks down the classical LASSO results? Up to what fraction of contamination can we allow in the data? This is equivalent to answering what the breakdown point of the proposed estimator would be.

The goal of this dissertation is to address the two questions above. Specifically, we

1. first, show that the estimators in (1.5) for this mean shift model yields similar properties on the new augmented design matrix $\mathbf{Z}$, in terms of studying the risk bounds for estimation/prediction and the rate of false discovered coefficients;

2. second, calculate its breakdown point of, i.e., the maximum contamination that can be allowed given that the estimation will not go to infinity, or alternatively, the maximum contamination can be allowed while the regularity conditions are not violated and the desirable properties established for our proposed estimators remain valid.

The organization of the dissertation is as follows: after a brief survey of literature review in section 2 for both model selection part and robust estimation on linear regression part, section 3 presents the main theorem regarding the rate consistency and risk bounds. The regularity conditions for these results are discussed in section 4 . Section 5 provides the breakdown point property and develop a generalized version for the proposed method. The optimality of the proposed method is discussed. Section 6 briefly introduces an iterative algorithm for computing efficiency improvement. After a simulation in section 7, we will apply our method to the aircraft landing data analysis in section 8, to identify the precursors for undesirable landings and to detect the potential runway overruns . The proof and lemmas are presented in section 9.

Chapter 2

Literature Review

This section gives a brief literature review on the development of robust estimation and outlier detection for linear regression model, the model selection in high dimension setting, and the simultaneous model selection and outlier detection.

2.1 Robust Estimation and Outlier Detection in Linear Regression

In multivariate linear regression, a data point which deviates from the linear pattern is called an outlier. Hampel et al. [22] write that 10% of outliers is quite common. Many literatures typically adopt $\epsilon = 0.25n$ as an upper bound for contamination, such as Hubert [27] and McCann–Welsh[45]. If the data is contaminated, then the multivariate estimates for the regression model would differ from the true parameters, or the estimates from the dataset without outliers. Meanwhile, the outlier detection diagnostics based on the model fitting would be problematic. Both masking and swamping effects would occur, where "masking" means the outliers would mask each other from being undetected and "swamping" refers to mistakenly detected outliers which are in fact regular observations.

A robust estimation method would help to raise the chance of detecting outliers correctly, and the accuracy of outlier detection would improve the robustness of the estimations. For robust estimators for regression, we provide a brief list here, including M-estimators [24] [25], generalized M-estimators [69], [10], R-esitmators[28], S-estimators [62], [12], MM-estimators [77], τ-esitmators, [40], CM-esitmators [29], L-estimators [32], least median of squares (LMS) in Rousseeuw [57], Rousseeuw and Hubert [59], least trimmed squares (LTS) and related algorithms [59], Rousseeuw and Van Driessen [64]. LTS, LMS and MM estimators are one of the first high-breakdown regression methods.

However, it has been shown that the convergence rate is slow and the asymptotic efficiency is zero. In contrast, LTS is asymptotically normal and can be computed much faster.

In multivariate data analysis, outlier detection techniques have also been greatly progressed, starting from minimum covariance determinant estimator (MCD) proposed by Rousseeuw [57] [58] with the efficient algorithm developed in Roussseeuw and Van Driessen [63], and minimum volume ellipsoid (MVE) estimators via Mahalanobis distances in Rousseeuw [57] [58]. Efficient estimators for both MCD and MVE are generated by Lopuhaä–Rousseeuw [42] and Lopuhaä [41]. Butler, Davies and Jhun[6] and Davies [**?**] study the asymptotic results for both MCD and MVE.

Nonparametric methods also plays an important role in outlier detection. An important example is data depth, the notion for the measurement of outlyingness of a data cloud. To name a few, estimators based on different notions of depth are proposed in Rousseeuw, Ruts,and Tukey [61], Liu [34] [38] [36] [39], Liu, Parelius and Singh [37], and Zuo and Serfling [86] [87] etc. These methods work very well for low diminutional datasets. The developed algorithm could also be time consuming when sample size is too large.

Diagnostics methods such as Weisberg's [75] leave-one-out approach is another popular way to determine the outlyingness for each datapoint in multivariate data analysis. It calculates the difference of a specific statistic of interest after a single observation is deleted. Atkinson and Riani [3], Atkinson et al., [4], and Riani, Atkinson and Cerioli [56] proposed forward search and producing series of plots for diagnostics for outlier identification. Similarly, a backward search is proposed by Menjoge and Welsch [47]. However, when there exists a group of outliers, the statistic of interests could be contaminated and this approach would fail.

A lot of methods mentioned above a robust initial start with high breakdown, and then updates the estimators using iterations. There are two challenges here: first, the breakdown point is not high many methods Typically the data used for simulation or real study datasets are low ranked. Bootstrap or cross validation methods may also be involved for updating stages, which is impractical for high dimension setting, due to

the "curse of dimension".

2.2 Variable Selection in High Dimension Datasets

Variable selection is fundamental in statistical research field. The earliest and methods include such as Akaike information criterion(AIC) [2] , Mallows' C_p [43], Bayesian information criterion (BIC) [65] and data-driven methods. These methods aim to select a best subset of the covariates which yields a minimized value for a pre- specified loss function. For dataset with large dimension, best- subset methods are not feasible due to computation efficiency. LASSO [67] is a successful penalized method for high dimensional dataset with good computing efficiency [52] [53] [13]. Its ℓ_1 penalty function can provide a continuous solution path and yield a sparse output model.

Under a strong irrepresentable condition proposed in [49] Meinshausen–Buhlmann, Tropp [68], Zhao and Yu [82] and Wainwright [73] proved that the LASSO is variable selection consistent. However, the strong irrepresentable condition is quite restrictive for moderately large size of the true model. Under sparse Riesz condition (SRC) on the ℓ_2 norm of sub-Gram matrices, Zhang–Huang [80] proved that the dimension for the LASSO selection is of the same order as the size of the true model.

There have been a great number of LASSO-typed or similar estimators using convex penalization functions, such as Elastic net [83], which is combination of ℓ_1 and ℓ_2 penalties, Adaptive LASSO which minimizes a weighted ℓ_1 loss function [84], Huang, Ma and Zhang [23] and Zou and Li [85], and Dantzig selector, the [9], Efron, Hastie and Tibshirani [14], and Meinshausen, Rocha and Yu [48].

Two of the main limitations of convex penalization functions are the estimation bias and the restrictive assumptions on model selection consistency. Remedies using concave penalized functions are proposed, such as SCAD [19], MCP [78], and capped ℓ_1 penalty [79]. These non-convex methods can remove the bias of estimation and yields elegant oracle inequalities Buhlmann and van deGeer [5], Kim, Choi and Oh [31] and Zhang [78]. However, it is hard to answer the questions such as whether the solution is global, and if not, then what would the relationship between the local and global solutions be.

Meanwhile, the oracle inequalities for high dimensional settings may not be as optimal as in lower dimension cases.

A lot of the LASSO-typed or LASSO-like estimators require ℓ_2 regularity conditions. They may have been discussed into different forms in the literature. A most well known example is the restricted isometry condition (RIP) introduced by Candés and Tao [8]. A related condition is the uniform uncertainty principle (UUP). Optimal error bounds for $\|\widehat{\beta} - \beta\|_2$ can be established with RIP and UUP [8] and Cai, Wang, Xu [11]. As for the generalized sparsity assumption, SRC conditions are also studied in Zhang and Huang [80] and Ye and Zhang [76].

A similar type of conditions include restricted eigenvalues as seen in Biskel, Ritov and Tsybakov [7], and Koltchinskii [33], and compatibility factor in [70] [72] can be viewed as a modified ℓ_2 regularity conditions. A weaker version RIF_q, the restricted invertibility factor [19] [76], [81], together with its sign restricted version [76] are found to be able to generate optimal order [55] and sharpen previous results.

2.3 Simultaneous Model Selection and Outlier Detection

There is increasing but limited literatures on simultaneous model selection and outlier detection in high dimensional setting. When the number of parameters are larger than the sample size, many classical robustified procedures can not be applied directly. Thus one possible remedy is to use multivariate robust estimator to replace the non-robust estimator. One example is the robust LARS [13] proposed by A Khan, van Aelst and Zamar [30]. At each selection step, a robust correlation matrix is calculated, for example, using the bivariate Winsorization method to shrunk data points towards the data bulk. Thus the correlation matrix is more robust against certain outliers. However, this "robustness" is only for the correlation matrix needed for the algorithm. The robustness for parameter estimation is not guaranteed.

Another direction of remedies is to modify the regression mold itself, incorporating the coefficient of outlyingness. A wavelet thresholding model, but with no sparsity assumptions on the X-covariates is adopted in Gannaz [20]. She–Owen [66] considered

the outlier detection problem together with variable selection with setting of a mean-shift model, in which observations may have different intercept terms. The sparsity for both outlier indicator variables and the covariates are assumed. The method proposed by She and Owen is to use a nonconvex penalty function, a hybrid of hard-thresholding and ridge regression method, with two tuning parameters for each part. Besides ℓ_1 penalized regression methods, Robust ridge regression for high-dimensional dataset is also studied in Maronna [44]. Wang and Li [74] also apply an Wilcoxon-typed smoothly clipped absolute deviation method as a non convex remedy for simultaneous variable selection and outlier detection. All these methods need data driven methods for the selection of multiple tuning parameters. And a few common limitations includes the unknown breakdown point, the lack of loss bound of estimation and/or the bound of false discovery.

Chapter 3

Loss bound in LASSO on Outlier Detection Model

3.1 Assumptions

In the setting of this dissertation, we assume the dataset $\mathbf{X}$ is drawn from random design experiments. Suppose that the n rows of the random matrix $X_{n\times p}$ are i.i.d. copies of a random vector drawn from multivariate Gaussian distributions with Σ being the covariance. And suppose the sequence of covariates x_j's satisfies the Reisz condition, i.e., if there exist fixed $0 < \rho_* < \rho^* < \infty$ such that

$$\rho_* \sum_{j=1}^{p} b_j^2 \leq E\left(\sum_{j=1}^{p} b_j \xi_j\right)^2 \leq \rho^* \sum_{j=1}^{p} b_j^2 \tag{3.1}$$

for all constants b_j.

Recall that a design matrix $\mathbf{X}$ satisfies the sparse Riesz condition (SRC) with rank q^* and spectrum bounds $0 < c_*(q^*) < c^*(q^*) < \infty$, if

$$c_*(q^*) \leq \frac{\|\boldsymbol{X_A v}\|^2}{\boldsymbol{n}\|\boldsymbol{v}\|^2} \leq c^*(q^*), \quad \forall A \text{ with } |A| = q^* \text{ and } v \in R^{q^*} \tag{3.2}$$

Essentially SRC imposes ℓ_2 regularity conditions on the design matrix, which plays a very important role to control the risk bound for LASSO and similar types of penalized regression estimators. This has been discussed in many literatures, e.g., as seen in [80] and [78]. And it has been shown to hold with large probability approaching to 1 when n is large, when the observed data is drawn from experiments of random matrices with Reisz condition and Gaussian design. Similarly, in this section, we will introduce a modified condition, so called joint SRC, which will hold with large probability as well, with detailed configurations in section 5 and proof shown in the appendix. This condition provides the spectral norm bound for the design matrix Z. Rather than extracting any q^* columns in X, we turn to extract an arbitrary submatrix from X with

q_1^* columns and extract an arbitrary sub-matrix from identify matrix multiplied by $\sqrt{n}$ with any q_2^* columns.

Definition 3.1.1. *We say that the design matrix* $\mathbf{Z} = \left(\mathbf{X}|\sqrt{\mathbf{n}}\mathbf{I_n}\right)$ *satisfies the joint sparse Riesz condition (SRC) with pair rank* (d^*, s^*) *and spectrum bounds* $0 < c_*(d^*, s^*) < c^*(d^*, s^*) < \infty$*, if for any* $A \subseteq \{1, \cdots, p\}$ *with* $|A| = d^*$*,* $S \subseteq \{1, \cdots, n\}$ *with* $|S| = s^*$ *and* $v \in R^{q^*}$*,*

$$c_*(d^*, s^*) \leq \frac{\|\boldsymbol{Z_{A\cup S}}\|^2}{\boldsymbol{n}\|\boldsymbol{v}\|^2} \leq c^*(d^*, s^*). \tag{3.3}$$

Before stating the main results of this dissertation, we will present some definitions and notations. Let

$$B \equiv \{j : \beta_j \neq 0, 1 \leq j \leq p\}, \quad T \equiv \{i : \gamma_i \neq 0, 1 \leq i \leq n\}.$$

For any $\mathbf{v} \in R^n$, let

$$\zeta_\beta(\mathbf{v}; m_1, m_2, B, T) \equiv \max_{A,S} \left\{ \frac{\|(P_{A\cup S} - P_{B\cup S})\mathbf{y}\|}{(m_1 n)^{1/2}} : B \subseteq A \subseteq \{1, \cdots, p\}, \right.$$
$$\left. |A| = m_1 + |B|, T \subseteq S \subseteq \{1, \cdots, n\}, |S| = m_2 + |T| \right\}, \tag{3.4}$$

and

$$\zeta_\gamma(\mathbf{v}; m_2, B, T) \equiv \max_{S} \left\{ \frac{\|(P_{B\cup S} - P_{B\cup T})\mathbf{y}\|}{(m_2 n)^{1/2}} : T \subseteq S \subseteq \{1, \cdots, n\}, |S| = m_2 + |T| \right\}, \tag{3.5}$$

where the $P_{M_1\cup M_2}$ is the projection matrix from $\mathbb{R}^n$ to $\mathbb{R}^{|M_1|+|M_2|}$, the linear span generated by column vectors $\{\mathbf{x}_j,\ j \in M_1 \subseteq \{1, \cdots, p\}\}$ and $\{\sqrt{n}\mathbf{e}_i, i \in M_2 \subseteq \{1, \cdots, n\}\}$. Here $\mathbf{e}_i$ is the i^{th} column vector of the $n \times n$ identity matrix. For example, the projection matrix $P_{B\cup T}$ is calculated by

$$P_{B\cup T} \equiv Z_{B\cup T}(Z'_{B\cup T} Z_{B\cup T})^{-1} Z'_{B\cup T}. \tag{3.6}$$

We denote $\widehat{\boldsymbol{\beta}}^o$ is the oracle estimator for $\boldsymbol{\beta}$, and $\widehat{\boldsymbol{\beta}}^o_B = \{b_j, j \in B\}$, and it is calculated based on the formula below:

$$\widehat{\boldsymbol{\theta}}^o = (\widehat{\boldsymbol{\beta}}^o, \widehat{\boldsymbol{\gamma}}^o) = \underset{\boldsymbol{\beta},\ \boldsymbol{\gamma}}{\arg\min} \left\{ \|y - X\boldsymbol{\beta} - \sqrt{n}\boldsymbol{\gamma}\|^2 : \widehat{\boldsymbol{\beta}}_j = \widehat{\boldsymbol{\gamma}}_i = 0, j \in B, i \in T \right\}. \tag{3.7}$$

Thus

$$\widehat{\boldsymbol{\beta}}^o_B = (X'_{T^c,B}X_{T^c,B})^{-1}X'_{T^c,B}\mathbf{y}_{T^c}, \tag{3.8}$$

$$\widehat{\boldsymbol{\gamma}}^o_T = \mathbf{0}, \quad \widehat{\boldsymbol{\gamma}}^o_{T^c} = \mathbf{y}_{T^c} - X_{T^c,B}\widehat{\boldsymbol{\beta}}^o_B. \tag{3.9}$$

Let $d^0 = |B|$, and $s^0 = |T|$. Let $\widetilde{p}_{\epsilon,\beta} \geq \sqrt{e}$ be the solution of (3.11) and $\widetilde{p}_{\epsilon,\gamma} \geq \sqrt{p}$ be the solution of

$$2\log\widetilde{p}_{\epsilon,\beta} - 1 - \log(2\log\widetilde{p}_{\epsilon,\beta}) \tag{3.10}$$

$$= \quad (2/m_1)\left\{\log\binom{p-d^0}{m_1} + \log\binom{n-s^o}{m_2} + \log(1/\epsilon)\right\}, \tag{3.11}$$

and

$$\frac{2\log\widetilde{p}_{\epsilon,\gamma}}{\log p} - 1 - \log\left(\frac{2\log\widetilde{p}_{\epsilon,\gamma}}{\log p}\right) = (2/m_2)\left\{\log\binom{n-s^o}{m_2} + \log(1/\epsilon)\right\} \tag{3.12}$$

respectively, for nonnegative integers $m_1 \in [1, p-d^0]$, $m_2 \in [1, n-s^0]$ and a real number $\epsilon \in (0,1]$. Define the following constants: $\widetilde{p}^*_{\epsilon,\beta}$ with $m_1 = m_1^*$, $m_2 = m_2^*$, $\widetilde{p}^{\,1}_{\epsilon,\beta}$ with $m_1 = 1$ and $m_2 = m_2^*$, and $\widetilde{p}^*_{\epsilon,\gamma}$ with $m_2 = m_2^*$. Define

$$\lambda_{\epsilon,\beta} \equiv \frac{\sqrt{r^*}}{\alpha}\sigma\left(\sqrt{2\log\widetilde{p}^*_{\epsilon,\beta}/\log n} \vee \sqrt{2\log\widetilde{p}^{\,1}_{\epsilon,\beta}/\log n}\right), \tag{3.13}$$

and

$$\lambda_{\epsilon,\gamma} \equiv \frac{\sqrt{r^*}}{\alpha}\left(\sigma\sqrt{2\log\widetilde{p}^*_{\epsilon,\gamma}}\right). \tag{3.14}$$

3.2 Main Results

We now proceed to the main theorem with two assumptions in (3.2) and (3.3) as follows:

(1) Suppose SRC in (3.2) holds for $\mathbf{X}$ with certain d^* and $c^* \geq c_* > 0$. For any subset $A \subset \{1, \cdots, p\}$,

$$c_* \leq \min_{|A|\leq d^*} c_{\min}(\Sigma_A) \leq \max_{|A|\leq d^*} c_{\max}(\Sigma_A) \leq c^*.$$

(2) Suppose joint SRC in (3.3) holds for $\mathbf{Z}$ with pair rank (d^*, s^*). For any subset $A \subset \{1, \cdots, p\}$ and any subset $S \subset \{1, \cdots, n\}$, we have

$$r_* \leq \min_{|A|\leq d^*, |S|\leq s^*} c_{\min}(\Sigma_{A\cup S}) \leq \max_{|A|\leq d^*, |S|\leq s^*} c_{\max}(\Sigma_{A\cup S}) \leq r^*,$$

where $s^* < n$ and $\lim_{n\to\infty} s^*/n \to a_0 < 1$.

The following is the main theorem of this dissertation.

Theorem 3.2.1. *Let B be a deterministic set of $\{1, \cdots, p\}$ and T be a deterministic set of $\{1, \cdots, n\}$. Let $\widehat{\theta}$ be the LASSO estimator in (1.5), and $\widehat{\theta}^o = (\widehat{\beta}^o, \widehat{\gamma}^o)$ be the oracle estimator as defined in (3.7). Let*

$$m_1^* = d^* - d^0,\ m_2^* = s^* - s^0,\ 0 < \alpha \leq 2/3.$$

Define

$$K_* \equiv \min_{w \geq 0} K_{*,w} = \min_{w \geq 0} \frac{((1-\alpha)(1+w) + w\alpha)r^*/r_* - (1-\alpha)}{(1-\alpha)(2 - 3\alpha + w - 2w\alpha^2)}. \tag{3.15}$$

Assume $|T| = s^0 < s^*$. Suppose

$$|B| = d^0 \leq d^*, \quad |T| = s^0 \leq s^*, \quad (K_* + 1)(|B| + \frac{|T|}{\log p}) \leq d^* \wedge \frac{s^*}{\log p} \tag{3.16}$$

Let $\lambda_{\epsilon,\beta}$ and $\lambda_{\epsilon,\gamma}$ be as in (3.13) and (3.14), and let

$$\lambda_0 \equiv \lambda_{\epsilon,\beta} \vee \lambda_{\epsilon,\gamma}. \tag{3.17}$$

Then for given constant $\epsilon < 1/\sqrt{2}$ and for any $\lambda \geq \lambda_0$, with probability at least $1 - \sqrt{2}\epsilon$, one has

$$\Big(\#\{j \notin B : \widehat{\beta}_j \neq 0\}\Big) \log p + (\#\{i \notin T : \widehat{\gamma}_i \neq 0\}) \\ < 1 \vee \big(K_*(|B| \log p + |T|)\big), \tag{3.18}$$

and

$$\begin{aligned} r_* \|\widehat{\theta} - \widehat{\theta}^*\| &\leq \|Z(\widehat{\theta} - \widehat{\theta}^o)\|/\sqrt{n} \\ &\leq \lambda\sqrt{(\log n)/n}\big(1 + \alpha\sqrt{2K_* r_*/r^*}\big)(\sqrt{|B|} + \sqrt{|T|/\log p}). \end{aligned} \tag{3.19}$$

The first conclusion in the theorem provides the bound for the false selected variables and outliers, and the weighted sum of both. The weights are $\log p$ for β and 1 for γ. The bound is proportional to the weighted sum of the true sets , if the order of $\log p$ is close to $O(n)$, then $|T|/\log p$ will be a small constant. Thus we know that the model selected (the non-zero $\widehat{\beta}'s$) thus is in the correct order of dimension. Explanation with more details on the estimation/prediction accuracy part will be given in next section.

3.3 The Orders of Penalty Levels

The second conclusion (3.19) in the theorem gives the loss bound of LASSO estimation/prediction in ℓ_2 format: $\|\widehat{\theta} - \widehat{\theta}^o\|$ and $\|Z\widehat{\theta} - Z\widehat{\theta}^o\|/\sqrt{n}$. Both terms are bounded by $\lambda\sqrt{\log n/n}(\sqrt{|B|} + \sqrt{|T|/\log p})$ up to a multiple. Thus we need to check the order of $\lambda\sqrt{\log n/n}$, which is also the penalty level for β. According to the assumption in the theorem, λ is greater than $\lambda_{\epsilon,\beta}$ and $\lambda_{\epsilon,\gamma}$. The orders for both thresholding values will be checked:

- By definitions for $\lambda_{\epsilon,\beta}$, its order can be calculated by Sterling's formula.

$$\log \widetilde{p}_{\epsilon,\beta} \approx \frac{1}{m_1}\left(\log\binom{p-d^0}{m_1} + \log\binom{n-s^0}{m_2}\right) \tag{3.20}$$

  thus we have for $m_1 = m_1^*$, $\log \widetilde{p}^*_{\epsilon,\beta} \asymp \log p + n/m_1^*$. For m_2 is a small integer, $\log \widetilde{p}^{\,1}_{\epsilon,\beta} \asymp \log p + n$. By (3.13),

$$\lambda_{\epsilon,\beta} = (\sqrt{2r^*}\sigma/\alpha)\left(\sqrt{\log p/\log n} \vee \sqrt{n/\log n}\right).$$

  Let C_0 be a constant determined by r^*, σ, and α only. Thus the penalty levels in the proposed model (1.5) are:

$$\text{for } \beta : \lambda_{\epsilon,\beta}\sqrt{(\log n)/n} = C_0(\sqrt{\log p/n}) \vee C_0, \tag{3.21}$$

$$\text{for } \gamma : \lambda_{\epsilon,\beta}\sqrt{(\log n)/(n\log p)} = C_0(n^{-1/2}) \vee C_0\big((\log p)^{-1/2}\big) \tag{3.22}$$

- Similarly, for $\lambda_{\epsilon,\gamma}$, by (3.12) and (3.14), we have

$$\log \widetilde{p}^*_{\epsilon,\gamma} = C_1^2 \log p, \quad \lambda_{\epsilon,\gamma} = C_1\sqrt{\log p}.$$

  Therefore, the penalty levels are:

$$\text{for } \beta : \lambda_{\epsilon,\gamma}\sqrt{(\log n)/n} = C_1\sqrt{(\log n)(\log p)/n}, \tag{3.23}$$

$$\text{for } \gamma : \lambda_{\epsilon,\gamma}\sqrt{(\log n)/(n\log p)} = C_1\sqrt{\log n/n}. \tag{3.24}$$

To summarize, the penalty level for γ_i is at most in the order $O(\sqrt{(\log n)/n})$. The largest order of penalty level on the parameter β_j is $O(\sqrt{(\log n)(\log p)/n})$. Recall the

oracle properties, to name a few, such as [70] and [78]) in regular settings, where the outliers are not present. The corresponding loss has ℓ_2 norm in the order of $\sqrt{\log p/n}|B|^{1/2}$. This implies the inflation factor of the loss here is $\sqrt{\log n}$. In the worst case with $\log p = \alpha_p n$, the penalty level for β is $O(\sqrt{\log n})$. This inflation ratio increases very slowly when $n \to \infty$. For example, when $n = 10^{20}$, $\sqrt{\log n} < 7$. This means that β's penalty level can stay low even when n is in millions or billions.

We will prove (3.3) after presenting the corollaries and lemmas. The proof of (3.19) will be provided in Appendix, which uses the notations and proof for Lemma 3.3.2.

As a direct result from Theorem , we obtain Corollary 3.3.1 as follows:

Corollary 3.3.1. *Assume the notations in Theorem 3.2.1. Assume that there exists a constant $c_p^* > 0$ such that $|T|/(|B|\log p) \le c_p^*$. Then for the same λ defined as above, and any given constant $\epsilon < 1/\sqrt{2}$, with probability at least $1 - \sqrt{2}\epsilon$, one has*

$$\Big(\#\{j \notin B : \widehat{\beta}_j \neq 0\}\Big) < 1 \vee \big(K_*(1 + c_p^*)|B|\big), \tag{3.25}$$

and

$$\begin{aligned} & r_*\|\widehat{\theta} - \widehat{\theta}^*\| \le \|Z(\widehat{\theta} - \widehat{\theta}^o)\|/\sqrt{n} \\ \le \quad & \lambda\sqrt{(\log n)/n}(1 + c_p^*)\big(1 + \alpha\sqrt{2K_* r_*/r^*}\big)|B|^{1/2}. \end{aligned} \tag{3.26}$$

Lemma 3.3.2. *Suppose SRC holds for $\mathbf{X}$ with certain d^* and $c^* \ge c_* > 0$, and suppose joint SRC also holds for $\mathbf{Z}$ with subset $A \subset \{1, \cdots, p\}$ and subset $S \subset \{1, \cdots, n\}$,. Let $\lambda > 0$, $\alpha \in (0, 2/3]$, and let K_* be defined as in (3.15). Suppose there exist two sets B and T satistfying: $B \subset \{1, \cdots, p\}$ with $|B| = d^0 \le d^*$, and $T \subset \{1, \cdots, n\}$ with $|T| = s^0 \le s^*$, and*

$$(1 + K_*)(|B| + |T|/\log p) \le d^* + s^*/\log p.$$

Let m_1 and m_2 be fixed integers satisfying $1 \le m_1 \le d^* - |B|$, $1 \le m_2 \le s^* - |T|$, and suppose $\mathbf{y} \in \mathbb{R}^n$ with

$$(\sqrt{c^*}/\alpha)\zeta_\beta(\mathbf{y}; m_1, m_2, B, T) \le \lambda\sqrt{\frac{\log n}{n}}, \tag{3.27}$$

$$(\sqrt{c^*}/\alpha)\zeta_\gamma(\mathbf{y}; m_2, B, T) \le \lambda\sqrt{\frac{\log n}{n\log p}}, \tag{3.28}$$

where $\zeta_\beta(\mathbf{y}; m_1, B, T)$ and $\zeta_\gamma(\mathbf{y}; m_2, A, T)$ are defined in (3.4) and (3.5) respectively. Let $\widehat{\beta}$ and $\widehat{\gamma}$ be the solution with λ in (1.5). Let A_1 and S_1 be the sets satisfying

$$B \cup \{j : \widehat{\beta}_j \neq 0\} \subseteq A_1 \subseteq B \cup \left\{ |\mathbf{x}_j'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \sqrt{n}\widehat{\boldsymbol{\gamma}})|/n = \lambda\sqrt{\frac{\log n}{n}} \right\}, \tag{3.29}$$

$$T \cup \{i : \widehat{\boldsymbol{\gamma}}_j \neq 0\} \subseteq S_1 \subseteq T \cup \left\{ |\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}} - \sqrt{n}\widehat{\boldsymbol{\gamma}}_i|/\sqrt{n} = \lambda\sqrt{\frac{\log n}{n \log p}} \right\}. \tag{3.30}$$

If

$$|A_1| = |B| + m_1,\ |A_1| \leq d^*, |S_1| = |T| + m_2,\ |S_1| \leq s^*,$$

then

$$|A_1| - |B| + \frac{(|S_1| - |T|)}{\log p} \leq K_* \left(|B| + \frac{|T|}{\log p} \right), \tag{3.31}$$

and

$$r_* \|\widehat{\theta} - \widehat{\theta}^*\| \leq \|Z(\widehat{\theta} - \widehat{\theta}^o)\|/\sqrt{n} \leq \lambda \left(1 + \alpha\sqrt{2K_* r_*/r^*}\right) (\sqrt{|B|} + \sqrt{|T|/\log p}), \tag{3.32}$$

where $\widehat{\theta}^o = (\widehat{\beta}^o, \widehat{\gamma}^o)$ is the oracle estimator as defined in (3.7).

Lemma 3.3.3. *Let $\zeta_\beta(\mathbf{v}; m_1, m_2, B, T)$ and $\zeta_\gamma(\mathbf{v}; m_2, B, T)$ be as in (3.4) and (3.5) with deterministic m_1, m_2, B and T. Let $d^0 = |B|$ and $s^0 = |T|$. Suppose $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$, and $\log \widetilde{p}_{\epsilon,\beta}$ and $\log \widetilde{p}_{\epsilon,\gamma}$ are as in (3.11) and (3.12). We have*

$$P\left\{ \zeta_\beta(\boldsymbol{\varepsilon}; m_1, m_2, B, T) \geq \sigma\sqrt{(2/n)\log \widetilde{p}_{\epsilon,\beta}} \text{ or } \right.$$
$$\left. \zeta_\gamma(\boldsymbol{\varepsilon}; m_2, B, T) \geq \sigma\sqrt{\frac{2\log \widetilde{p}_{\epsilon,\gamma}}{n \log p}} \right\} \leq \sqrt{2}\epsilon. \tag{3.33}$$

Lemma 3.3.4. *Assume the notations in Lemma 3.3.3, and that $\lambda_{\epsilon,\beta}$ and $\lambda_{\epsilon,\gamma}$ are as in (3.13) and (3.14). If $\lambda \geq \lambda_{\epsilon,\beta} \vee \lambda_{\epsilon,\gamma}$, then*

$$P\left\{ \sqrt{\frac{\log n}{n}}\lambda < \frac{\sqrt{c^*}}{\alpha} \left(\zeta_\beta(\mathbf{y}; m_1, m_2, B, T) \vee \sqrt{\log p}\ \zeta_\gamma(\mathbf{y}; m_2, B, T) \right) \right\} \leq \sqrt{2}\epsilon \tag{3.34}$$

for any $\epsilon \in (0, \sqrt{1/2})$, $m_1 \leq m_1^* \equiv d^* - d^0$ and $m_2 \leq m_2^* \equiv s^* - s^0$.

3.4 Proof of Theorem 3.2.1.

Proof. Define two sets below and their sizes as

$$s_1^{(\lambda)} \equiv \#\left\{T \cup i \notin T : |y_i - \mathbf{x}_i\widehat{\beta} - \sqrt{n}\widehat{\gamma}_i| = \lambda\sqrt{(\log n)/\log p}\right\} \tag{3.35}$$

$$d_1^{(\lambda)} \equiv \#\left\{B \cup j \notin B : |x_j'(y - X\widehat{\beta} - \sqrt{n}\gamma)|/n = \lambda\sqrt{(\log n)/n}\right\}. \tag{3.36}$$

If we can show that

$$s_1^{(\lambda)} \leq s^*, \quad d_1^{(\lambda)} \leq d^* \tag{3.37}$$

, then the statement of (3.3) and (3.19) is a direct result by applying Lemma 3.3.2, Lemma 3.3.3 and Lemma 3.3.4.

When $\lambda \to \infty$, both $d_1^{(\lambda)}$ and $s_1^{(\lambda)}$ are small, thus they are bounded by d^* and s^* respectively. When the λ increases, the both $d_1^{(\lambda)}$ and $s_1^{(\lambda)}$ increase as well. Then if for some $\lambda_1 \geq \lambda_0$, either $d_1^{(\lambda)}$ is greater than d^* or $s_1^{(\lambda)}$ is greater than s^*, or both of them are greater. Suppose we can add variables one at a time. This means that

$$d_1^{(\lambda_2)} + s_1^{(\lambda_2)}/\log p \geq d_1^{(\lambda_2)} \wedge s_1^{(\lambda_2)}/\log p.$$

In case (3.37) is violated, then

$$d_1^{(\lambda_2)} + s_1^{(\lambda_2)}/\log p \geq d^* \wedge s^*/\log p \tag{3.38}$$

Thus there must exist a $\lambda_2 \geq \lambda_1$, such that

$$d_1^{(\lambda_2)} + s_1^{(\lambda_2)}/\log p \approx d_1^{(\lambda_2)} \wedge s_1^{(\lambda_2)}/\log p,$$

where the sign "$\approx$" means either equal or slightly larger than, but with a small about of difference up to 1.

This yields that $d_1^{(\lambda_2)} \leq d^*$ and $s_1^{(\lambda_2)} \leq s^*$. Let $m_1 = d_1^{(\lambda_2)} - d^0$ and $m_2 = s_1^{(\lambda_2)} - s^0$, then $m_1 \leq m_1^*$ and $m_2 \leq m_2^*$. Therefore by applying lemma 3.3.2, lemma 3.3.4, and (3.35) in step 1, it follows that

$$|A_1^{(\lambda_2)}| + |S_1^{(\lambda_2)}|/\log p < (1 + K_*)(|B| + |T|/\log p) \leq d^* \wedge s^*/\log p.$$

This leads to $d^* + s^* \log p \geq |A_1^{(\lambda_2)}| + |S_1^{(\lambda_2)}|/\log p < d^* \wedge s^*/\log p$. Therefore, for $\lambda \geq \lambda_0$, $|A_1^{\lambda}| < d^*$ and $S_1^{\lambda}| < s^*$.

Now all the conditions for Lemma 3.3.2 are satisfied, so we can applying the results in Lemma 3.3.2 to 3.3.3. With carefully chosen λ as described in the theorem, it follows that

$$\begin{aligned} & \#\{j \notin B : \widehat{\beta}_j \neq 0\} + \big(\#\{i \notin T : \widehat{\gamma}_i \neq 0\}\big)/\log p \\ < \quad & 1 \vee \big(K_*(|B| + |T|/\log p)\big), \end{aligned}$$

and

$$\begin{aligned} r_*\|\widehat{\theta} - \widehat{\theta}^*\| &\leq \|Z'(\widehat{\theta} - \widehat{\theta}^o)\|/\sqrt{n} \\ &\leq \lambda\big(1/\sqrt{r_*} + \alpha\sqrt{2K_*/r^*}\big)\sqrt{|B| + |T|/\log p}. \end{aligned}$$

□

Chapter 4

Sufficient conditions on SRC/ Joint SRC for Design Matrices in Mean Shift Model

Before stating the main results of this section, we will define the following quantities which are determined by a constant $a_0 \in [0, 1)$ only. These quantities are relevant to trimmed distribution of a standard normal variable, including trimmed means, trimmed standard deviations, and quantities developed by previous two. These quantities are used to derive the inequalities for spectrum bounds of design matrix $\mathbf{Z}$ in our model.

Denote

$$\begin{aligned}
\mu_{a_0,-} &\equiv \frac{1}{1-a_0}\int_0^{1-a_0} Q(x)dx, && (4.1)\\
\sigma^2_{a_0,-} &\equiv \frac{1}{1-a_0}\int_0^{1-a_0} Q^2(x)dx - (\mu_{a_0,-})^2, && (4.2)\\
\mu_{a_0,+} &\equiv \frac{1}{1-a_0}\int_{a_0}^{1} Q(x)dx, && (4.3)\\
\sigma^2_{a_0,+} &\equiv \frac{1}{1-a_0}\int_{a_0}^{1} Q^2(x)dx - (\mu_{a_0,+})^2, && (4.4)\\
t_{a_0} &\equiv \frac{(\sigma_{a_0,+}\mu_{a_0,-} + \sigma_{a_0,-}\mu_{a_0,+})(\rho^* - \rho_*/\sigma_{a_0,+})}{\sigma_{a_0,-}(\rho^*\sigma_{a_0,+} + 3\rho_*\sigma_{a_0,-})}, && (4.5)\\
c^*_+(a_0) &\equiv \frac{\sigma_{a_0,+}}{(1-a_0)(\sigma_{a_0,+}\mu_{a_0,-} + \sigma_{a_0,-}\mu_{a_0,+})}, && (4.6)\\
c^*_-(a_0) &\equiv \frac{\sigma_{a_0,-}}{(1-a_0)(\sigma_{a_0,+}\mu_{a_0,-} + \sigma_{a_0,-}\mu_{a_0,+})}. && (4.7)
\end{aligned}$$

Here $Q(x)$ is the inverse function or quantile function of a random variable with χ^2_1-distribution, the chi-square distribution with degree of freedom 1. The rest of this section is to formulate the sufficient conditions of joint SRC being hold on the design matrix $\mathbf{Z}$.

Proposition 4.0.1. *Suppose there are infinitely many possible covariates* $\{\xi_j, j = 1, 2, \cdots, \}$, *and the covariate sequence satisfies Reisz condition. Namely, there exist*

fixed constants $0 < \rho_* < \rho^* < \infty$ *such that*

$$\rho_* \sum_{j=1}^{\infty} b_j^2 \le E\left(\sum_{j=1}^{\infty} b_j \xi_j\right)^2 \le \rho^* \sum_{j=1}^{\infty} b_j^2 \tag{4.8}$$

for all constants b_j*. Suppose the row vector of* X *is from a Gaussian distribution. Let* $\epsilon_0, \epsilon_1, \epsilon_2, \epsilon_3$*, and* a_0 *be positive constants in* $(0,1)$ *satisfying* $\epsilon_1 + \epsilon_2 < 1$ *and* $\epsilon_3 < \epsilon_2^2/2$*. Then for any set* $S \subset \{1, \ldots, n\}$ *with size* $|S| = a_0 n$*, and for all* (a_0, m, n, p) *satisfying*

$$m \le \min(p, \epsilon_1^2 n), \quad \text{and } \log \binom{p}{m} \le (\epsilon_3 \wedge c_{a_0}^2) n,$$

we have

$$P\left(\rho_* \tau_* \le \min_{\|u\|^2 + \|v\|_S^2 = 1} \min_{P_m} f(u,v) \le \max_{\|u\|^2 + \|v\|_S^2 = 1} \max_{P_m} f(u,v) \le (2 \vee \rho^* \tau^*)\right) \to 1 \tag{4.9}$$

as $n \to \infty$*, where*

$$\tau^* = (1 + \epsilon_1 + \epsilon_2)^2, \quad \tau_* = \frac{1}{3c_+(a_0) + (\sqrt{3c_+(a_0)} + \sqrt{\rho_*})^2 + \epsilon_0},$$

and

$$f(u,v) = \frac{\|X_m u\|^2}{n} + 2\frac{v' X_m u}{\sqrt{n}} + \|v\|^2, \tag{4.10}$$

with

$$u \in R^m, \quad v \in R^n, \quad X_m = X P_m'.$$

$$S = \{i : v_i \ne 0, 1 \le j \le n\}, \quad \|u\|_2^2 + \|v_S\|_2^2 = 1.$$

Here c_{a_0} *is a constant determined by* ρ_*, τ_*, a_0 *only, which is defined as follows. Denote*

$$t_1^* = \frac{1}{3c_+(a_0)\left(\left(\sqrt{2^{1/2}(1-a_0)t_{a_0} + 1} + \sqrt{1/(3c_+(a_0))}\right)^2 + 1/\rho^*\right)}. \tag{4.11}$$

If $\rho_* \tau_* \ge t_1^*$*, then*

$$c_{a_0} = \frac{(1-a_0)t_{a_0}}{\sqrt{2}} = \frac{(1-a_0)(\sigma_{a_0,+}\mu_{a_0,-} + \sigma_{a_0,-}\mu_{a_0,+})(\rho^* - \rho_*/\sigma_{a_0,+})}{\sqrt{2}\sigma_{a_0,-}(\rho^*\sigma_{a_0,+} + 3\rho_*\sigma_{a_0,-})}. \tag{4.12}$$

If $\rho_* \tau_* < t_1^*$*, then*

$$c_{a_0} = \frac{1}{2}\left(\left(\sqrt{\frac{1}{3c_+(a_0)\rho_*\tau_*} - \frac{1}{\rho^*}} - \sqrt{\frac{1}{3c_+(a_0)}}\right)^2 - 1\right) > 0. \tag{4.13}$$

Remark. The number t_1^* satisfies the equation,

$$\frac{1}{2}\left(\left(\sqrt{\frac{1}{3c_+(\alpha)t_1^*}-\frac{1}{\rho^*}}-\sqrt{\frac{1}{3c_+(\alpha)}}\right)^2-1\right)=(1-\alpha)t_\alpha/\sqrt{2},$$

which means it corresponds to the turning point's solution.

Remark. According to the theorem, the only constraint on a_0 is $a_0 \neq 1$. Thus s^* can be take any integer value but smaller than n. The penalty level on outlying coefficients are much smaller than that of covariates in X.

Remark. The upper bound of $(c_{a_0}^*)^2$ is $1/2$. Consider the special case when $a_0 \to 0$. Then $c_{a_0}^* \to \dfrac{\rho^*-\rho_*/2}{\sqrt{2}(\rho^*+3\rho_*)}$. When ρ^*/ρ_* is large, $(c_{a_0}^*)^2 \approx 1/2$. This bound is comparable to the bound of ϵ_3, which is bounded by $\epsilon_2^2/2 < 1/2$ as well. This means the new constraint on d^* in joint SRC is similar to the constrain for q^* in the original SRC.

Chapter 5

Breakdown Point and Generalized Penalty Weight Factor

5.1 Breakdown Point

The concept of breakdown point is a measure of the degree of robustness of an estimate in the presence of outlier [77]. It is equivalent to the maximum fraction of outliers which a given sample may allow without generating an extremely bad output. For example, the sample average, is not a robust estimator at all. It only has zero breakdown point value. This is because it may be unbounded with contamination on a single observation. As long as we change one value into a sufficiently large number, then the sample mean will breaks down. Another example is the sample median, which can has a high breakdown point value of 50%. A high breakdown point is a desired property for a robust estimator. However, many of the traditional estimators do not have high breakdown any more when the data dimension p is high. The sample size is not large enough to remove the effects of the outliers.

We claim that our estimator in (1.5) holds a high breakdown point value. A mild sufficient condition for positive breakdown point when both n and p are large is presented in next section.

5.2 Sufficient Conditions for Positive Breakdown.

Proposition 5.2.1. *Assume the notations and assumptions in Theorem 3.2.1. If there exists constants $M_1 < \infty$ and $M_2 > 0$ such that*

$$(\log p)(\log n) \leq M_1 n, \quad \frac{d^* \log p}{n} = M_2, \tag{5.1}$$

then estimator $\widehat{\theta}$ from (1.5) has the breakdown point value

$$\alpha^* = \frac{a_0 \wedge M_2}{K_* + 1}.$$

Proof. Assumption Proposition 5.2.1 indicates that for a positive breakdown: $|T| = a_* n$ with $a_* > 0$, the order of d^* has to be at least $O(n/\log p)$. If d^* is at least of order $O(n/\log p)$, then the proposed estimator can allow a fixed fraction of outliers. Thus it has a finite breakdown point.

Meanwhile, a necessary condition to control the rate consistency and/or estimation loss is to control the thresholding level. In other words, the factor in the bound $\lambda\sqrt{\log n/n}$ has to be finite. This requires $(\log p)(\log n) = O(n)$.

Now we prove the result in the proposition: Since $s^0/n \leq s^*/n = a_0$, combining the assumption (3.16) and in Theorem 3.2.1 and assumption in (5.1), the results on breakdown point's value follows immediately. □

Remark. In previous section we have provided the sufficient conditions for joint SRC being hold. It is required that $\log \binom{p}{d^*} = O(n)$. It indicates that the probability that joint SRC being hold when d^* has order up to $O(n)/\log p$. This conclusion is similar to that of the sufficient conditions for SRC, which is studied in [80]. Again, d^* is allowed to take order up to $n/\log p$. Thus the conditions proposed (5.1) are mild and they do not violate either of the two sufficient conditions.

5.3 Generalized Estimation Scheme by Varying Penalty Weight Factor

The proposed method in (1.5) can be viewed as a weighted ℓ_1 penalized regression. The penalty weight factor for β and γ is $\sqrt{\log p}$. Intuitively, using weight functions in greater order will eliminate many of the β's because it penalizes β much more heavily than γ's, while a not-so-low-order penalty weight factor is reluctant to identify the true outliers, especially when the number of outliers grows as quickly as the sample size.

In general, one can always use their own penalty weight factor and to obtain similar

statement as in Theorem 3.2.1. An estimator in such form is defined by:

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &\equiv (\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\gamma}}(\lambda)) \\ &\equiv \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\arg\min} \left\{ \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sqrt{n}\boldsymbol{\gamma}\|^2}{2n} + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{\beta}\|_1 + \frac{\lambda}{\sqrt{nF_{p,n}}}\|\boldsymbol{\gamma}\|_1 \right\} \end{aligned} \tag{5.2}$$

We claim that the weight function $F_{p,n} = \log p$ is optimistic in the order when there is a fixed contamination portion when $p >> n$.

The first advantage it brings is the sufficiency for positive breakdown point. Let $F_{p,n}$ be the squared weight ratio for penalty levels. For example, in our original settings of (1.5), $F_{p,n} = \log p$. We have shown that a sufficient condition for positive breakdown, when $n \to \infty$, is: $d^* F_{p,n} \propto n$. If we use another weight function $f_1(p, n)$, then the corresponding sufficient condition becomes $d^* f_1(p, n) \geq \alpha_2 n$.

Meanwhile, a guarantee for SRC/joint SRC being hold can allow d^* being in the order up to $n \log p$. This implies that in order NOT to violate SRC/joint SRC, one would consider the penalty weigh factor at least in the order of $\log p$.

By replacing the $\log p$ by $F_{p,n}$, it is not hard to generate all the analogous results in Section 3. The detailed formula and equations are omitted here, but we will highlight a few results. First we generate the modified $\widetilde{p}_{\epsilon,\gamma}$ as follows and keep $\widetilde{p}_{\epsilon,\beta}$ or $\lambda_{\epsilon,\beta}$ unchanged.

$$\frac{2\log \widetilde{p}_{\epsilon,\gamma}}{F_{p,n}} - 1 - \log\left(\frac{2\log \widetilde{p}_{\epsilon,\gamma}}{F_{p,n}}\right) = (2/m_2)\left\{\log\binom{n-s^o}{m_2} + \log(1/\epsilon)\right\} \tag{5.3}$$

And then generate the corresponding $\lambda_{\epsilon,\gamma}$ via $\log \widetilde{p}_{\epsilon,\gamma}$ by the equation (3.14). Then we obtain a generalized version of Theorem 3.2.1.

Theorem 5.3.1. *Let B be a deterministic set of $\{1, \cdots, p\}$ and T be a deterministic set of $\{1, \cdots, n\}$. Let $\widehat{\theta}$ be the LASSO estimator in (5.2), and $\widehat{\theta}^o = (\widehat{\beta}^o, \widehat{\gamma}^o)$ be the oracle estimator as defined in (3.7). Let*

$$m_1^* = d^* - d^0,\ m_2^* = s^* - s^0,\ 0 < \alpha \leq 2/3.$$

And let K_* be defined as before. Assume $|T| = s^0 < s^*$. Suppose

$$(K_* + 1)(|B| + \frac{|T|}{F_{p,n}}) \leq d^* \wedge \frac{s^*}{F_{p,n}} \tag{5.4}$$

Let $\lambda_{\epsilon,\beta}$ and $\lambda_{\epsilon,\gamma}$ be as in (3.13) and (3.14), with updated $\log \widetilde{p}_{\epsilon,\gamma}$. Let

$$\lambda_0 \equiv \lambda_{\epsilon,\beta} \vee \lambda_{\epsilon,\gamma}. \tag{5.5}$$

Then for given constant $\epsilon < 1/\sqrt{2}$ and for any $\lambda \geq \lambda_0$, with probability at least $1-\sqrt{2}\epsilon$, one has

$$\Big(\#\{j \notin B : \widehat{\beta}_j \neq 0\}\Big) F_{p,n} + (\#\{i \notin T : \widehat{\gamma}_i \neq 0\}) \ < \ 1 \vee \big(K_*(|B|F_{p,n} + |T|)\big), \tag{5.6}$$

and

$$\begin{aligned} r_*\|\widehat{\theta} - \widehat{\theta}^*\| &\leq \|Z(\widehat{\theta} - \widehat{\theta}^o)\|/\sqrt{n} \\ &\leq \lambda/\sqrt{n}\big(1 + \alpha\sqrt{2K_* r_*/r^*}\big)(\sqrt{|B|} + \sqrt{|T|/F_{p,n}}). \end{aligned} \tag{5.7}$$

Similarly, by replacing $\log p$ by $F_{p,n}$, one can obtain the analogue for all the corollaries and lemmas in Section 3. As for the order of the penalty levels, the thresholding effects brought by the two λ's: $\lambda_{\epsilon,\beta}$, and $\lambda_{\epsilon,\gamma}$, their maximum penalty levels on β's are summarized here:

- For $\lambda_{\epsilon,\beta}$

$$\begin{aligned} &\text{for } \beta : \lambda_{\epsilon,\beta}/\sqrt{n} = C_0\sqrt{\log p/n} \vee C_0, \\ &\text{for } \gamma : \lambda_{\epsilon,\beta}/\sqrt{nF_{p,n}} = C_0\sqrt{\frac{\log p}{F_{p,n}n}} \vee C_0\big(F_{p,n}^{-1/2}\big) \end{aligned}$$

- For $\lambda_{\epsilon,\gamma}$,

$$\begin{aligned} &\text{for } \beta : \lambda_{\epsilon,\gamma}/\sqrt{n} = C_1\sqrt{(\log n)(F_{p,n})/n}, \\ &\text{for } \gamma : \lambda_{\epsilon,\gamma}/\sqrt{nF_{p,n}} = C_1\sqrt{\log n/n}. \end{aligned}$$

5.4 The Order of Penalty Weight for Generalized Estimator and the Breakdown Point

Similarly, we can develop the sufficient conditions for positive breakdown property being hold for the estimator in (5.2).

Proposition 5.4.1. *Assume the notations and assumptions in Theorem 3.2.1. If there exists constants $M_1 < \infty$ and $M_2 > 0$ such that*

$$(F_{p,n})(\log n) \leq M_1 n, \quad \frac{d^* F_{p,n}}{n} = M_2, \tag{5.8}$$

then estimator $\widehat{\theta}$ from (5.9) has the breakdown point value

$$\alpha^* = \frac{a_0 \wedge M_2}{K_* + 1}.$$

In order to bound the thresholding levels, which has order up to $O\left(\sqrt{(\log n)F_{p,n}/n}\right) \vee O(1)$, a necessary condition is that $F_{p,n} \log n = O(n)$. This implies that the maximum order for $F_{p,n}$ is $n/\log n$. Together with the discussion on positive breakdown in Subsection 5.2.1, a sufficient condition for non-zero breakdown is that d_* is at least in the order of $O(n/F_{p,n})$. Without breaking the sufficient condition for SRC and joint SRC, a $F_{p,n}$ with order at least $O(\log p)$ is recommended. In summary, a good candidate of $F_{p,n}$ is in the order between $O(\log p)$ and $O(n/\log n)$, provided that $F_{p,n} \log n$ is up to the order of n.

In summary, if the weight is lower than $\log p$, then the positive breakdown point may not be guaranteed. If the weight has order of $O(n/\log n)$, the both loss bound of estimation and the false selected rates are out of control.

5.5 Choice of Weight Function for Datasets with Small Contamination

As we have discussed in previous section, the penalty weight function $\sqrt{\log p}$ will not breakdown for fixed contamination as $n \to \infty$. In real data applications, it is likely that only a few observations are outliers and the contamination is close to zero. Thus this penalty weight of $\log p$ would be conservative. Consider the case that only few observations, less than 1% , or rare events with fraction $< 0.1\%$, one only expect a small number of outliers. A modified version of (1.5) can be obtained by adjusting $F_{p,n}$ to a lower level, thus we will not select too many outliers in.

The weight adjustment is aimed to control the relative size ratio of the selected variables. For example, if one has prior information on the relative order between the number of nonzero β's and γ's, the squared root of that order ratio can be adopted. If

one is only aware of the contamination close to zero, then a good candidate is to use the penalty weight factor $\sqrt{\log p/\log n}$. The penalized regression problem becomes:

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &\equiv (\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\gamma}}(\lambda)) \\ &\equiv \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\arg\min} \left\{ \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sqrt{n}\boldsymbol{\gamma}\|^2}{2n} + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{\beta}\|_1 + \lambda\sqrt{\frac{\log n}{n \log p}}\|\boldsymbol{\gamma}\|_1 \right\} \end{aligned} \tag{5.9}$$

An advantage of this weight is to make the two λ's orders almost even:

$$\lambda_{\epsilon,\beta} = O(1) \vee O(\sqrt{\log p/n}), \quad \lambda_{\epsilon,\gamma} = O(\sqrt{\log p/n}).$$

Therefore the rate consistancy and loss bound will still hold while the dimension of p is up to the order of $e^{\alpha_1 n}$, where α_1 is a small positive number.

5.6 Discussion

A most important goal of a robust model selection procedures is to avoid over fitting. Suppose the signals of β's and γ's do not differ substantially. The larger the penalty weight factor $F_{p,n}$ is adopeted (at least $O(1)$), the faster the γ's would be selected in the selection sequence. Thus the noise is more likely to be absorbed by $\widehat{\gamma}_i$'s. By slowing down the selection speed of β's, we can shrink the size of the set $\widehat{A} \equiv \{j, \widehat{\beta}_j \neq 0\}$ to prevent over fitting, and to achieve the goal of robustness. In case one knows the contaminated sample size is large, one could use a larger penalty weight factor. In the case that the contamination is rare and/or the total sample size is low, one can use a smaller factor.

The weighting scheme here is different from the other weighted LASSO-typed methods in literature, such as adaptive LASSO [83]. The adapted weights essentially generated through data-driven methods, based on the individual signal strength. However, the goal of the penalty weight factor adopted here is to adjust the selection speed for the two groups of covariates, thus to balance the overall sizes and estimation errors between the two groups.This is crucial in the setting when the number of parameters are data dependent, such as the mean-shift model, where the number of outliers can grows as quickly as the sample size.

Meanwhile, suppose we do not incorporate a penalty weight factor. Consider a simpler alternative estimator, using the original LASSO instead. It is true that as long as the SRC/joint SRC conditions are satisfied, then the existed results from LASSO can be all applied automatically. For example, we can easily obtain the bound of $|\widehat{A} \setminus B| + |\widehat{S} \setminus T|$, which is proportional to the $|B| + |T|$. However, the explanation for this added-up result for β together with γ would be problematic, due to the cancellation between the two sets of coefficients: β and γ. This result is not informative enough. Though our proposed method is still in a joint fashion, by picking suitable weight factor, the cancellation can be controlled. For example, if we let $F_{p,n} = n/\log n$. When n is large, $|T|/\sqrt{F_{p,n}}$ and $|T|/\sqrt{n}$ will not differ much, where the latter is at most a small fraction. Thus this cancellation can be greatly removed.

Chapter 6

An Iterative Algorithm

In terms of the implementation procedures, one can always use LASSO or LARS algorithms to obtain the estimators in (1.5) and (5.9) and a selection sequence along with a series of tuning parameter λ. As for a fixed tuning parameter level, one may also develop an iterative procedure so that less inverting matrix will be needed, thus the computing time is reduced and the computing efficiency is improved. This will bring more benefits when the model dimension, either p or n are very large.

Suppose the solution of this equation for a fixed λ is unique, then we can start from an initial estimate of $\widehat{\beta}$, and do the following iterations:

Step 0. Obtain an initial estimate of $\widehat{\beta}^{(0)}$.

Step 0.1. Threshold the residuals $r_i = y_i - x_i\widehat{\beta}^{(0)}$, and obtain the initial estimate of γ's by the formula

$$\widehat{\gamma}_i^{(0)} = \mathrm{sgn}(y_i - \boldsymbol{x_i}\widehat{\boldsymbol{\beta}}) \left(|y_i - \boldsymbol{x_i}\widehat{\boldsymbol{\beta}}|/\sqrt{n} - \lambda \right)_+.$$

Step 0.2. Obtain the sets $\widehat{T}^{(0)} = \{i : \widehat{\gamma}_i^{(0)} \neq 0\}$ and $\widehat{T}^{c,(0)} = \{i : \widehat{\gamma}_i^{(0)} = 0\}$

Step 0.3 Use LASSO for the observations whose indices do not belong to $\widehat{T}^{(0)}$ to update $\widehat{\beta}$, i.e.,

$$\widehat{\beta}^{(1)} = \arg\min_{\beta} \frac{1}{2n} \|y_{\widehat{T}^{c,(0)}} - X_{\widehat{T}^{c,(0)}}\beta\|^2 + \lambda\|\beta\|_1.$$

Step 1.1. Go to step 0.1 and replace $\widehat{\beta}^{(0)}$ by its updated version, $\widehat{\beta}^{(0)}$.

If the solution is unique then the algorithm is guaranteed to converge to its global minimizer. This is because each step will decrease the value of the loss function, which

is convex. However, if the solution is not unique, the convergence might yield a local minimizer and the updating process is not stable. This iteration can be used for a pre-selected tuning parameter λ.

Chapter 7

Simulation

In this chapter, we apply our proposed method on simulated dataset.

7.1 Simulation Setting

- Design matrix X.

  The design matrix, X, has dimension of 100 samples and 500 covariate. $n = 100,\ p = 500$. And they are randomly drawn from the multivariate normal distribution with covariance matrix follows an AR(1) structure. The correlation coefficient of the AR(1) process is 0.5. This implies we assume moderate association among covariates.

- Parameters

  Suppose the true model only contains 5 nonzero β's. And their indices are a uniform random sample drawn between 1 and $p = 500$. Generate the true β's values from the uniform distribution $[-10,\ 10]$.

- Outlying Coefficients

  Suppose the contamination is α. And only a randomly selected αn γ's are non zeros. These non-zero γ's are i.i.d samples generated Cauchy distribution with scale coefficients 5.

- Repetitions

  Let the noise follows $N(0, \sigma^2 I_n)$, where $\sigma = 1$. Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\epsilon}$. Repeat the experiments for α is taking the following values:

$$\alpha = 0,\ 0.05,\ 0.1,\ 0.2, 0.3, \cdots, 0.9.$$

And repeat the whole set for 100 times.

- Estimator Candidates

  Five estimators are applied: the original LASSO, LASSO for augmented design matrix Z, estimator in (1.5), and estimator in (5.2) with $F_{p,n} = \log p / \log n$, and the last one is the estimator in (5.2) with $F_{p,n} = \log p / n$.

7.2 Results Comparison

We calculate the square root of MSE for both $\|\widehat{\beta} - \beta\|$ and $\|X\widehat{\beta} - X\beta\|$. The first quantity measures the estimation accuracy and the second one represents the prediction accuracy.

The results are listed in the tables below:

Table 7.1: Square root of MSE for $\|\widehat{\beta} - \beta\|$

LASSO-X	LASSO-Z	Z-log p	Z-log $p/\log n$	Z-log (p/n)	Contamination
0.04	0.04	0.06	0.04	0.04	0.00
0.60	0.05	0.07	0.05	0.05	0.05
0.50	0.05	0.08	0.05	0.05	0.10
1.22	0.06	0.10	0.06	0.06	0.20
6.59	0.10	0.13	0.09	0.08	0.30
2.70	0.13	0.15	0.11	0.10	0.40
4.87	0.17	0.19	0.15	0.14	0.50
4.31	0.21	0.23	0.18	0.17	0.60
8.98	0.25	0.26	0.21	0.20	0.70
5.19	0.30	0.29	0.26	0.24	0.80
6.69	0.33	0.31	0.29	0.28	0.90
7.31	0.38	0.35	0.34	0.31	0.95

The results indicate that indeed the best model among all is the one using factor $F_{p,n} = \log p/n$, since the size of p is not substantially larger than n. And the comparison is as follows:

- If no outliers exist, then original LASSO is best among all. This is as expected since the other models will impose larger model size by augmenting the design matrix X.

- If the dataset is contaminated with outliers, then the original LASSO is worst,

Table 7.2: Square root of MSE for $\|X\widehat{\beta} - X\beta\|$

LASSO-X	LASSO-Z	Z-log p	Z-log $p/\log n$	Z-log (p/n)	Contamination
0.81	0.81	1.36	0.83	0.85	0.00
15.02	0.91	1.56	0.93	0.95	0.05
11.67	0.97	1.65	0.98	0.99	0.10
29.98	1.21	2.01	1.14	1.14	0.20
172.71	1.77	2.72	1.57	1.51	0.30
67.57	2.36	3.06	1.99	1.85	0.40
124.52	3.37	3.93	2.77	2.53	0.50
110.27	4.15	4.78	3.48	3.20	0.60
228.65	5.03	5.37	4.14	3.75	0.70
136.57	6.13	6.12	5.12	4.65	0.80
173.49	7.05	6.53	5.98	5.49	0.90
186.06	7.98	7.32	6.70	6.06	0.95

and Z-log (p/n) is best among all.

- The differences among the estimators are increasing when contamination increases.

- Explanation is that small penalty allows outliers indicators being selected out easily, which absorbs part of the noise. Thus the β's is less likely to end up with fitting the noise.

- One remark here is that: if we use too heavy penalty on outliers. Then a lot more outliers will be selected, and it may end up with a small fraction of the data is useful for β's estimation. In other words, the samples that are used to capture the signals by β's is significantly decreased.

Chapter 8

Application to Aircraft Landing Data Set

In this chapter, we apply our proposed penalized method to a project of analyzing an aircraft landing data set. Before applying the penalized estimation scheme, we describe the background of this project, the problem setting and the data set.

8.1 Motivating Example and Data Set

The Federal Aviation Administration (FAA) is responsible for regulating air transportation and aviation safety. With the reported safety incidents in the terminal area continuing to increase, U.S. Government Accountability Office (GAO) recommends that the FAA (1) extend oversight of terminal area safety to include runway and ramp areas, (2) develop risk-based measures for runway safety incidents, and (3) improve information sharing about incidents, including runway overruns [21]. Runway overruns are defined as situations when aircrafts on take-off or landing roll extend beyond the end of the runway. To pursue these goals, the FAA has launched several research projects to study aircraft landing performance and to develop strategies for reducing the runway overrun rate.

A typical aircraft landing consists of a touchdown, deceleration to a maneuverable speed to leave runway at an exit, or to make a full stop before the end of the runway. A critical task in studying landing performance for improving runway safety is to identify the contributing factors for predicting the touchdown distance and for reducing the risk of runway overruns. Flight data from quick-access recorders may provide useful insight to the study of runway safety.

The present study in this chapter uses a data set collected and simulated from some collaborating airlines. All the flights contained in this data set were supposed to operate

with a representative type of aircrafts on a typical airport. More than 200 performance measures (also referred to as factors or variables) are recorded in the forms of time series. To facilitate the analysis of all the flights under a suitable framework, the flight data need to be properly organized. We choose to organize the flight data into time series which are aligned by the touchdown time, with equal time lag. Specifically, we extract the 1-Hz data set, and thus the time series all have the length of 600 seconds. It should be stressed that our study of this data set is only preliminary and the findings are subject to further examinations by aviation subject matter experts.

"Touchdown point" is the location where the main gear of an aircraft touches the runway. In our data analysis, we are interested in the airborne distance which indicates the distance between runway threshold and the touchdown point. An ideal landing includes a smooth touchdown at the target point, which is generally defined as a touchdown point at approximately 1,000 feet down the runway. If the touchdown point is too close to threshold, there is a risk of runway undershoot. If the touchdown distance is far beyond the maximum, the risk of overruns is greatly increased. In order to detect the potentially undesired landings or the outliers, statistically speaking, reliable estimation and prediction for airborne distance model is needed.

Since airborne distance is not recorded routinely in the flight data recorder, we use the methods introduced in [51] and [46] to estimate the airborne distance for our study. Clearly, not all the 200+ factors are of equal importance. There are often redundant measurements, such as various forms of speeds or measuring the same factor by using different machines or techniques. It is well known in statistical modeling literature, such a model with redundant factors tends to cause over-fitting and thus yields large estimation/prediction errors. We can apply model selection tool to reduce over-fitting and provide a solution to the overall modeling of the landing data. The goal of this project is to develop statistical models to i) identify key contributing factors for the airborne distance using information observed from the whole landing time series data, ii) detect undesired landings (or outliers in the context of statistical inference), and iii) provide recommendations or guidelines for monitoring landings.

8.2 Model Bank

We begin by applying model selection to each fixed time point (all aligned up with touchdown point). At each time point, a subset of contributing factors are identified, based on which a regression model is obtained, and, consequently, the parameters and the airborne distance can both be estimated. We then "connect" the landing interval which is considered safe/acceptable to form a tolerance band for the landing performance measures. Finally, we draw the predictiontrajectory over the time and connect all the tolerance bands over the time to form a tolerance tube/cone for monitoring each landing trace. A flight with predicted airborne distance trajectory which falls within the tolerance tube/cone would be considered as acceptable performance. On the other hand, if a flight is predicted with a non-negligible probability to be out of the tolerance tube, particularly during the time nearing touchdown, the landing would be considered as undesired. In the context of statistical analysis, this amounts to detecting an outlier. This study requires the additional effort on building a model "bank" that stores the prediction model with each time index.

Using the time series data set which records information of the whole landing process, we are able to identify the contributing factors via various statistical tools, such as AIC, BIC, LASSO, MCPlus and other penalized regression methods.

Once the subset of contributing factors for airborne distances is determined, one can store and estimate the parameters for these models, and thus a model bank over time can be established. In order to identify the undesired landings, one can estimate the airborne distance and update it at each subsequent time point. Draw a prediction trajectory and compare it with the tolerance tube.

Our data show that, on average, all the flights have touchdown around 2500 ft, with the minimum of 714 ft, the maximum of 5661 ft, and the standard deviation is 616 ft. The model bank we stored is recorded by time index, from 1 second before touchdown (TD) to 200 second before TD. The R^2 value increases as the aircraft approaches the runway threshold, and it can reach the level of 0.9 through the last 20 seconds period. During the last 50 seconds, the selected model can yield $R^2 > 0.8$.

Note that the aforementioned time-based approach, however, has the following critical problem in implementation: the estimation and model selection in the approach are based on the alignment with the time to touchdown, which means that the method in fact utilizes information at touchdown. For an incoming flight, its real time to touchdown during the landing procedure is clearly unknown. Thus it seems impossible to predict touchdown since we do not observe the real touchdown time in advance. To address this issue, a next step is to estimate the time index (when the aircraft will touchdown).

8.3 Time Index Estimation

The time-based model selection process aligns all the data at the time of touchdown, which in fact is unobservable for incoming flights before their touchdowns. Therefore, we need to estimate the current time index, or the time to touchdown.

The key idea is to apply the K-nearest neighbor method to obtain the time index when the aircraft is passing a fixed distance to ILS-antenna, say, 1000 ft away from the runway ILS-antenna which is located under about 954 ft away from the runway threshold.

We also split the data into two parts, training and testing, to check the accuracy of the time-to-TD estimation. The testing data set comprises of 300 flights, which is a random subset of roughly 10% of the data set. And the remaining 2840 flights are used for fitting data set. Here is a simple illustration for time estimation procedures:

- First, we view a flight in testing data set as an approaching flight. Assume that its ground distance to the ILS-antenna is approximately 800 ft (not touchdown yet). Record the corresponding height at that time point, denoted by h_*. And denote the unknown time index be t_*, which is the quantity we need to estimate.

- Look up the fitting data set which is comprised of 2840 flights (about 90%). Record the time indices and height information, for each of the historical flight as it passed say 800 ft before the ILS-antenna. Denote the time indices by $t_1, \cdots, t_{2840}$, and heights by $h_1, \cdots, h_{2840}$.

- Identify the k, say, $k = 10$ flights from the above 2840 flights which have the most similar heights, that is, the 10 nearest neighbors of the value of h_*. Here the distance is measured by the absolute difference between h_i and h_*. Denote the indices of the k flights by $n_1, \cdots, n_{10}$.

- Collect the time indices of the $k = 10$ nearest neighbors $t_{n_1}, \cdots, t_{n_{10}}$, and use the average, $\widehat{t_*} = \frac{1}{10}(t_{n_1} + \cdots + t_{n_{10}})$, as the estimate for t_*.

The time estimation procedure turns out to have a high prediction accuracy. We test on the following ground distances to ILS-antenna: 500ft, 800ft, 954ft, 1500ft, 1500ft, 3000ft, 8000ft, 10000ft, 15000ft, and 20000ft. The estimation error, measured in square root of MSE, is around 0.5 second as far as 3000 ft away. And even at 20000 ft to ILS-antenna, the time estimation error is about 1second.

8.4 Model Prediction and Outliers Detection

Once we obtain the time index for the incoming new flight, we can apply the model selected accordingly from the model bank and predict the touchdown distances. Recall that the R^2 of the model in the fitted data set can reach at least 90% level during the last 20 seconds, which is the average time when the aircrafts passing over 3000 ft before ILS-antenna. The model prediction accuracy thus is guaranteed since the time index estimation is accurate. The square root of MSE on prediction error is ranged between 150 and 230. Compared to the 614ft-standard deviation of touchdown distances, this prediction error range is equivalent to R^2 level of 94% and 86%, respectively. Even at 20000ft away from the ILS-antenna, the approximated R^2 is more than 50%.

An interesting observation is that there are several flights which consistently hold large prediction errors. More specifically, their prediction error is much larger than the normal scale, and/or with unstable signs when different time index models are applied. These are possible outliers in the dataset.

By checking the largest prediction errors, there are four flights consistently singled out: flight A, with TD=3289 ft, flight B, with TD=1100 ft and flights C and D, both of which have TD more than 5000 ft.

On average, the whole population flights have TD 2500 ft and the typical range of TD is from 1500ft to 3500ft. Further examinations indicate that flights B, C, and D are indeed outlying, but flight A should not be viewed as an outlier. Note that flight A was detected by the model bank as an outlier merely reflects the fact that the detection method based on model bank is in the same spirit as the classical t-test on residuals, which tends to be overly aggressive in selecting outliers.

We then apply the proposed estimators in (1.5) and (5.2) for simultaneous outlier detection and variable selection. As for the data preparation, we can either organize the testing flights in time aligned fashion by estimating the time indices first, or distance-to-ILS-antenna aligned fashion, by just calculating the distance to ILS-antenna from the raw data. A quick type of implementation is to use the second one. We collect the covariates for all the testing flights passing over 800 ft to ILS-antenna and use "glmnet" package of R for estimation.

The total number of parameters of the covariates is slightly over 200, and the sample size is 300. According to the previous prediction error, the contamination here is too low. This implies that the estimator in (5.9) is able to detect the right number of outliers.

As expected, the estimator in (1.5), using penalty weight factor $1/\sqrt{\log p}$ tends to select more outliers than that in (5.9) where the factor is $\sqrt{\log n}/\sqrt{\log p}$. With a certain stage of tuning parameter λ, the model selects flights B, C, and D as outliers, all of which enter the selection sequence quickly. Meanwhile, another 8 variables from the X-covariates are selected, including aircraft speed, wind speed, pitch angle and so on. This subset of precursors are consistent with their high selection frequencies when using the model bank. However, the outlying indicator variable for flight A enters the selection sequence almost in the end, after more than 100 X-covariates and more than 100 outlying indicator variables. This suggests that flight A might not be an outlier. The reason why the previous method yields a large prediction error could be due to the non-robustness of the fitted models stored in the model bank.

Over all, our proposed approach appears to model well the given landing data set and to identify effectively the outlying landing flights.

Chapter 9

Appendix

9.1 Proof of Lemma 3.3.2

Proof. First of all, we introduce the following notations:

$$A_2 \equiv \{1, \cdots, p\} \setminus A_1, \quad A_3 \equiv B, \quad A_4 \equiv A_1 \setminus A_3.$$

$$S_2 \equiv \{1, \cdots, n\} \setminus S_1, \quad S_3 \equiv T, \quad S_4 \equiv S_1 \setminus S_3.$$

Denote the negative gradient

$$\mathbf{g} \equiv Z'(y - Z\boldsymbol{\theta})/n. \tag{9.1}$$

Since $\widehat{\boldsymbol{\beta}}_{A_2} = 0,$ and $\widehat{\boldsymbol{\gamma}}_{S_2} = 0$, the $A_1 \cup S_1$-component of the negative gradient satisfies

$$\mathbf{g}_{A_1 \cup S_1} = \frac{1}{n} Z'_{A_1 \cup S_1}(y - Z_{A_1 \cup S_1} \widehat{\boldsymbol{\theta}}_{A_1 \cup S_1}).$$

Let $\widetilde{\boldsymbol{\varepsilon}} \equiv \mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\theta}}^o = (\mathbf{I}_n - \mathbf{P}_{B \cup T})\mathbf{y}$ be the residual from the oracle estimator. Then

$$\mathbf{g}_{A_1 \cup S_1} = Z'_{A_1 \cup S_1} \widetilde{\boldsymbol{\varepsilon}}/n + \boldsymbol{\Sigma}_{A_1 \cup S_1}(\widehat{\boldsymbol{\theta}}^{\boldsymbol{o}}_{\boldsymbol{A_1 \cup S_1}} - \widehat{\boldsymbol{\theta}}_{\boldsymbol{A_1 \cup S_1}}) \tag{9.2}$$

$$\iff \quad \Sigma^{-1}_{A_1 \cup S_1} \mathbf{g}_{A_1 \cup S_1} + (\widehat{\boldsymbol{\theta}}_{\boldsymbol{A_1 \cup S_1}} - \widehat{\boldsymbol{\theta}}^{\boldsymbol{o}}_{\boldsymbol{A_1 \cup S_1}}) = \Sigma^{-1}_{A_1 \cup S_1} Z'_{A_1 \cup S_1} \widetilde{\boldsymbol{\varepsilon}}/n, \tag{9.3}$$

where the covariance matrix $\boldsymbol{\Sigma}_{A_1 \cup S_1} = Z'_{A_1 \cup S_1} Z_{A_1 \cup S_1}/n$.

Let $\mathbf{v}_1 \equiv \boldsymbol{\Sigma}^{-1/2}_{A_1 \cup S_1} \mathbf{g}_{A_1 \cup S_1}$ and $\mathbf{v}_k \equiv \boldsymbol{\Sigma}^{-1/2}_{A_1 \cup S_1} \mathbf{g}_{A_k \cup S_k}$ for $k = 3, 4$. Let Q_k be the matrix that selection of $\{A_k \cup S_k\}$ from $\{A_1 \cup S_1\}$. Then

$$\mathbf{g}'_{A_k \cup S_k} Q_{A_k \cup S_k} \Sigma^{-1}_{A_1 \cup S_1} Z'_{A_1 \cup S_1} \widetilde{\boldsymbol{\varepsilon}}/n \leq \|\mathbf{v}_k\| \cdot \|P_{A_1 \cup S_1} \widetilde{\boldsymbol{\varepsilon}}\|/\sqrt{n} \tag{9.4}$$

By the triangle inequality, we have

$$\begin{aligned} & \|P_{A_1 \cup S_1} \widetilde{\boldsymbol{\varepsilon}}\|/\sqrt{n} \\ = \; & \|(P_{A_1 \cup S_1} - P_{B \cup T})\mathbf{y}\|/\sqrt{n} \\ \leq \; & \|(P_{B \cup S_1} - P_{B \cup T})\mathbf{y}\|/\sqrt{n} + \|(P_{A_1 \cup S_1} - P_{B \cup S_1})\mathbf{y}\|/\sqrt{n}. \end{aligned}$$

Together with (3.27) and (3.28), we then have

$$\begin{aligned}\|P_{A_1\cup S_1}\widetilde{\boldsymbol{\varepsilon}}\|/\sqrt{n} &\leq \sqrt{|S_4|}\zeta_\gamma(\mathbf{y};|S_4|,B,T)+\sqrt{|A_4|}\zeta_\beta(\mathbf{y};|A_4|,|S_4|,B,T)\\ &\leq \frac{\alpha\lambda}{\sqrt{r^*}}\left(\sqrt{|S_4|/\log p}+\sqrt{|A_4|}\right). \end{aligned} \tag{9.5}$$

Thus for $k=3,4$,

$$\begin{aligned}&\mathbf{g}'_{A_k\cup S_k}Q_{A_k\cup S_k}\Sigma^{-1}_{A_1\cup S_1}Z'_{A_1\cup S_1}\widetilde{\boldsymbol{\varepsilon}}/n \leq \|v_k\|\frac{\alpha\lambda}{\sqrt{r^*}}\left(\sqrt{|A_4|}+\sqrt{|S_4|/\log p}\right)\\ \leq\ & 2\alpha\|v_k\|^2+\frac{\alpha\lambda^2}{r^*}(|A_4|+|S_4|/\log p). \end{aligned} \tag{9.6}$$

Since

$$\mathbf{Q_4}(\widehat{\theta}^{\mathbf{o}}_{\mathbf{A_1}\cup\mathbf{S_1}}-\widehat{\theta}_{\mathbf{A_1}\cup\mathbf{S_1}})=\widehat{\theta}^{\mathbf{o}}_{\mathbf{A_4}\cup\mathbf{S_4}}-\widehat{\theta}_{\mathbf{A_4}\cup\mathbf{S_4}}=-\widehat{\theta}_{\mathbf{A_4}\cup\mathbf{S_4}},$$

$$\mathbf{v}_3=\mathbf{v}_1-\mathbf{v_4},$$

by (9.3) and (9.6), we have

$$\begin{aligned}&\|v_4\|^2-\|v_3\|^2+\|v_1\|^2\\ =\ & 2v_4'v_1=2g'_{A_4\cup S_4}Q_4\Sigma^{-1}_{A_1\cup S_1}g_{A_1\cup S_1}\\ =\ & 2g'_{A_4\cup S_4}Q_4\Sigma^{-1}_{A_1\cup S_1}Z'_{A_1\cup S_1}\widetilde{\varepsilon}/n-2g'_{A_4\cup S_4}\widehat{\theta}_{A_4\cup S_4} \qquad (9.7)\\ \leq\ & 2\alpha\|v_4\|^2+\frac{\alpha(|A_4|+|S_4|/\log p)\lambda^2}{r^*}. \qquad (9.8)\end{aligned}$$

Thus

$$(1-2\alpha)\|v_4\|^2+\|v_1\|^2+2g'_{A_4\cup S_4}\theta_{A_4\cup S_4}\leq\|v_3\|^2+\frac{\alpha(|A_4|+|S_4|/\log p)\lambda^2}{r^*}. \tag{9.9}$$

Similarly, we have the second inequality

$$\begin{aligned}&\|v_4\|^2+2g'_{A_4\cup S_4}\widehat{\theta}_{A_4\cup S_4}+\|\Sigma^{1/2}_{A_1\cup S_1}(\widehat{\theta}_{A_1\cup S_1}-\widehat{\theta}^o_{A_1\cup S_1})\|^2\\ \leq\ & \|v_3\|^2+2\|v_3\|\frac{\alpha\lambda}{\sqrt{r^*}}\left(\sqrt{|A_4|}+\sqrt{|S_4|/\log p}\right)+\frac{2\alpha^2\lambda^2}{r^*}(|A_4|+|S_4|/\log p).\end{aligned}$$

Combine the two inequalities together by weighted sum. We obtain

$$\begin{aligned}LHS \equiv\ & (1-2\alpha+w)\|v_4\|^2+\|v_1\|^2+2(1+w)g_4'\theta_4\\ &+w\|\Sigma^{1/2}_{A_1\cup S_1}(\widehat{\theta}_1-\widehat{\theta}^o_1)\|^2\\ \leq\ & (1+w)\|v_3\|^2+(\alpha+2w\alpha^2)\lambda^2(|A_4|+|S_4|/\log p)/r^*\\ &+2w\|v_3\|\alpha\lambda\left(\sqrt{|A_4|}+\sqrt{|S_4|/\log p}\right)/\sqrt{r^*}. \qquad (9.10)\end{aligned}$$

By KKT conditions and (9.1),

$$\begin{aligned}\|g_{A_4\cup S_4}\|^2 &= \|g_{A_4}\|^2+\|g_{S_4}\|^2=\sum_{j\in A_4}\lambda^2+\sum_{i\in S_4}(\lambda^2/\log p)\\ &= \lambda^2\left(|A_4|+\frac{|S_4|}{\log p}\right),\end{aligned} \tag{9.11}$$

$$\begin{aligned}\|g_{A_3\cup S_3}\|^2 &= \|g_{A_3}\|^2+\|g_{S_3}\|^2=\sum_{j\in A_3}\lambda^2+\sum_{i\in S_3}(\lambda^2/\log p)\\ &= \lambda^2\left(|A_3|+\frac{|S_3|}{\log p}\right)=\lambda^2\left(|B|+\frac{|T|}{\log p}\right).\end{aligned}$$

Meanwhile, KKT conditions in (**??**) yield $\widehat{\theta}^o_j=0$ and $0\le \widehat{\theta}'_{A_j\cup S_j}g_j=|\widehat{\theta}'_{A_j\cup S_j}g_{A_j\cup S_j}|$ for $j\in A_4\cup S_4$. Thus we can bound LHS from both sides and obtain

$$\begin{aligned}LHS &\equiv (1-2\alpha+w)\|v_4\|^2+\|v_1\|^2+2(1+w)g'_{A_4\cup S_4}\theta_{A_4\cup S_4}\\ &\quad +w\|\Sigma^{1/2}_{A_1\cup S_1}(\widehat{\theta}_1-\widehat{\theta}^o_{A_1\cup S_1})\|^2\\ &\ge (1-2\alpha+w)\|g_{A_4\cup S_4}\|^2/r^*+\|g_{A_1\cup S_1}\|^2/r^*+2(1+w)|g'_{A_4\cup S_4}\theta_{A_4\cup S_4}|\\ &\quad +wr_*\|\widehat{\theta}_{A_1\cup S_1}-\widehat{\theta}^o_{A_1\cup S_1}\|^2\\ &\ge (2-2\alpha+w)\|g_{A_4\cup S_4}\|^2/r^*+\|g_{A_3\cup S_3}\|^2/r^*.\end{aligned} \tag{9.12}$$

Insert (9.12) into (9.10), we have

$$\begin{aligned}&(2-2\alpha+w)\|g_{A_4\cup S_4}\|^2/r^*+\|g_{A_3\cup S_3}\|^2/r^*\\ \le\ &(1+w)\|v_3\|^2+(\alpha+2w\alpha^2)\lambda^2(|A_4|+|S_4|/\log p)/r^*\\ &+2w\|v_3\|\alpha\lambda\left(\sqrt{|A_4|}+\sqrt{|S_4|/\log p}\right)/\sqrt{r^*}.\end{aligned} \tag{9.13}$$

It is easy to have

$$\begin{aligned}&\frac{(2-3\alpha+w-2w\alpha^2)}{r^*}\|g_4\|^2\\ \le\ &\left(\frac{1+w}{r_*}-\frac{1}{r^*}\right)\|g_3\|^2+w\alpha\left(\frac{\|v_3\|^2}{(1-\alpha)}+2(1-\alpha)\|g_4\|^2/r^*\right)\\ \Longleftrightarrow\quad &\frac{1}{r^*}\Big((2-3\alpha+w-2w\alpha^2-2w\alpha(1-\alpha)\Big)\|g_{A_4\cup S_4}\|^2\\ &\le\left(\frac{1+w}{r_*}-\frac{1}{r^*}\right)\|g_{A_3\cup S_3}\|^2+\frac{w\alpha}{(1-\alpha)r_*}\|g_{A_3\cup S_3}\|^2.\end{aligned} \tag{9.14}$$

Here the last inequality is based on the fact that

$$\|v_k\|^2=\|\Sigma^{-1/2}_{A_1\cup S_1}g_{A_k\cup S_k}\|^2\le\frac{1}{r_*}\|g_{A_k\cup S_k}\|^2.$$

Inequality (9.14) is equivalent to

$$K_{*,w}\|g_{A_4\cup S_4}\|^2 \le \|g_{A_3\cup S_3}\|^2, \tag{9.15}$$

where

$$K_{*,w} = \frac{((1-\alpha)(1+w)+w\alpha)r^*/r_* - (1-\alpha)}{(1-\alpha)(2-3\alpha+w-2w\alpha)}.$$

Let $K_* \equiv \min_{w\ge 0} K_{*,w}$. Then

$$|A_1| - |B| + \frac{|S_1|-|T|}{\log p} < K_*\left(|B| + \frac{|T|}{\log p}\right).$$

$K_{*,w}$ is a constant that is determined by α and w only. By taking suitable values of α and w, the value of K_* can be as small as possible, thus the bound can be optimal. For example, take $w = 1$, $\alpha = 1/2$, then $K_* \le 6r^*/r_* - 2$. We cam also use $6r^*/r_* - 2$ as a rough estimation for the bound.

Notice that our conclusion is in strict inequality. By (9.12) we know if the equality in (3.31) holds, then $\|g_{A_4\cup S_4}\| = 0$, which implies that $|A_4| = |S_4| = m_1 = m_2 = 0$, which contradicts with the assumption for m_1 and m_2 being positive integers.

Next, we show the loss bound of $\widehat{\theta}$ and $Z\widehat{\theta}$ in (3.32). Since

$$|A_1| = m_1 + d^0 \le (1+K_*)|B| \le d^*, \quad S_1 = m_2 + s^0 \le (1+K_*)|T| \le s^*,$$

the matrix $\Sigma_{A_k\cup S_k}$ is invertible for $k = 1, 2, 3, 4$. Then

$$\begin{aligned}
& \frac{1}{n}\|Z(\widehat{\theta} - \widehat{\theta}^o)\|^2 && (9.16)\\
= \;& (\widehat{\theta} - \widehat{\theta}^o)' \frac{Z'_{A_1\cup S_1} Z_{A_1\cup S_1}}{n} (\widehat{\theta} - \widehat{\theta}^o) \\
= \;& (\widehat{\theta}_{A_1\cup S_1} - \widehat{\theta}^o_{A_1\cup S_1})' Z'_{A_1\cup S_1}\widehat{\varepsilon}/n - (\widehat{\theta}_{A_1\cup S_1} - \widehat{\theta}^o_{A_1\cup S_1})' g_{A_1\cup S_1} \\
\le \;& (\widehat{\theta}_{A_1\cup S_1} - \widehat{\theta}^o_{A_1\cup S_1})' Z'_{A_1\cup S_1}\widehat{\varepsilon}/n - (\widehat{\theta}_{A_3\cup S_3} - \widehat{\theta}^o_{A_3\cup S_3})' g_{A_3\cup S_3}. && (9.17)
\end{aligned}$$

The last inequality is due to

$$\begin{aligned}
& (\widehat{\theta}_{A_1\cup S_1} - \widehat{\theta}^o_{A_1\cup S_1})' g_{A_1\cup S_1} \\
= \;& (\widehat{\theta}_{A_3\cup S_3} - \widehat{\theta}^o_{A_3\cup S_3})' g_{A_3\cup S_3} + (\widehat{\theta}_{A_4\cup S_4} - \widehat{\theta}^o_{A_4\cup S_4})' g_{A_4\cup S_4} \\
= \;& (\widehat{\theta}_{A_3\cup S_3} - \widehat{\theta}^o_{A_3\cup S_3})' g_{A_3\cup S_3} + (\widehat{\theta}'_{A_4\cup S_4} g_{A_4\cup S_4} \\
\ge \;& (\widehat{\theta}_{A_3\cup S_3} - \widehat{\theta}^o_{A_3\cup S_3})' g_{A_3\cup S_3}.
\end{aligned}$$

The first term in the formula in (9.17) above has norm:

$$\begin{aligned}
&\|(\widehat{\theta}_{A_1\cup S_1}-\widehat{\theta}^o_{A_1\cup S_1})'Z'_{A_1\cup S_1}\widehat{\varepsilon}/n\| \le \|Z'_{A_1\cup S_1}(\widehat{\theta}_{A_1\cup S_1}-\widehat{\theta}^o_{A_1\cup S_1})'P'_{A_1\cup S_1}\widehat{\varepsilon}/n\| \\
\le\ & \|\Sigma^{1/2}_{A_1\cup S_1}(\widehat{\theta}_{A_1\cup S_1}-\widehat{\theta}^o_{A_1\cup S_1})\|\|P'_{A_1\cup S_1}\widehat{\varepsilon}\|/\sqrt{n} \\
\le\ & \|\Sigma^{1/2}_{A_1\cup S_1}(\widehat{\theta}_{A_1\cup S_1}-\widehat{\theta}^o_{A_1\cup S_1})\|\alpha\lambda/\sqrt{r^*}(\sqrt{|A_4|}+\sqrt{|S_4|/\log p}) \\
\le\ & \sqrt{2/r^*}\alpha\lambda\|\Sigma^{1/2}_{A_1\cup S_1}(\widehat{\theta}_{A_1\cup S_1}-\widehat{\theta}^o_{A_1\cup S_1})\|\sqrt{|A_4|+|S_4|/\log p},
\end{aligned}$$

where the third inequality is derived from (9.5). By (3.31), we have

$$\sqrt{|A_4|+|S_4|/\log p}\le\sqrt{K_*(|B|+|T|/\log p)}. \tag{9.18}$$

The second term also has bounded norm by

$$\begin{aligned}
&\|(\widehat{\theta}_{A_3\cup S_3}-\widehat{\theta}^o_{A_3\cup S_3})'g_{A_3\cup S_3}\| \le \frac{1}{\sqrt{r_*}}\|\Sigma^{1/2}_{A_1\cup S_1}(\widehat{\theta}_{A_1\cup S_1}-\widehat{\theta}^o_{A_1\cup S_1})\|\|g_{A_3\cup S_3}\| \\
=\ & \frac{\lambda}{\sqrt{r_*}}\|\Sigma^{1/2}_{A_1\cup S_1}(\widehat{\theta}_{A_1\cup S_1}-\widehat{\theta}^o_{A_1\cup S_1})\|\sqrt{|B|+|T|/\log p}.
\end{aligned} \tag{9.19}$$

Plug (9.18) and (9.19) in (9.17). We obtain

$$\|\Sigma^{1/2}_{A_1\cup S_1}(\widehat{\theta}_{A_3\cup S_3}-\widehat{\theta}^o_{A_3\cup S_3})\|\le\lambda\big(1/\sqrt{r_*}+\alpha\sqrt{2K_*/r^*}\big)\sqrt{|B|+|T|/\log p}. \tag{9.20}$$

Since

$$\|\Sigma^{1/2}_{A_1\cup S_1}(\widehat{\theta}-\widehat{\theta}^o)\|=Z'(\widehat{\theta}-\widehat{\theta}^o)\|/\sqrt{n},\ \text{and}\ \sqrt{r_*}\|u\|\le\|\Sigma^{1/2}_{A_1\cup S_1}u\|\le\sqrt{r^*}\|u\|,$$

this leads to (3.32):

$$r_*\|\widehat{\theta}-\widehat{\theta}^*\|\le\|Z'(\widehat{\theta}-\widehat{\theta}^o)\|/\sqrt{n}\le\lambda\big(1/\sqrt{r_*}+\alpha\sqrt{2K_*/r^*}\big)\sqrt{|B|+|T|/\log p}.$$

□

9.2 Proof of Lemma 3.3.3

Proof. The first step is to show that

$$P\left\{\zeta_\gamma(\varepsilon;m_2,B,T)\ge\sigma\sqrt{(2/(n\log p))\log\widetilde{p}_{\epsilon,\gamma}}\right\}\le\epsilon/\sqrt{2}. \tag{9.21}$$

Since m_2, B and T are deterministic, the set S_1 has $\binom{n-s^0}{m_2}$ possibilities. Thus $\{nm_2\zeta_\gamma^2(\boldsymbol{\varepsilon}; m_2, B, T)\}$ is the maximum of $\binom{n-s^0}{m_2}$ variables with the $\chi^2_{m_2}$ distribution, so

$$\begin{aligned} & P\left\{\sqrt{\log p}\ \zeta_\gamma(\varepsilon; m_2, B, T) \geq \sigma\sqrt{\frac{2}{n}\log \widetilde{p}_{\epsilon,\gamma}}\right\} \\ = \ & P\left\{\zeta_\gamma(\varepsilon; m_2, B, T) \geq \sigma\sqrt{\frac{2\log \widetilde{p}_{\epsilon,\gamma}}{n\log p}}\right\} \\ \leq \ & \binom{n-s_0}{m_2} P\{\chi^2_{m_2} \geq m_2(1+x)\}, \end{aligned} \tag{9.22}$$

where $x = (2\log \widetilde{p}_{\varepsilon,\gamma})/\log p - 1 > 0$.

Notice that $\chi^2_{m_2}/(1+x)$ has gamma distribution with parameters $(m/2, (1+x)/2)$. Using the derivation and inequality as seen in Zhang (2010), we obtain

$$P\{\chi^2_{m_2} \geq m_2(1+x)\} \leq \frac{e^{-m_2x/2}(1+x)^{m_2/2}}{2\log \widetilde{p}_{\varepsilon,\gamma}/\log p}. \tag{9.23}$$

Therefore, take the logarithm and we can prove the inequality in (9.21) by

$$\begin{aligned} & \log\left(P\Big\{\zeta_\gamma(\varepsilon; m_2, B, T) \leq \sigma\sqrt{\frac{2\log \widetilde{p}_{\epsilon,\gamma}}{n\log p}}\Big\}\right) \\ \leq \ & \log\binom{n-s_0}{m_2} - \frac{m_2x}{2} + \frac{m_2\log(1+x)}{2} - \log\frac{2\log \widetilde{p}_{\varepsilon,\gamma}}{\log p} \\ \leq \ & \log\binom{n-s_0}{m_2} - \frac{m_2}{2}\left(\frac{2\log \widetilde{p}_{\varepsilon,\gamma}}{\log p} - 1 - \log\left(\frac{2\log \widetilde{p}_{\varepsilon,\gamma}}{\log p}\right)\right) \\ = \ & \log\left(\frac{\epsilon}{\sqrt{2}}\right). \end{aligned}$$

By the Bonferroni inequality, it suffices to show that

$$P\left\{\zeta_\beta(\varepsilon; m_1, m_2, B, T) \geq \sigma\sqrt{(2/n)\log \widetilde{p}_{\epsilon,\beta}}\right\} \leq \epsilon/\sqrt{2}. \tag{9.24}$$

The proof is similar to the statement above. Notice that $\{nm_1\zeta_\beta^2(\boldsymbol{\varepsilon}; m_1, m_2, B, T)\}$ is the maximum of $\binom{p-d^0}{m_1}\binom{n-s^0}{m_2}$ variables with the $\chi^2_{m_1}$ distribution, so

$$P\left\{\zeta_\beta(\varepsilon; m_1, m_2, B, T) \geq \sigma\sqrt{\frac{2\log \widetilde{p}_{\epsilon,\beta}}{n}}\right\} \leq \binom{p-d_0}{m_1}\binom{n-s_0}{m_2} P\{\chi^2_{m_1} \geq m_1(1+x)\}, \tag{9.25}$$

with $x = (2\log \widetilde{p}_{\varepsilon,\beta})/\log p - 1 > 0$. □

9.3 Proof of Lemma 3.3.4

Proof. The outline of the proof is as follows. Step 1 shows

$$\zeta_\beta(y; m_1, m_2, B, T) \le \zeta_\beta(\varepsilon; m_1, m_2, B, T)..$$

Step 2 looks for the range of $\log \widetilde{p}_{\epsilon,\beta}$ and $\log \widetilde{p}_{\epsilon,\gamma}$ while m_1 and m_2 vary. In the last step , together with Lemma 3.3.3, we will prove (3.34) in the last step.

Step 1. By the triangle inequality, for any $m_1 \le m_1^* = d^* - d^0$, and any $m_2 \le s^* - s^0$,

$$\zeta_\beta(y; m_1, m_2, B, T) \le \zeta_\beta(\varepsilon; m_1, m_2, B, T) + \zeta_\beta(Z\theta; m_2, B, T), \tag{9.26}$$

where

$$\zeta_\beta(Z\theta; m_1, m_2, B, T) = \frac{\|(P_{A\cup S} - P_{B\cup T})Z\theta\|}{(m_1 n)^{1/2}}. \tag{9.27}$$

Since the linear span by column vectors from index set $\{B \cup T\}$ is a subspace of that from index set $\{A \cup S\}$, and $Z\theta = Z_{B\cup T}\theta$ lies in the smaller linear span, we have (9.27) equals 0 and

$$\zeta_\beta(y; m_1, m_2, B, T) \le \zeta_\beta(\varepsilon; m_1, m_2, B, T). \tag{9.28}$$

Step 2. Recall that $\log \widetilde{p}_{\epsilon,\beta}$ and $\log \widetilde{p}_{\epsilon,\gamma}$ are solutions of (3.11) and (3.12). Thus for deterministic $m_1 \le m_1^*$ and $m_2 \le m_2^*$, we have the approximation

$$\log \widetilde{p}_{\epsilon,\beta} \approx \frac{1}{m_1}\left(\log \binom{p-d^0}{m_1} + \log \binom{n-s^0}{m_2}\right) \approx \log \frac{p}{m_1} + \frac{m_2 \log(n/m_2)}{m_1}. \tag{9.29}$$

Notice that $m_2 \log(n/m_2)$ is in the order between $O(\log n)$ and $O(n)$. If $m_2 = (n - s^0)/2$, then it reaches the maximum value at $(n \log 2)/2$. It reaches the minimum value of $\log n$ when $m_1 = 1$.

Therefore, with $m_1 = 1$ and $m_2 = m_2^*$,

$$\log \widetilde{p}^{\,1}_{\epsilon,\beta} \approx (\log p) \vee n,$$

and with $m_1 = m_1^*,\ m_2 = m_2^*$,

$$\log \widetilde{p}^{\,*}_{\epsilon,\beta} \approx \log(p/m_1^*) \vee (n/m_1^*).$$

Then for any $m_1 \le m_1^*$ and $m_2 = m_2^*$,

$$\log \widetilde{p}_{\epsilon,\beta} \le \log \widetilde{p}^{\,1}_{\epsilon,\beta} \vee \log \widetilde{p}^{\,*}_{\epsilon,\beta}.$$

Similarly, by (3.12), for deterministic m_2, we have

$$\frac{\log \widetilde{p}_{\epsilon,\gamma}}{\log p} \approx \frac{2}{m_2} \log \binom{n - s^0}{m_2} \approx \log \frac{n}{m_2},$$

which is between $O(1)$ and $\log n$. Here the lower bound is obtained when $m_2 \propto m_2^* \propto n$ and the maximum is obtained when $m_2 = 1$. Therefore, for any $m_2 \leq m_2^*$, we have

$$\log \widetilde{p}_{\epsilon,\gamma} \leq \log \widetilde{p}^*_{\epsilon,\gamma} \log n. \tag{9.30}$$

Step 3. By Lemma 3.3.3, for any $m_1 \leq m_1^*$ and $m_2 \leq m_2^*$, we have

$$P\left(\zeta_\beta(\varepsilon; m_1, m_2, B, T) \geq \sigma\sqrt{(2/n)\log \widetilde{p}_{\epsilon,\beta}}\right) \leq \epsilon/\sqrt{2} \tag{9.31}$$

Thus $\log \widetilde{p}^*_{\epsilon,\beta} \geq \log \widetilde{p}_{\epsilon,\beta}$, and

$$P\left(\zeta_\beta(\varepsilon; m_1, m_2, B, T) \geq \sigma\sqrt{\frac{2}{n}(\log \widetilde{p}^{\,1}_{\epsilon,\beta} \vee \log \widetilde{p}^*_{\epsilon,\beta})}\right) \leq \epsilon/\sqrt{2}. \tag{9.32}$$

By inequality (9.28) and the definition of $\lambda_{\epsilon,\beta}$ in (3.13), we obtain for any deterministic m_1 and m_2,

$$P\left\{\zeta_\beta(y; m_1, m_2, B, T) \geq \sqrt{\frac{\log n}{n}}\lambda_{\epsilon,\beta}\right\} \tag{9.33}$$

$$\leq P\left\{\zeta_\beta(\varepsilon; m_1, m_2, B, T) \geq \sqrt{\frac{\log n}{n}}\lambda_{\epsilon,\beta}\right\}$$

$$\leq P\left\{\zeta_\beta(\varepsilon; m_1, m_2, B, T) \geq \frac{\sqrt{r^*}}{\alpha}(\sigma\sqrt{(2/n)\log \widetilde{p}_{\epsilon,\beta}})\right\} \leq \epsilon/\sqrt{2}, \tag{9.34}$$

with $\log \widetilde{p}_{\epsilon,\beta}$ calculated from given m_1 and m_2.

Similarly, using Lemma 3.3.3, for any $m_2 \leq m_2^*$, we have

$$P\left(\zeta_\gamma(\varepsilon; m_2, B, T) \geq \sigma\sqrt{\frac{2\log \widetilde{p}_{\epsilon,\gamma}}{n \log p}}\right) \leq \epsilon/\sqrt{2}. \tag{9.35}$$

Insert (9.30) into the definition of $\lambda_{\epsilon,\gamma}$, we obtain

$$\lambda_{\epsilon,\gamma} > \frac{\sqrt{r^*}}{\alpha}\left(\sigma\sqrt{\frac{2\log \widetilde{p}_{\epsilon,\gamma}}{\log n}}\right). \tag{9.36}$$

With (9.28), (9.35), and (9.36), we have

$$P\left(\zeta_\gamma(\varepsilon; m_2, B, T) \geq \sqrt{\frac{\log n}{n \log p}}\lambda_{\epsilon,\gamma}\right) \tag{9.37}$$

$$\leq P\left(\zeta_\gamma(\varepsilon; m_2, B, T) \geq \sigma\sqrt{\frac{2\log \widetilde{p}_{\epsilon,\gamma}}{n \log p}}\right) \leq \epsilon/\sqrt{2}. \tag{9.38}$$

Combine (9.34) and (9.38), and for any $\lambda > \lambda_{\epsilon,\beta} \vee \lambda_{\epsilon,\gamma}$, (3.34) is proved. □

9.4 Proof of Proposition 4.0.1

Proof. We first will show the upper bound of (4.10). If $\max_u \max_{P_m} \frac{\|X_m u\|^2}{n\|u\|^2} \le \rho^*\tau^*$,

$$\begin{aligned}
\max_{\|u\|^2+\|v\|_A^2=1} \max_{P_m} f(u,v) &= \max_{\|u\|^2+\|v\|_A^2=1} \max_{P_m} \frac{\|X_m u\|^2}{n} + 2\frac{v' X_m u}{\sqrt{n}} + \|v\|_{A^*}^2 \\
&\le \max_u \max_{P_m} \frac{2\|X_m u\|^2}{n} + 2\|v\|_{A^*}^2 \\
&\le (\rho^*\tau^* \vee 2)(\|u\|^2 + \|v\|_{A^*}^2) \\
&= (\rho^*\tau^* \vee 2), \qquad (9.39)
\end{aligned}$$

Thus by Zhang's theorem, let $\epsilon_k, k = 1,2,3,4$, be positive constants in $(0,1)$, satisfying $m \le \min(p, \epsilon_1^2 n)$, $\epsilon_1 + \epsilon_2 < 1$, and $\epsilon_3 + \epsilon_4 = \epsilon_2^2/2$. Then for all (m,n,p) satisfying $\log \binom{p}{m} \le \epsilon_3 n$,

$$P\{c^*(m) \ge \tau^*\rho^*\} \le e^{-n\epsilon_4}, \qquad (9.40)$$

where $\tau^* = (1+\epsilon_1+\epsilon_2)^2$.

Therefore

$$\begin{aligned}
& P\{\max_{\|u\|^2+\|v\|_A^2=1} \max_{P_m} f(u,v) \le (\rho^*\tau^* \vee 2)\} \qquad (9.41) \\
\ge\ & P\left\{\max_u \max_{P_m} \frac{\|X_m u\|^2}{n\|u\|^2} \le \rho^*\tau^*\right\} \qquad (9.42) \\
\ge\ & 1 - e^{-\epsilon_4 n} \qquad (9.43)
\end{aligned}$$

9.4.1 Lower Bound

In this subsection, we are to look for the probability bound of

$$P\{\min_{\|u\|^2+\|v\|_{A^*}^2=1} f(u,v) \le \rho_*\tau_*\}$$

for a fixed P_m, this is because

$$\begin{aligned}
& P\{\min_{P_m} \min_{\|u\|^2+\|v\|_{A^*}^2=1} f(u,v) \ge \tau_*\rho_*\} \\
=\ & 1 - P\{\min_{P_m} \min_{\|u\|^2+\|v\|_{A^*}^2=1} f(u,v) \le \rho_*\tau_*\} \\
\ge\ & 1 - \binom{p}{m} P\{\min_{\|u\|^2+\|v\|_{A^*}^2=1} f(u,v) \le \tau_*\rho_*\}.
\end{aligned}$$

Let $z_i = X_{i,\cdot}P_m'u$,then z_i's are i.i.d random variables with mean zero and variance

$$\sigma_z^2 = u'\Sigma_m u, \tag{9.44}$$

$$\rho_*\|u\|^2 \leq \sigma_z^2 \leq \rho^*\|u\|^2. \tag{9.45}$$

Thus

$$\begin{aligned} f(u,v) &= \frac{\sum_{i=1}^n z_i^2}{n} + 2\frac{\sum_{i\in A} v_i z_i}{\sqrt{n}} + \|v_{A^*}\|^2 \\ &= \frac{\|z_{(A^*)^c}\|^2}{n} + \|\frac{z_{A^*}}{\sqrt{n}} + v_{A^*}\|^2 \\ &\geq \frac{\|z_{(A^*)^c}\|^2}{n} + (\frac{\|z_{A^*}\|}{\sqrt{n}} - \|v_{A^*}\|)^2 \\ &= \frac{\|z_{(A^*)^c}\|^2}{n} + (\frac{\|z_{A^*}\|}{\sqrt{n}} - \sqrt{1-\|u\|^2})^2 \end{aligned} \tag{9.46}$$

In (9.47), the lower bound could be dominated by either term, depending on the L^2 norm of $\|u\|$. We will discuss it in two cases: given P_m, $\forall t_1$ with certain constraints,

$$P\{\min_{\|u\|^2+\|v\|_{A^*}^2=1} f(u,v) \leq t_1\} \leq P\{\min_{u,v,\|u\|^2\leq c_0} f(u,v) \leq t_1\} \vee P\{\min_{u,v,\|u\|^2\geq c_0} f(u,v) \leq t_1\} \tag{9.47}$$

The structure of the derivation is given as below: first of all, we review the conceptions of net and covering numbers which are common techniques applied in random matrices. Then the properties for strict sub-Gaussian distributions are also summarized, together with lemmas needed for deviations later on. All the lemma are also applicable to Gaussian distributions. After that, we are to show the probabilities in the two components as shown in (9.47) separately. In the end, we combine and discuss the two cases, and look for a sharper bound for the probability in (9.47).

9.4.2 Net and Covering numbers

A typical and alternative way to derive the probability that the eigenvalue being bounded is to use the nets and covering numbers. Let $S_\epsilon^{n+m-1} \in S^{n+m-1}$ to be an ϵ net with radius ϵ, i.e., use the balls with radius ϵ to cover the S^{n+m-1}. We give an upper bound of minimum balls needed for covering. If we are allowed to pack $N_{\epsilon/2}$ balls of radius $\epsilon/2$ into the sphere S^{n+m-1}. All of the balls have centers on the sphere and

are thus contained by the $(1+\epsilon/2)$ ball $\in R^{n+m}$. Then

$$N_{\epsilon/2} \times (\epsilon/2)^{n+m} \leq (1+\epsilon/2)^{n+m}$$
$$N_{\epsilon/2} \leq (\frac{1+\epsilon/2}{\epsilon/2})^{n+m} = (1+2/\epsilon)^{n+m}$$

It is not hard to show that, if we replace the $\epsilon/2$ radius balls by ϵ balls with the same centers, then we are allow to cover the whole ball thus the whole unit sphere S^{t+m-1}, and denote $S_\epsilon^{t+m-1} = \{w_1, w_2, \cdots, w_{(1+2/\epsilon)^{t+m}}\}$ where $w_j \in S^{t+m-1}$ and they are centers of covering balls.

In Cai et al (2010), it has been shown that $\forall m \times m$ symmetric matrix $M = X'X/n$, where n is the number of the rows of X, we have

$$-\|u-\tilde{u}\|\|M\|\|u+\tilde{u}\| \leq |u'Mu| - |\tilde{u}'M\tilde{u}| \leq \|u-\tilde{u}\|\|M\|\|u+\tilde{u}\|$$

Let $S_{1/8}^{m-1}$ be a 1/8 net of the unit sphere S^{m-1} in Euclidean distance in R^m. We have

$$\|M\| \leq \sup_{\|u\|\in S^{m-1}} |u'Mu| \leq \max_{\tilde{u}\in S_{1/8}^{m-1}} |\tilde{u}'M\tilde{u}| + \frac{1}{4}\|M\| \tag{9.48}$$

$$\|M\| \geq \inf_{\|u\|\in S^{m-1}} |u'Mu| \geq \min_{\tilde{u}\in S_{1/8}^{m-1}} |\tilde{u}'M\tilde{u}| - \frac{1}{4}\|M\|, \tag{9.49}$$

$$\text{thus,} \quad \sup_{u\in S^{m-1}} |u'Mu| \leq \frac{4}{3} \max_{\tilde{u}\in S_{1/8}^{m-1}} |\tilde{u}'M\tilde{u}|, \tag{9.50}$$

$$\inf_{u\in S^{m-1}} |u'Mu| \geq \min_{\tilde{u}\in S_{1/8}^{m-1}} |\tilde{u}'M\tilde{u}| - \frac{1}{4} \sup_{u\in S^{m+1}} |u'Mu| \tag{9.51}$$

where d_0 and ϵ_1 are both constants determined by ρ_* and ρ^*, and d_0^m is the upper bound of the cardinality of net $S_{r_0}^{m-1} \in S^{m-1}$, the unit sphere. Thus,

$$P(\max_{\|u\|^2\leq c_0} \frac{\|X_{A,m}u\|}{\sqrt{n}} \geq \sqrt{\tau^*\rho^*}\|u\|) \tag{9.52}$$

$$\leq P(\max_{\|u\|^2\leq c_0} \frac{\|X_m u\|}{\sqrt{n}} \geq \sqrt{\tau^*\rho^*}\|u\|) \tag{9.53}$$

$$\leq (1+\frac{2}{1/8})^m P\left(\frac{\|X_m\tilde{u}\|}{\sqrt{n}\|\tilde{u}\|} \geq \sqrt{\tau^*\rho^*}\right) \tag{9.54}$$

$$= 17^m P\left(\frac{\|X_m\tilde{u}\|^2}{n\|\tilde{u}\|^2} \geq \tau^*\rho^*\right) \tag{9.55}$$

Since the rows of X are drawn from the sub-Gaussian/Gaussian distributions, $\frac{\|X_m\tilde{u}\|^2}{n\|\tilde{u}\|^2}$ is the average of n i.i.d random variables with finite mean and bounded variance.

Lemma 9.4.1. *Suppose a random variable $x \sim N(0, \sigma^2)$, then the following large deviation holds:*

$$P(|x| > t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$$

Proof.

$$\begin{aligned} & P(x > t) = P(\lambda x > \lambda t) \leq e^{-\lambda t} E(e^{\lambda x}) \\ = \quad & e^{-\lambda t + \lambda^2 \sigma^2 / 2} \leq e^{\frac{1}{2}(\lambda\sigma - \frac{t}{\sigma})^2 - \frac{t^2}{2\sigma^2}} \end{aligned}$$

Thus if we take λ such that $\lambda\sigma^2 = t$, then we have

$$P(x > t) \leq e^{-\frac{t^2}{2\sigma^2}}.$$

Similarly,

$$P(x < -t) \quad = \quad P(-\lambda x > \lambda t) \leq e^{-\lambda t} E(e^{-\lambda x}) \leq e^{-\frac{t^2}{2\sigma^2}}$$

□

Lemma 9.4.2. *Assume $x_1, \cdots, x_n$ are n i.i.d normal random variable with mean zero and variance σ^2, then $\forall 0 < t < \sigma^2$,*

$$P\left(\frac{1}{n}|\sum_{i=1}^{n} x_i^2 - \sigma^2| > t\right) \leq 2e^{-\frac{nt^2}{4\sigma^4}}.$$

Proof. Denote $Y = X^2$, where $X \sim N(0, 1)$. Then $\mu_y = E(Y) = 1$. We can show the tail probability of Y.

Next we will derive the large deviation inequality for squared normal random variable with the use of moment generating function. For any $0 < t < 1$, the tail probability

$$P(y > 1 + t) = P(\lambda y > \lambda t + \lambda) \leq \frac{E(e^{\lambda y})}{e^{\lambda t + \lambda}}$$

By

$$E \exp(\lambda y) = \frac{1}{\sqrt{1 - 2\lambda}},$$

we have

$$\begin{aligned} & P(y > 1 + t) \leq \frac{e^{-\lambda t - \lambda}}{\sqrt{1 - 2\lambda}} \\ = \quad & \exp\left\{-(\lambda(t + 1) + \frac{1}{2}\ln(1 - 2\lambda))\right\} \end{aligned} \tag{9.56}$$

The inequality above holds for $\forall\lambda$, therefore

$$P(y > 1+t) \leq \inf_{\lambda} \exp\left\{-(\lambda(t+1) + \frac{1}{2}\ln(1-2\lambda))\right\}$$

Take the derivative in the exponential term with respect to λ, we have

$$\lambda = \frac{1}{2} - \frac{1}{2(t+1)} = \frac{t}{2(t+1)}.$$

Plug it into (9.56) and expand $\ln(1+x)$ around $x=0$ by Taylor expansion, we have

$$\begin{aligned} & \lambda(t+1) + \frac{1}{2}\ln(1-2\lambda) \\ = \quad & \frac{t}{2} + \frac{1}{2}\ln(1-2\lambda) = \frac{t}{2} - \frac{1}{2}\ln t + 1 \\ \approx \quad & \frac{t}{2} - \frac{t}{2} + \frac{t^2}{4} = \frac{t^2}{4} \end{aligned}$$

Thus

$$P(y > 1+t) \leq \inf_{\lambda} \exp\left\{-(\lambda(t+1) + \frac{1}{2}\ln(1-2\lambda))\right\} = e^{-\frac{t^2}{4}}$$

For i.i.d random samples $y_1, \cdots, y_n$ drawn from chi-square distribution, the large deviation inequality for their average is as follows:

$$\begin{aligned} & P\left\{\frac{1}{n}(y_1 + \cdots + y_n) - 1 > t\right\} = P\left(\lambda(y_1 + \cdots + y_n) > \lambda n(t+1)\right) \\ \leq \quad & \frac{E(e^{\lambda(y_1+\cdots y_n)})}{e^{\lambda n(t+1)}} \leq \left(\frac{e^{-\lambda(t+1)}}{\sqrt{1-2\lambda}}\right)^n \leq e^{-\frac{nt^2}{4}} \end{aligned}$$

Therefore if x is a normal random variable with mean of 0 and variance of σ^2, then we have $\forall\ 0 < t < \sigma^2$,

$$P(\frac{x_1^2 + \cdots + x_n^2}{n} - \sigma^2 > t) \leq e^{-\frac{nt^2}{4\sigma^4}}$$

Similarly, we can also show that

$$P(\frac{x_1^2 + \cdots + x_n^2}{n} - \sigma^2 < -t) \leq \ e^{-\frac{nt^2}{4\sigma^4}}$$

Thus $\forall\ 0 < t < \sigma^2$,

$$P(|\frac{x_1^2 + \cdots + x_n^2}{n} - \sigma^2| > t) \leq 2e^{-\frac{nt^2}{4\sigma^4}}$$

□

9.4.3 Case 1: $\|\mathbf{u}\|^2 < \mathbf{c_0}$

Following (9.46), we know that

$$f(u,v) \geq \left(\sqrt{1-\|u\|^2} - \|\frac{z_{A^*}}{\sqrt{n}}\|\right)^2 \tag{9.57}$$

Thus

$$P\{\min_{\|u\|^2\leq c_0, v} f(u,v) \leq t_1\} \leq P\left\{\min_{\|u\|^2\leq c_0}\left(\sqrt{1-\|u\|^2} - \|\frac{z_{A^*}}{\sqrt{n}}\|\right)^2 \leq t_1\right\} \tag{9.58}$$

Now we will apply lemma 9.4.2 to (9.55). For any vector $a \in R^m$, the variance of $a'X_m^{(i)}$ is $a'\Sigma a$.

Now consider

$$\frac{a'X_m'X_m a}{n} = \frac{\sum_{i=1}^n a'(X_m^{(i)})'X_m^{(i)}a}{n},$$

which is the average of n i.i.d random variables. Each random variable is generated by taking the square of a normal random variable. Thus the average has expectation of the individual expectaion

$$E(a'(X_m^{(i)})'X_m^{(i)}a) = a'\Sigma a.$$

Since $\tilde{u}'\Sigma\tilde{u} \leq \rho^*$, $\forall \tau^* > 1$, we turn to consider the probability on the right side of the formula below

$$P\left(\frac{1}{n}\tilde{u}'X_m'X_m\tilde{u} > \tau^*\rho^*\right) \leq P\left(\frac{1}{n}\tilde{u}'(X_m'X_m - \Sigma)\tilde{u} > (\tau^*-1)\tilde{u}'\Sigma\tilde{u}\right)$$

By lemma 9.4.2, we have

$$\begin{aligned} & P\left(\frac{1}{n}\tilde{u}'(X_m'X_m - \Sigma)\tilde{u} > (\tau^*-1)\tilde{u}'\Sigma\tilde{u}\right) \\ \leq \quad & \exp\left(-\frac{n(\tau^*-1)^2(\tilde{u}'\Sigma\tilde{u})^2}{4(\tilde{u}'\Sigma\tilde{u})^2}\right) = \exp\left(-\frac{n(\tau^*-1)^2}{4}\right) \end{aligned}$$

Now we have shown that, $\forall\ 1 < \tau^* < 2$,

$$P\left(\frac{1}{n}\tilde{u}'X_m'X_m\tilde{u} > \tau^*\rho^*\right) \leq \exp\left(-\frac{n(\tau^*-1)^2}{4}\right). \tag{9.59}$$

Thus for any $u \in S^{m-1}$,

$$P\left(\frac{1}{n}u'X_m'X_m u > \tau^*\rho^*\right) \leq 17^m \exp\left(-\frac{n(\tau^*-1)^2}{4}\right). \tag{9.60}$$

and for any $u \in R^m$. Specifically, for any $\|u\| < c_0$ and $u \in S^{m-1}$,

$$P\left(\frac{u'X_m'X_m u}{n\|u\|^2} > \tau^*\rho^*\right) \le \exp\left(-\frac{n(\tau^*-1)^2}{4}\right). \tag{9.61}$$

Use the above large deviation inequality in (9.61), we can obtain the bound for (9.58),

$$\begin{aligned}
& P\left\{\min_{\|u\|^2\le c_0,v}\left(\sqrt{1-\|u\|^2}-\|\frac{z_{A^*}}{\sqrt{n}}\|\right)^2 < t_1\right\} && (9.62)\\
= \;& P\left\{\min_{\|u\|^2\le c_0,v}\left(\sqrt{1-\|u\|^2}-\|\frac{z_{A^*}}{\sqrt{n}}\|\right)^2 < t_1, \max_{\|u\|^2\le c_0}\frac{\|z\|}{\sqrt{n}\|u\|} > \sqrt{\tau^*\rho^*}\right\}\\
& +P\left\{\min_{\|u\|^2\le c_0,v}\left(\sqrt{1-\|u\|^2}-\frac{\|z_{A^*}\|}{\sqrt{n}}\right)^2 < t_1, \max_{\|u\|^2\le c_0}\frac{\|z\|}{\sqrt{n}\|u\|} \le \sqrt{\tau^*\rho^*}\right\}\\
\le \;& P\left\{\max_{\|u\|^2\le c_0}\frac{\|z\|}{\sqrt{n}\|u\|} > \sqrt{\tau^*\rho^*}\right\} +\\
P \;& \left\{\min_{\|u\|^2\le c_0,v}\left(\sqrt{1-\|u\|^2}-\frac{\|z_{A^*}\|}{\sqrt{n}}\right)^2 < t_1, \max_{\|u\|^2\le c_0}\frac{\|z\|}{\sqrt{n}\|u\|} \le \sqrt{\tau^*\rho^*}\right\} && (9.63)
\end{aligned}$$

By (9.61), the first term in (9.63) above is bounded by $(17^m)\exp\left(-\frac{n(\tau^*-1)^2}{4}\right)$. Next will show that the second term vanishes when τ^* takes suitable values.

In the event that $\|z\|/\sqrt{n} \le \sqrt{\tau^*\rho^*}\|u\|$, and given that

$$\|u\| \le \sqrt{c_0} \le 1/\sqrt{(1+\tau^*\rho^*)}, \tag{9.64}$$

we have

$$\sqrt{1-\|u\|^2} \ge \sqrt{1-c_0} \ge \sqrt{\tau^*\rho^* c_0} \ge \frac{\|z\|}{\sqrt{n}} \ge \frac{\|z_{A^*}\|}{\sqrt{n}}.$$

Thus

$$\min_{\|u\|\le c_0,v}\left(\sqrt{1-\|u\|^2}-\frac{\|z_{A^*}\|}{\sqrt{n}}\right)^2 \ge \left(\sqrt{1-c_0}-\sqrt{\tau^*\rho^* c_0}\right)^2.$$

If the following equality holds:

$$\sqrt{1-c_0}-\sqrt{\tau^*\rho^* c_0} = \sqrt{t_1}, \tag{9.65}$$

i.e., we take τ^* with the value

$$\tau^* = \frac{(\sqrt{1-c_0}-\sqrt{t_1})^2}{\rho^* c_0} \tag{9.66}$$

Then the second term in (9.63) vanishes.

The probability in (9.58) now is bounded by

$$17^m \exp\left(-\frac{n(\tau^*-1)^2}{4}\right) = 17^m \exp\left(-\frac{n}{4}\left(\frac{(\sqrt{1-c_0}-\sqrt{t_1})^2}{\rho^* c_0}-1\right)^2\right).$$

Remark 1. When c_0 is large, the probability in the above formula is large. Thus the worst case occurs when c_0 is taken the largest value.

Remark 2. When $\rho^* < t_1 < 1$, algebra shows that $\omega_2 < 1/(1+\rho^*)$. When $\rho^* >> 1 > t_1$, then $\omega_2 > 1/(1+\rho^*)$.

Remark 3. To summarize, in case 1, given P_m, we have shown the following proposition:

Proposition 9.4.3. *Need to revise!!!For any positive constant $t_1 < \rho^*/(1+\rho^*)$, and $\forall\ c_0 > 0$ which satisfies:*

$$c_0 < (\omega_1 \wedge \frac{1}{1+\rho^*}) \text{ or } \omega_2 < c_0 < \frac{1}{1+\rho^*},$$

where

$$\omega_1 = \frac{(\sqrt{\rho^*+1-t_1}-\sqrt{\rho^* t_1})^2}{(\rho^*-1)^2}, \quad \omega_2 = \frac{(\sqrt{\rho^*+1-t_1}+\sqrt{\rho^* t_1})^2}{(\rho^*-1)^2}$$

We have

$$P\left\{\min_{\|u\|^2\le c_0, v} f(u,v) \le t_1\right\} \le 17^m \exp\left(-\frac{n}{4}\left(\frac{(\sqrt{1-c_0}-\sqrt{t_1})^2}{\rho^* c_0}-1\right)^2\right) \tag{9.67}$$

Case 2: $\|\mathbf{u}\|^2 \ge \mathbf{c_0}$

If $\|u\| \ge \sqrt{c_0}$, then for a fixed P_m, the target probability satisfies:

$$P\left\{\min_{\|u\|^2\ge c_0, v} f(u,v) \le t_1\right\} \le P\left\{\min_{\|u\|^2\ge c_0, v}\min_{|A|=\alpha n} \frac{\|z_{(A^*)^c}\|^2}{n} \le t_1\right\} \tag{9.68}$$

$$\le \quad P\left\{\min_{\|u\|\in S^{m-1}}\min_{|A|=\alpha n} \frac{\|X_{(A^*)^c,m} u\|^2}{n} \le \frac{t_1}{c_0}\right\}. \tag{9.69}$$

For a given $\tilde{u} \in S^{m-1}_{1/8}$, denote

$$\tilde{z} = X_{(A^*)^c}\tilde{u} = (z_1, \cdots, z_n)'.$$

Order z_i's by absolute values

$$|z_{[1]}| \le |z_{[2]}| \le \cdots \le |z_{[n]}|$$

We now look for the bound in (9.69) based on the coverings and the 1/8 radius net. Recall the two inequalities (9.50) and (9.51), $\forall\ m \times m$ matrix M,

$$\sup_{u \in S^{m-1}} |u'Mu| \le \frac{4}{3} \max_{\tilde{u} \in S_{1/8}^{m-1}} |\tilde{u}'M\tilde{u}|,$$
$$\inf_{u \in S^{m-1}} |u'Mu| \ge \min_{\tilde{u} \in S_{1/8}^{m-1}} |\tilde{u}'M\tilde{u}| - \frac{1}{4} \sup_{u \in S^{m+1}} |u'Mu|$$

Denote $m \times m$ matrix $M_{(A^*)^c} = \dfrac{X'_{(A^*)^c} X_{(A^*)^c}}{(1-\alpha)n}$. For any two constants $0 < t^*/3 < t_* < t^*$,

$$\begin{aligned}
& P\left\{ \min_{|(A^*)^c|=(1-\alpha)n} \min_{\|u\|=1} \frac{\|X_{(A^*)^c,m} u\|^2}{(1-\alpha)n} \le t_* - t^*/3 \right\} \\
\le\ & P\left\{ \min_{|(A^*)^c|=(1-\alpha)n} \left(\min_{\tilde{u} \in S_{1/8}^{m-1}} |\tilde{u}' M_{(A^*)^c} \tilde{u}| - \frac{1}{4} \sup_{u \in S^{m-1}} |u' M_{(A^*)^c} u| \right) \le t_* - t^*/3 \right\} \\
\le\ & P\left\{ \min_{|(A^*)^c|=(1-\alpha)n} \left(\min_{\tilde{u} \in S_{1/8}^{m-1}} |\tilde{u}' M_{(A^*)^c} \tilde{u}| - \frac{1}{3} \max_{\tilde{u} \in S_{1/8}^{m-1}} |\tilde{u}' M_{(A^*)^c} \tilde{u}| \right) \le t_* - t^*/3 \right\} \\
\le\ & P\left\{ \min_{|(A^*)^c|=(1-\alpha)n} \min_{\tilde{u} \in S_{1/8}^{m-1}} |\tilde{u}' M_{(A^*)^c} \tilde{u}| - \frac{1}{3} \max_{|(A^*)^c|=(1-\alpha)n} \max_{\tilde{u} \in S_{1/8}^{m-1}} |\tilde{u}' M_{(A^*)^c} \tilde{u}| \le t_* - t^*/3 \right\} \\
\le\ & P\left\{ \min_{|(A^*)^c|=(1-\alpha)n} \min_{\tilde{u} \in S_{1/8}^{m-1}} |\tilde{u}' M_{(A^*)^c} \tilde{u}| \le t_* \right\} + P\left\{ \max_{|(A^*)^c|=(1-\alpha)n} \max_{\tilde{u} \in S_{1/8}^{m-1}} |\tilde{u}' M_{(A^*)^c} \tilde{u}| \ge t^* \right\} \\
\le\ & P\left\{ \min_{\tilde{u} \in S_{1/8}^{m-1}} \frac{\sum_{i=1}^{(1-\alpha)n} z_{[i]}^2}{(1-\alpha)n} \le t_* \right\} + P\left\{ \max_{\tilde{u} \in S_{1/8}^{m-1}} \frac{\sum_{i=\alpha n+1}^{n} z_{[i]}^2}{(1-\alpha)n} \ge t^* \right\} \\
\le\ & |S_{1/8}^{m-1}| \left(P\left\{ \frac{\sum_{i=1}^{(1-\alpha)n} z_{[i]}^2}{(1-\alpha)n} \le t_* \right\} + P\left\{ \frac{\sum_{i=\alpha n+1}^{n} z_{[i]}^2}{(1-\alpha)n} \ge t^* \right\} \right) \qquad (9.70)
\end{aligned}$$

Denote

$$T_{n,\alpha,-} = \frac{\sum_{i=1}^{(1-\alpha)n} z_{[i]}^2}{(1-\alpha)n\tilde{u}'\Sigma_m\tilde{u}},$$
$$T_{n,\alpha,+} = \frac{\sum_{i=\alpha n+1}^{n} z_{[i]}^2}{(1-\alpha)n\tilde{u}'\Sigma_m\tilde{u}}.$$

Since $z_i = X_m^{(i)}\tilde{u}$ all has zero expectation and variance of $\tilde{u}'\Sigma\tilde{u}$. Thus $T_{n,\alpha,-}$ and $T_{n,\alpha,+}$ are both now standardized with mean zero and unit variance.

Stigler's theorem (Stigler, 1973)shows that

$$\sqrt{n}(T_{n,\alpha,-} - \mu_{\alpha,-}) \xrightarrow{\mathcal{L}} N(0, \frac{\sigma^2_{\alpha,-}}{1-\alpha}), \tag{9.71}$$

$$\sqrt{n}(T_{n,\alpha,+} - \mu_{\alpha,+}) \xrightarrow{\mathcal{L}} N(0, \frac{\sigma^2_{\alpha,+}}{1-\alpha}). \tag{9.72}$$

Thus following convergence holds ,$\forall t \geq 0$, when $n \to \infty$,

$$P\{\sqrt{(1-\alpha)n}(\frac{T_{n,\alpha,-} - \mu_{\alpha,-}}{\sigma_{\alpha,-}}) < -t\} \to 1 - \Phi(t) \tag{9.73}$$

$$P\{\sqrt{(1-\alpha)n}(\frac{T_{n,\alpha,+} - \mu_{\alpha,+}}{\sigma_{\alpha,+}}) > t\} \to 1 - \Phi(t) \tag{9.74}$$

where

$$\mu_{\alpha,-} = \frac{1}{1-\alpha}\int_0^{1-\alpha} Q(x)dx \tag{9.75}$$

$$\sigma^2_{\alpha,-} = \frac{1}{1-\alpha}\int_0^{1-\alpha} Q^2(x)dx - (\mu_{\alpha,-})^2 \tag{9.76}$$

$$\mu_{\alpha,+} = \frac{1}{1-\alpha}\int_\alpha^{1} Q(x)dx \tag{9.77}$$

$$\sigma^2_{\alpha,+} = \frac{1}{1-\alpha}\int_\alpha^{1} Q^2(x)dx - (\mu_{\alpha,+})^2 \tag{9.78}$$

Here Q(x) is its inverse function or quantile function of χ^2_1, the chi-square distribution with degree of freedom 1.

Now going back to the two probabilities in (9.70),

$$\begin{aligned} & P\left\{\frac{\sum_{i=1}^{(1-\alpha)n} z^2_{[i]}}{(1-\alpha)n} \leq t_*\right\} \leq P\left\{\frac{\sum_{i=1}^{(1-\alpha)n} z^2_{[i]}}{(1-\alpha)n\tilde{u}'\Sigma_m\tilde{u}} \leq \frac{t_*}{\rho_*}\right\} \\ = \quad & P\left\{T_{n,\alpha,-} \leq \frac{t_*}{\rho_*}\right\} \leq (1+o(1))\exp\left\{-\frac{n(1-\alpha)(\mu_{\alpha,-} - t_*/\rho_*)^2}{2\sigma^2_{\alpha,-}}\right\} \end{aligned} \tag{9.79}$$

The first inequality is due to the assumption that Σ has eigen values bounded between ρ^* and ρ_*. Similarly, we have

$$\begin{aligned} & P\left\{\frac{\sum_{i=\alpha n+1}^{n} z^2_{[i]}}{(1-\alpha)n} \geq t^*\right\} \leq P\left\{\frac{\sum_{i=\alpha n+1}^{n} z^2_{[i]}}{(1-\alpha)n\tilde{u}'\Sigma_m\tilde{u}} \geq \frac{t_*}{\rho^*}\right\} \\ = \quad & P\left\{T_{n,\alpha,+} \geq \frac{t^*}{\rho^*}\right\} \leq (1+o(1))\exp\left\{-\frac{n(1-\alpha)(t^*/\rho^* - \mu_{\alpha,+})^2}{2\sigma^2_{\alpha,+}}\right\} \end{aligned} \tag{9.80}$$

Let ϵ_2 and ϵ_3 be constants such that:

$$0 < \epsilon_2 < 1, \ \epsilon_3 > 1.$$

For t_* and t^* satisfies that:

$$t_* = \epsilon_2 \rho_* \mu_{\alpha,-},\ t^* = \epsilon_3 \rho^* \mu_{\alpha,+},$$

$$(1-\alpha)c_0(t_* - t^*/3) = t_1. \tag{9.81}$$

Together with (9.79) and (9.80), it follows that

$$\begin{aligned}
& P\left\{\min_{\|u\|^2 \geq c_0, v} f(u,v) \leq t_1\right\} \leq P\left\{\min_{\|u\| \in S^{m-1}} \min_{|A|=\alpha n} \frac{\|X_{(A^*)^c,m} u\|^2}{n} \leq \frac{t_1}{c_0}\right\} \\
= & P\left\{\min_{|(A^*)^c|=(1-\alpha)n} \min_{\|u\|=1} \frac{\|X_{(A^*)^c,m} u\|^2}{(1-\alpha)n} \leq t_* - t^*/3\right\} \\
\leq & |S_{r_0}^{m-1}| \left(P\left\{\frac{\sum_{i=1}^{(1-\alpha)n} z_{[i]}^2}{(1-\alpha)n} \leq t_*\right\} + P\left\{\frac{\sum_{i=\alpha n+1}^{n} z_{[i]}^2}{(1-\alpha)n} \geq t^*\right\}\right) \qquad (9.82) \\
\leq & |S_{r_0}^{m-1}| \left(\exp\left\{-\frac{n(1-\alpha)^2(\mu_{\alpha,-} - t_*/\rho_*)^2}{2\sigma_{\alpha,-}^2}\right\} + \exp\left\{-\frac{n(1-\alpha)^2(t^*/\rho^* - \mu_{\alpha,+})^2}{2\sigma_{\alpha,+}^2}\right\}\right) \\
\leq & 2 \times 17^m \exp\left(-\frac{n(1-\alpha)^2 t_\alpha^2}{2}\right) \qquad (9.83)
\end{aligned}$$

where

$$\begin{aligned}
t_\alpha &= \frac{(\mu_{\alpha,-} - t_*/\rho_*)}{\sigma_{\alpha,-}} \wedge \frac{(t^*/\rho^* - \mu_{\alpha,+})}{\sigma_{\alpha,+}} \\
&= \frac{(1-\epsilon_2)\mu_{\alpha,-}}{\sigma_{\alpha,-}} \wedge \frac{(\epsilon_3 - 1)\mu_{\alpha,+}}{\sigma_{\alpha,+}}
\end{aligned}$$

When α is given, the sharpest bound in (9.83) is obtained when t_α is achieved its maximum. By (9.81), we know that ϵ_2 and ϵ_3 are collinear and the correlation is positive. $(1-\epsilon_2)\mu_{\alpha,-}/\sigma_{\alpha,-}$ is decreasing with ϵ_2, and $(\epsilon_3 - 1)\mu_{\alpha,+}/\sigma_{\alpha,+}$ is increasing with ϵ_3, equivalently, with ϵ_2.

When ϵ_3 is close to 1, ϵ_2 is small, and $t_\alpha = (\epsilon_3 - 1)\mu_{\alpha,+}/\sigma_{\alpha,+}$, which is smaller than $(1-\epsilon_2)\mu_{\alpha,-}/\sigma_{\alpha,-}$. When ϵ_2 and ϵ_3 are both getting larger, $t_\alpha = (\epsilon_3 - 1)\mu_{\alpha,+}/\sigma_{\alpha,+}$ until

$$\frac{(1-\epsilon_2)\mu_{\alpha,-}}{\sigma_{\alpha,-}} = \frac{(\epsilon_3 - 1)\mu_{\alpha,+}}{\sigma_{\alpha,+}}.$$

If ϵ_2 and ϵ_3 continue to increase, $(1-\epsilon_2)\mu_{\alpha,-}/\sigma_{\alpha,-}$ would be smaller than $(\epsilon_3 - 1)\mu_{\alpha,+}/\sigma_{\alpha,+}$. Thus $t_\alpha = (1-\epsilon_2)\mu_{\alpha,-}/\sigma_{\alpha,-}$. The value t_α then decreases when ϵ_2 is approaching 1.

Therefore, we have shown that the maximum value of t_α is achieved when

$$\frac{(\mu_{\alpha,-} - t_*/\rho_*)}{\sigma_{\alpha,-}} = \frac{(t^*/\rho^* - \mu_{\alpha,+})}{\sigma_{\alpha,+}}$$

Using a more generalized notation ϵ_0 to replace the fraction $1/3$ in (9.81)(see the remark in the end of this subsection), we can solve

$$\begin{aligned} t_* &= \frac{\epsilon_0\rho^*\rho_*(\sigma_{\alpha,+}\mu_{\alpha,-} + \sigma_{\alpha,-}\mu_{\alpha,+}) + \frac{t_1}{c_0}\rho_*\sigma_{\alpha,-}/(1-\alpha)}{\epsilon_0\rho^*\sigma_{\alpha,+} + \rho_*\sigma_{\alpha,-}} && (9.84)\\ t^* &= \frac{\rho^*\rho_*(\sigma_{\alpha,+}\mu_{\alpha,-} + \sigma_{\alpha,-}\mu_{\alpha,+}) - \frac{t_1}{c_0}\rho^*\sigma_{\alpha,+}/(1-\alpha)}{\epsilon_0\rho^*\sigma_{\alpha,+} + \rho_*\sigma_{\alpha,-}} && (9.85)\\ t_\alpha &= \frac{\epsilon_0\rho^*(\sigma_{\alpha,+}\mu_{\alpha,-} + \sigma_{\alpha,-}\mu_{\alpha,+}) - \frac{t_1}{c_0}\sigma_{\alpha,-}/(1-\alpha)}{\sigma_{\alpha,-}(\epsilon_0\rho^*\sigma_{\alpha,+} + \rho_*\sigma_{\alpha,-})} && (9.86) \end{aligned}$$

It is easy to check that $t_* - \epsilon_0 t^* > 0$.

If $\sigma_{\alpha,-} < \sigma_{\alpha,+}$ holds, then a sufficient condition for $t_* < t^*$ is:

$$\begin{aligned} \iff \epsilon_0 &\leq 1 - \frac{2t_1}{c_0\rho_*}c_+^*(\alpha), && (9.87)\\ \text{where } c_+^*(\alpha) &= \frac{\sigma_{\alpha,+}}{(1-\alpha)(\sigma_{\alpha,+}\mu_{\alpha,-} + \sigma_{\alpha,-}\mu_{\alpha,+})}. && (9.88) \end{aligned}$$

Lemma (**??**) shows $\forall\alpha > 0, \sigma_{\alpha,-} < \sigma_{\alpha,+}$. Thus,

$$\frac{(\rho^*\sigma_{\alpha,+} + \rho_*\sigma_{\alpha,-})t_1}{(1-\alpha)c_0} < \frac{2\rho^*\sigma_{\alpha,+}t_1}{(1-\alpha)c_0} \leq (1-\epsilon_0)\rho^*\rho_*(\sigma_{\alpha,+}\mu_{\alpha,-} + \sigma_{\alpha,-}\mu_{\sigma,+})$$

Therefore we obtain

$$\begin{aligned} & t_*\epsilon_0\rho^*\sigma_{\alpha,+} = \epsilon_0\rho^*\rho_*(\sigma_{\alpha,+}\mu_{\alpha,-} + \sigma_{\alpha,-}\mu_{\alpha,+}) + \frac{t_1\rho_*\sigma_{\alpha,-}}{c_0(1-\alpha)}\\ < \quad & \rho^*\rho_*(\sigma_{\alpha,+}\mu_{\alpha,-} + \sigma_{\alpha,-}\mu_{\alpha,+}) - \frac{t_1\rho^*\sigma_{\alpha,+}}{(1-\alpha)c_0} = t^*\epsilon_0\rho^*\sigma_{\alpha,+}\\ \iff \quad & t_* < t^*. \end{aligned}$$

Remark 1. *Here we take the net with radius* $1/8$ *and we obtain the fractions* $1/4$ *in (9.51) and thus* $1/3$ *in the first row in (9.70) and afterwards. If we choose another value for radius, then this fraction would be changed accordingly. As for a generalization, we use* ϵ_0 *to replace the faction* $1/3$*, and the radius and cardinality of the net now are changed to*

$$r_0 = \frac{\epsilon_0}{2(\epsilon_0+1)},\ |S_{r_0}^{m-1}| \leq (5 + \frac{4}{\epsilon_0})^m.$$

This also shows the choice of r_0 could be any real number lies in (0, 0.5). And ϵ_0 could be ant positive real number.

Remark 2. *$\mu_{\alpha,-}$ is a decreasing function of α, $\mu_{\alpha,+}$ is an increasing function of α. Furthermore, $\forall 0 < \alpha < 1$,*

$$0 = \mu_{1,-} < \mu_{\alpha,-} \leq \mu_{0,-} = 1 = \mu_{0,+} < \mu_{\alpha,+} \leq \mu_{1,+} = \infty$$

The two tails are obtained by using L'Hopital's Rule.

The derivation is to take the derivative with respect to α:

$$\frac{\partial \mu_{\alpha,+}}{\partial \alpha} = \frac{1}{(1-\alpha)^2}\left(\int_{\alpha}^{1} Q(x)dx - (1-\alpha)Q(\alpha)\right) > 0,$$

since $Q(x)$ is increasing with x. Similarly

$$\frac{\partial \mu_{\alpha,-}}{\partial \alpha} = \frac{1}{(1-\alpha)^2}\left(\int_{0}^{1-\alpha} Q(x)dx - (1-\alpha)Q(1-\alpha)\right) < 0.$$

To summarize, in this subsection, we have shown that: given projection P_m, ρ^*, ρ_*, if the design matrix has contamination portion α, then $\forall t_1 > 0$ and $\forall c_0$ and ϵ_0, such that $0 < c_0 < 1$, and

$$0 < \epsilon_0 \leq 1 - \frac{2t_1}{c_0\rho_*}c_+^*(\alpha),$$

then

$$P\left\{\min_{\|u\|^2 \geq c_0, v} f(u,v) \leq t_1\right\} \leq 2(5 + \frac{4}{\epsilon_0})^m \exp\left(-\frac{n(1-\alpha)^2 t_\alpha^2}{2}\right),$$

where

$$t_\alpha = \frac{\epsilon_0\rho^*(\sigma_{\alpha,+}\mu_{\alpha,-} + \sigma_{\alpha,-}\mu_{\alpha,+}) - \frac{t_1}{c_0}\sigma_{\alpha,-}/(1-\alpha)}{\sigma_{\alpha,-}(\epsilon_0\rho^*\sigma_{\alpha,+} + \rho_*\sigma_{\alpha,-})}. \tag{9.89}$$

9.4.4 The value of parameters: c_0 and ϵ_0.

If we denote

$$A_\alpha = \frac{1}{(1-\alpha)(\epsilon_0\rho^*\sigma_{\alpha,+} + \rho_*\sigma_{\alpha,-})} \tag{9.90}$$

$$B_\alpha = \frac{\epsilon_0\rho^*(\sigma_{\alpha,+}\mu_{\alpha,-} + \sigma_{\alpha,-}\mu_{\alpha,+})}{\sigma_{\alpha,-}(\epsilon_0\rho^*\sigma_{\alpha,+} + \rho_*\sigma_{\alpha,-})} \tag{9.91}$$

$$t_\alpha = -A_\alpha\frac{t_1}{c_0} + B_\alpha > 0 \tag{9.92}$$

From (9.92), we have

$$c_0 \quad > \quad \frac{A_\alpha}{B_\alpha} t_1 = \frac{\sigma_{\alpha,-}}{(1-\alpha)\epsilon_0 \rho^* (\sigma_{\alpha,+}\mu_{\alpha,-} + \sigma_{\alpha,-}\mu_{\alpha,+})} t_1 \tag{9.93}$$

Denote

$$c_-^*(\alpha) = \frac{\sigma_{\alpha,-}}{(1-\alpha)(\sigma_{\alpha,+}\mu_{\alpha,-} + \sigma_{\alpha,-}\mu_{\alpha,+})}.$$

Then

$$c_0 > \frac{c_-^*(\alpha) t_1}{\epsilon_0 \rho^*} \tag{9.94}$$

Together with (9.87), we obtain

$$\frac{c_-^*(\alpha) t_1}{\rho^* c_0} < \epsilon_0 \leq 1 - \frac{2c_+^*(\alpha) t_1}{\rho_* c_0}, \tag{9.95}$$

which requires that

$$\frac{t_1}{c_0} < \left(\frac{c_-^*(\alpha)}{\rho^*} + \frac{2c_+^*(\alpha)}{\rho_*} \right)^{-1}.$$

Thus, given t_1, c_0 can take any number that is greater than $\left(\frac{c_-^*(\alpha)}{\rho^*} + \frac{2c_+^*(\alpha)}{\rho_*} \right) t_1$, for example, let $c_0 = 3c_+^*(\alpha) t_1 / \rho_*$.

Then we obtain that

$$\frac{\rho_* c_-^*(\alpha)}{3\rho^* c_+^*(\alpha)} < \epsilon_0 \leq \frac{1}{3},$$

Here we take the largest ϵ_0 given t_1/c_0 is due to the fact that given α, t_1 and c_0, t_α is an increasing function with respect to ϵ_0 in (9.89). The optimum bounded probability(minimum) is obtained when t_α and/or ϵ_0 reaches its maximum.

Remark. The range of $c_-^*(\alpha)$.

Figures **??** obtained from the simulation results shows how the value of $c_-^*(\alpha)$ changes along with α. The function $c_-^*(\alpha)$ is a not strictly increasing function with α. The order of $c_-^*(\alpha)$ is between 10^{-1} and 10^2 if $\alpha \leq 0.9999$. It is not hard to derive that if $\alpha \to 1$, $c_-^*(\alpha) \to \infty$ using L' Hopital's Rule. The plot of $c_-^*(\alpha)$ v.s. α shows that $c_-^*(\alpha)$ is decreasing slowly when α is ranged between 0 and 0.1, hitting the minimum 0.45 around $\alpha = 0.1$ and then increasing afterwards. The value of $c_-^*(\alpha)$ is between 0.45 and 0.51. When $\alpha < 0.1$, it lies between 0.45 and 1 when $\alpha < 0.8$.

9.4.5 Combining the two cases

For the convenience of the calculation, we can use the same net and covering for calculations in both case 1 and case 2. To summarize the both cases, we have shown that, for fixed P_m, if we take $\epsilon_0 = 1/3$, and subsequently let

$$r_0 = \frac{1}{2/\epsilon_0 + 2} = \frac{1}{8}, \text{ and } c_0 = \frac{3c_+^*(\alpha)t_1}{\rho_*}$$

In this part, we combine the two cases and show the main theorem of this dissertation:

Theorem 9.4.4. $\forall t_1 < \dfrac{\rho_*}{3c_+(\alpha) + (\sqrt{3c_+(\alpha)} + \sqrt{\rho_*})^2}$, *if* $\dfrac{\log \binom{p}{m}}{n} < C_\alpha^2$, *where* C_α *is defined as follows:*

$$\text{if } t_1 < t_1^*,\ C_\alpha = t_\alpha = \frac{(\sigma_{\alpha,+}\mu_{\alpha,-} + \sigma_{\alpha,-}\mu_{\alpha,+})(\rho^* - \rho_*/\sigma_{\alpha,+})}{\sigma_{\alpha,-}(\rho^*\sigma_{\alpha,+} + 3\rho_*\sigma_{\alpha,-})} \tag{9.96}$$

else,

$$C_\alpha = \frac{1}{2}\left(\frac{(\sqrt{1-c_0} - \sqrt{t_1})^2}{\rho^* c_0} - 1\right) = \frac{1}{2}\left(\left(\sqrt{\frac{1}{3c_+(\alpha)t_1} - \frac{1}{\rho^*}} - \sqrt{\frac{1}{3c_+(\alpha)}}\right)^2 - 1\right) \tag{9.97}$$

Here t_1^* *is the* t_1 *that satisfies the following:*

$$\frac{1}{2}\left(\left(\sqrt{\frac{1}{3c_+(\alpha)t_1} - \frac{1}{\rho^*}} - \sqrt{\frac{1}{3c_+(\alpha)}}\right)^2 - 1\right) = (1-\alpha)t_\alpha/\sqrt{2}$$

then

$$P\{\min_{P_m} \min_{\|u\|^2 + \|v\|_{A^*}^2 = 1} f(u,v) \le t_1\} \le 2\binom{p}{m} 17^m \exp\left(-nC_\alpha^2\right) \to 0$$

(9.83) can be re-written as:

$$P\left\{\min_{\|u\|^2 \ge c_0, v} f(u,v) \le t_1\right\} \le 2 \times 17^m \exp\left(-\frac{n(1-\alpha)^2 t_\alpha^2}{2}\right),$$

where

$$t_\alpha = \frac{(\sigma_{\alpha,+}\mu_{\alpha,-} + \sigma_{\alpha,-}\mu_{\alpha,+})(\rho^* - \rho_*/\sigma_{\alpha,+})}{\sigma_{\alpha,-}(\rho^*\sigma_{\alpha,+} + 3\rho_*\sigma_{\alpha,-})} \tag{9.98}$$

Lemma 9.4.5. $\sigma_{\alpha,+}^2$ *is an increasing function with* α.

Proof. Since $Q(x)$ is an increasing function with α,

$$\int_\alpha^1 Q^2(x)dx > (1-\alpha)Q^2(\alpha).$$

$$\begin{aligned} \Longrightarrow \qquad \frac{\partial \sigma_{\alpha,+}}{\partial \alpha} &= \frac{1}{(1-\alpha)^2}\int_\alpha^1 Q^2(x)dx - \frac{1}{1-\alpha}Q^2(\alpha) \\ + \quad \frac{2}{(1-\alpha)^3}&\left(\int_\alpha^1 Q(x)dx\right)^2 + \frac{2Q(\alpha)}{(1-\alpha)^2}\int_\alpha^1 Q(x)dx > 0. \end{aligned}$$

□

By Lemma 9.4.5, we have that $\sigma_{\alpha,+} \geq \sigma_{0,+} = \sqrt{2}$, thus $\rho^* - \rho_*/\sigma_{\alpha,+} > 0$,

$$t_\alpha \geq \frac{(\sigma_{\alpha,+}\mu_{\alpha,-} + \sigma_{\alpha,-}\mu_{\alpha,+})(1-1/\sigma_{\alpha,+})}{\sigma_{\alpha,-}(\rho^*\sigma_{\alpha,+} + 3\rho_*\sigma_{\alpha,-})}\rho_* \tag{9.99}$$

Next, combining the two cases, $\|u\|^2 \leq c_0$ and $\|u\|^2 > c_0$, we have

$$\begin{aligned} & P\{\min_{\|u\|^2+\|v\|^2_{A^*}=1} f(u,v) \leq t_1\} \\ \leq \quad & P\{\min_{u,v,\|u\|^2\leq c_0} f(u,v) \leq t_1\} \vee P\{\min_{u,v,\|u\|^2\geq c_0} f(u,v) \leq t_1\} \\ = \quad & 2\cdot 17^m \left[\exp\left(-\frac{n}{4}\left[\frac{(\sqrt{1-c_0}-\sqrt{t_1})^2}{\rho^* c_0} - 1\right]^2\right) \vee \exp\left(-\frac{n(1-\alpha)^2 t_\alpha^2}{2}\right)\right] \end{aligned}$$

The above inequality is for fixed P_m. Thus for all the enumerations of P_m, we have

$$P\{\min_{P_m}\min_{\|u\|^2+\|v\|^2_{A^*}=1} f(u,v) \leq t_1\} \leq 2\binom{p}{m}17^m \exp\left(-nC_\alpha^2\right) \tag{9.100}$$

$$\text{where } C_\alpha = \frac{1}{2}\left(\frac{(\sqrt{1-c_0}-\sqrt{t_1})^2}{\rho^* c_0} - 1\right) \wedge (1-\alpha)t_\alpha/\sqrt{2}. \tag{9.101}$$

Let $t_1 = \rho_*\epsilon_t$, as long as

$$\epsilon_t = \frac{t_1}{\rho_*} < \frac{1}{3c_+(\alpha) + (\sqrt{3c_+(\alpha)} + \sqrt{\rho_*})^2},$$

then

$$\tau^* = \frac{(\sqrt{1-c_0}-\sqrt{t_1})^2}{\rho^* c_0} > 1.$$

This is because

$$\frac{(\sqrt{1-c_0}-\sqrt{t_1})^2}{\rho^* c_0} > 1 \Longleftrightarrow \left(\sqrt{\frac{1}{c_0}-1}-\sqrt{\frac{t_1}{c_0}}\right)^2 > \rho^* \tag{9.102}$$

$$\Longleftrightarrow \quad \frac{\rho_*}{3c_+(\alpha)t_1} > 1+\left(\sqrt{\rho^*}+\sqrt{\frac{\rho_*}{3c_+(\alpha)}}\right)^2 \tag{9.103}$$

$$\Longleftrightarrow \quad t_1 < \frac{\rho_*}{3c_+(\alpha)+(\sqrt{3c_+(\alpha)}+\sqrt{\rho_*})^2} \tag{9.104}$$

Meanwhile, t_1 has to satisfy the following condition:

$$\sqrt{t_1} < \sqrt{1-c_0} \Longleftrightarrow t_1 < \frac{\rho^*}{\rho^*+3c_+(\alpha)}$$

If (9.104) holds, then the above inequality holds automatically.

Therefore, we have shown that $\forall \epsilon_t < \dfrac{1}{3c_+(\alpha)+(\sqrt{3c_+(\alpha)}+\sqrt{\rho_*})^2}$, where

$$c_+^*(\alpha) = \frac{\sigma_{\alpha,+}}{(1-\alpha)(\sigma_{\alpha,+}\mu_{\alpha,-}+\sigma_{\alpha,-}\mu_{\alpha,+})}, then$$

$$P\{\min_{P_m} \min_{\|u\|^2+\|v\|^2_{A^*}=1} f(u,v) \le t_1\} \le 2\binom{p}{m} 17^m \exp\left(-nC_\alpha^2\right)$$

$$\text{where } C_\alpha = \frac{1}{2}\left(\frac{(\sqrt{1-c_0}-\sqrt{t_1})^2}{\rho^* c_0}-1\right) \wedge (1-\alpha)t_\alpha/\sqrt{2}.$$

Since $\dfrac{1}{2}\left(\dfrac{(\sqrt{1-c_0}-\sqrt{t_1})^2}{\rho^* c_0}-1\right)$ is a decreasing function with c_0 and/or t_1, we know that

$$\text{if } t_1 < t_1^*,\ C_\alpha = \frac{(1-\alpha)t_\alpha}{\sqrt{2}} = \frac{(1-\alpha)(\sigma_{\alpha,+}\mu_{\alpha,-}+\sigma_{\alpha,-}\mu_{\alpha,+})(\rho^*-\rho_*/\sigma_{\alpha,+})}{\sqrt{2}\sigma_{\alpha,-}(\rho^*\sigma_{\alpha,+}+3\rho_*\sigma_{\alpha,-})} \tag{9.105}$$

else,

$$C_\alpha = \frac{1}{2}\left(\frac{(\sqrt{1-c_0}-\sqrt{t_1})^2}{\rho^* c_0}-1\right) = \frac{1}{2}\left(\left(\sqrt{\frac{1}{3c_+(\alpha)t_1}-\frac{1}{\rho^*}}-\sqrt{\frac{1}{3c_+(\alpha)}}\right)^2-1\right) \tag{9.106}$$

Here t_1^* is the t_1 that satisfies the following:

$$\frac{1}{2}\left(\left(\sqrt{\frac{1}{3c_+(\alpha)t_1}-\frac{1}{\rho^*}}-\sqrt{\frac{1}{3c_+(\alpha)}}\right)^2-1\right) = (1-\alpha)t_\alpha/\sqrt{2}$$

Thus we can obtain the theorem in the beginning of the section. □

References

[1] Air Traffic Organization Operations Planning Office of Aviation Research and Development. (2007). *Identification of Aircraft Touchdown Point in Commercial Operations*. DOT/FAA/AR-06/52. Washington, DC 20591. Available at *ftp://aosftp.jccbi.gov/230/public/SF21/*.

[2] Akaike, H., (1973) *Information theory and an extension of the maximum likelihood principle*. International Symposium on Information Theory, 2nd, Tsahkadsor, Armenian SSR; Hungary; 2-8 Sept. 1971. pp. 267-281.

[3] Atkinson, A. C., and Riani, M. (2000), *Robust Diagnostic Regression Analysis*, New York: Springer-Verlag.

[4] Atkinson, A. C., Riani, M., and Cerioli, A. (2004), *Exploring Multivariate Data With the Forward Search*, New York: Springer-Verlag.

[5] Bühlmann, P., and van de Geer, S., (2011) *Statistics for high-dimensional data: Methods, theory and applications*, Springer, New York.

[6] Butler, R. W., Davies, P. L. and Jhun, M. (1993). *Asymptotics for the Minimum Covariance Determinant estimator*. Ann. Statist. (21),1385-1400.

[7] Bickel, P., Ritov, Y., and Tsybakov. A, (2009) *Simultaneous analysis of Lasso and Dantzig selector*, Annals of Statistics, 37(4), 1705-1732.

[8] Candès, E. J. and Tao, T. (2005). *Decoding by linear programming*. IEEE Trans. Inform. Theory (51), 4203-4215.

[9] Candès, E. J. and Tao, T. (2007). *The Dantzig Selector: Statistical Estimation When p Is Much Larger than n*, The Annals of Statistics, 35 (6), 2313-2351.

[10] Coakley, C. W. and Hettmansperger, T. P. (1993). *A bounded influence, high breakdown, efficient regression estimator*. J. Amer. Statist. Assoc. (88), 872-880.

[11] Cai, T.T., Wang, L., and Xu, G. (2010) *New bounds for restricted isometry constants*, IEEE Trans. Inf. Theory, 56(9), 4388 -4394.

[12] Davies, P. L. (1987). *Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices*. Ann. Statist. (15),1269-1292.

[13] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). *Least angle regression*. Ann. Statist. (32), 407-451.

[14] Efron, B., Hastie,T., and Tibshirani, R. (2007) *Discussion: The Dantzig selector: Statistical estimation when p is much larger than n* Ann. Statist. 35(6), 2358-2364.

[15] ESRI shapefile technical description. An ESRI White Paper. (1998). Available at *http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf*.

[16] FAA William J. Hughes Technical Center. *Terminal Area Safety Program Fact Sheet*. AJP-6350.

[17] Federal Aviation Administration. (1998). *Advisory Circular 25-7A: Flight Test Guide for Certification of Transport Category Airplane*.

[18] Federal Aviation Administration, Office of Aviation Research. *Requirements for Landing in Federal Aviation Regulations*. Title 14 Code of Federal Regulations Part 25.125.

[19] Fan J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. (96), 1348-1360.

[20] Gannaz, I. (2007) *Robust estimation and wavelet thresholding in partially linear models*. Statistics and Computing, 17(4), 293-310.

[21] United States Government Accountability Office. (2011). *Aviation Safety. Enhanced Oversight and Improved Availability of Risk- Based Data Could Further Improve Safety*.

[22] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986), *Robust Statistics. The Approach Based on Influence Functions*. New York: John Wiley and Sons.

[23] Huang, J., Ma, S. and Zhang, C.-H. (2008). *Adaptive LASSO for sparse high-dimensional regression models*. Statist. Sinica. (18), 1603-1618.

[24] Huber, P. J. (1973). *Robust regression: Asymptotics, conjectures and Monte Carlo*. Ann. Statist. (1), 799-821.

[25] Huber, P. J. (1981). *Robust Statistics*, New York: Wiley.

[26] Hoeting, J., Raftery, A. E. and Madigan, D. (1996). *A method for simultaneous variable selection and outlier identification in linear regression*. Comput. Stat. Data Anal., (22) 252-270.

[27] Hubert, M., Rousseeuw, P., Van Aelst, S. (2008). *High-breakdown robust multivariate methods*. Statistical science, 23 (1), 92-119.

[28] Jurecková, J. (1971). *Nonparametric estimate of regression coefficients*. Ann. Math. Statist. (42), 1328-1338.

[29] Kent, J. T. and Tyler, D. E. (1996). *Constrained M-estimation for multivariate location and scatter*. Ann. Statist. (24) 1346-1370.

[30] Khan, J. and Van A., and Zamar, R. (2010). *Fast robust estimation of prediction error based on resampling*. Computational Statistics & Data Analysis. 54(12), 3121-3130.

[31] Kim, Y., Choi, H., and Oh, H-S., (2008) *Smoothly clipped absolute deviation on high dimensions*, Journal of American Statistical Association, (103), 1665-1673.

[32] Koenker, R. and Portnoy, S. (1987). *L-estimation for linear models*. J. Amer. Statist. Assoc. (82), 851-857.

[33] Koltchinskii, V. (2009) *The dantzig selector and sparsity oracle inequalities*, Bernoulli (15), 799-828.

[34] Liu, R. Y. , (1990). *On a Notion of Data Depth Based on Random Simplices*. The Annals of Statistics, (18), 405-414.

[35] Liu, R. Y, (1992). *Data depth and multivariate rank tests*. In L-1 Statistics and Related Methods Z. Y. Dodge, ed. 279-294. North-Holland, Amsterdam. Z.

[36] Liu, R. Y., (1995). *Control charts for multivariate processes*. J. Amer. Statist. Assoc. (90) 1380-1388.

[37] Liu, R. Y., Parelius, J. M. and Singh, K. (1999). *Multivariate analysis by data depth: Descriptive statistics, graphics and inference*. Ann. Statist. (27), 783840.

[38] Liu, R. Y., and Singh, K. (1993), *A Quality Index Based on Data Depth and Multivariate Rank Tests*, Journal of the American Statistical Association, (88), 252-260.

[39] Liu, R. and Singh, K. (1997). *Notions of limiting P-values based on data depth and bootstrap*. J. Amer. Statist. Assoc. (91), 266-277.

[40] Lopuhaä, H. P. (1991). *Multivariate τ-estimators for location and scatter*. Canad. J. Statist. (19) 307-321.

[41] Lopuhaä, H. P. (1999). *Asymptotics of reweighted estimators of multivariate location and scatter*. Ann. Statist. (27) 1638-1665.

[42] Lopuhaä, H. P. and Rousseeuw, P. J. (1991). *Breakdown points of affine equivariant estimators of multivariate location and covariance matrices*. Ann. Statist. (19), 229-248.

[43] Mallows, C. L. (1973) *Some Comments on CP*. Technometrics, 15(4), 661-675.

[44] Maronna R.A. (2011). *Robust Ridge Regression for High-Dimensional Data*. Technometrics . 53(1), 44-53.

[45] McCann, L., and Welsch, R. E. (2007), *Robust Variable Selection Using Least Angle Regression and Elemental Set Sampling*, Computational Statistics &Data Analysis, 52 (1), 249-257.

[46] Air Traffic Organization Operations Planning Office of Aviation Research and Development.(2008). *Analysis of Aircraft Touchdown Point and the Associated Uncertainty*. DOT/FAA/AR-07/67, Washington, DC.

[47] Menjoge, R. S., and Welsch, R. E. (2010), *A Diagnostic Method for Simultaneous Feature Selection and Outlier Identification in Linear Regression*, Computational Statistics & Data Analysis, (54), 3181-3193.

[48] Meinshausen, N., Rocha, G. and Yu, B. (2007). *A tale of three cousins: Lasso, L2Boosting and Dantzig*. Ann. Statist. (35) 2373-2384.

[49] Meinshausen, N. and Yu, B. (2006). *Lasso-type recovery of sparse representations for high-dimensional data.* Ann. Statist. 37(1), 246-270.

[50] Müller, S. and Welsh, A. H. (2005). *Outlier robust model selection in linear regression.* J. Amer. Statist. Assoc. (100), 1297-1310.

[51] Air Traffic Organization Operations Planning Office of Aviation Research and Development. (2007). *A Study of Normal Operational Landing Performance on Subsonic, Civil, Narrow-Body Jet Aircraft During Instrument Landing System Approaches.* DOT/FAA/AR-07/7, Washington, DC.

[52] Osborne, M., Presnell, B. and Turlach, B. (2000a). *A new approach to variable selection in least squares problems.* IMA J. Numer. Anal. (20) 389-404.

[53] Osborne, M., Presnell, B. and Turlach, B. (2000b). *On the lasso and its dual.* J. Comput. Graph. Statist. (9), 319-337.

[54] Perrotta, D., Riani, M., and Torti, F. (2009). *New robust dynamic plots for regression mixture detection.* Adv Data Anal Classif (3) 263-279.

[55] Raskutti, G., Wainwright, M. J. and Yu, B.(2010). *Restricted eigenvalue conditions for correlated Gaussian designs.* Journal of Machine Learning Research, (11) 2241-2259.

[56] Riani, M., Atkinson, A. C., and Cerioli, A. (2009), *Finding an Unknown Number of Multivariate Outliers,* Journal of the Royal Statistical Society, Ser. B, (71), 447-466.

[57] Rousseeuw, P. J. (1984). *Least median of squares regression.* J. Amer. Statist. Assoc. (79) 871-880.

[58] Rousseeuw, P. J. (1985). *Multivariate estimation with high breakdown point.* In Mathematical Statistics and Applications, B (W. Grossmann, G. Pflug, I. Vincze and W. Wertz, eds.). Reidel Publishing Company, Dordrecht.

[59] Rousseeuw, P. and Hubert M. (1997) *Recent developments in PROGRESS.* In L1-Statistical Procedures and Related Topics, ed Y. Dodge, IMS Lecture Notes (31), pp. 201-214.

[60] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* Wiley-Interscience, New York.

[61] Rousseeuw, P. J., Ruts, I. and Tukey, J. W. (1999a). The bagplot: A bivariate boxplot. American Statistician 53 382387.

[62] Rousseeuw, P. and Yohai, V. (1984). *Robust regression by means of S-estimators.* In Robust and Nonlinear Time Series Analysis. Lecture Notes in Statist. (26) 256-272. Springer, New York.

[63] Rousseeuw, P. J. and Van Driessen, K. (1999). *A fast algorithm for the minimum covariance determinant estimator.* Technometrics. (41) , 212-223.

[64] Rousseeuw, P. J. and Van Driessen, K. (2006). *Computing LTS regression for large data sets.* Data Mining and Knowledge Discovery (12), 29-45.

[65] Schwarz, G., (1978) *Estimating the Dimension of a Model*. The Annals of Statistics. 6(2), 461-464.

[66] She, Y., and Owen A.B. (2011) *Outlier detection using nonconvex penalized regression*. Journal of the American Statistical Association, 106 (494), 626-639.

[67] Tibshirani, R. (1996) *Regression Shrinkage and Selection via the Lasso.*, Journal of the Royal Statistical Society. Series B (Methodological) , 58(1), 267-288.

[68] Tropp,J. A. (2006). *Just relax: convex programming methods for identifying sparse signals in noise.* IEEE Transactions on Information Theory, (52), 10301051.

[69] Simpson, D. G., Ruppert, D. and Carroll, R. J. (1992). *On one-step GM-estimates and stability of inferences in linear regression*. J.Amer. Statist. Assoc. (87) 439-450.

[70] van de Geer, S. (2008). *High-dimensional generalized linear models and the lasso*. Ann. Statist. 36(2), 614-645.

[71] van de Geer, S. and B:uhlmann, B. (2009). *On the conditions used to prove oracle results for the lasso*. Electronic Journal of Statistics (3), 1360-1392.

[72] van de Geer, S. and B:hlmann, P. (2009). *On the conditions used to prove oracle results for the lasso*. Electronic Journal of Statistics, (3)1360-1392.

[73] Wainwright.M.J., (2009) *Sharp thresholds for noisy and highdimensional recovery of sparsity using l1constrained quadratic programming (lasso)*. IEEE Transactions on Information Theory, (55),2183-2202.

[74] Wang, L., and Li, R. (2009). *Weighted Wilcoxon-Type Smoothly Clipped Absolute Deviation Method*. Biometrics, 65(2), 564-571.

[75] Weisberg, S. (1985), *AppliedLinear Regression,(2nd ed.)*, New York: JohnWiley.

[76] Ye, F. and Zhang, C.-H. (2010) *Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls*. Journal of Machine Learning Research, 11:35193540.

[77] Yohai, V. J. (1987). *High breakdown point and high efficiency robust estimates for regression*. Ann. Statist. (15) 642-656.

[78] Zhang, C.-H. (2010). *Nearly unbiased variable selection under minimax concave penalty*. Annals of Statistics 38, 894-942.

[79] T. Zhang. Some sharp performance bounds for least squares regression with l1 regularization. Annals of Statistics, 37:21092144, 2009.

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

[81] Zhang, C.H. and Zhang, T. (2011) *A general theory of concave regularization for high dimensional sparse estimation problems*, Tech. Report, arXiv:1108.4988, arXiv.

[82] Zhao, P. and Yu, B. (2006). *On model selection consistency of LASSO*. J. Machine Learning Research. (7), 2541-2567.

[83] Zou,. H. and Hastie,. T (2005). *Regularization and variable selection via the elastic net*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.

[84] Zou, H. (2006) *The adaptive Lasso and its oracle properties*. J. Am. Stat. Assoc. 101(476), 1418-1429.

[85] Zou, H., Li, R.(2008) *One-step sparse estimates in nonconcave penalized likelihood models*. Ann. Stat. 36(4), 1509-1533.

[86] Zuo, Y. and Serfling, R. (2000a). *General notions of statistical depth function*. Ann. Statist. (28), 461-482.

[87] Zuo, Y. and Serfling, R. (2000b). *Nonparametric notions of multivariate "scatter measure" and "more scattered" based on statistical depth functions*. J. Multivariate Anal. (75), 62-78.

Vita

Wei Li

2000-2004 B. Sc. in Cilvil Engineeing, Tsinghua University

2004-2006 M. Engr. in Civil Engineering, City University of New York

2006-2012 Ph. D. in Statistics, Rutgers University