

## ABSTRACT

YANG, PENG. Penalized Regression and Model Selection in High Dimensions. (Under the direction of Soumendra Lahiri.)

This dissertation focuses on the penalized regression and model selection criteria in high dimensions. In particular, we investigate the methods which enjoy the Oracle Property and the suitable model selection criteria used to achieve such property. In addition, we point out the limitation of existing results, such as the requirement on the smallest signal and the distribution of errors, and we proposed new method/theory to expand the literature.

In Chapter 2, we propose a general family of nonconcave penalty functions which includes a few existing penalties, and we show that this family enjoys the Oracle Property in high dimensions. This family expands the choice of penalty function, and explains how the oracle property is achieved. Furthermore, we compare the asymptotic bias and variance of a few popular methods.

In Chapter 3, we construct an interesting example which shows that BIC fails to select the true model consistently in high dimensions. In fact, BIC favors an overfitted model, thus a modified version is necessary in high dimensions.

In Chapter 4, we propose the Partially Penalized Regression (PPR) which aims to solve the problems associated with small coefficients. In the literature, all the theoretical results on selection consistency are based on an inevitable assumption, that the smallest signal must be greater than a threshold. We discuss the limitation of this assumption and the issues associated with small signals. And we propose PPR, which is a two step procedure: it selects large coefficients in the first step and only penalizes small coefficients in the second step. Ideally, it combines the advantages of LASSO and nonconcave penal-

ties, and has a uniformly smaller risk than both types of penalties. In numerical studies, PPR demonstrates an advantage over the existing methods, in terms of both selection consistency and estimation accuracy.

In Chapter 5, we investigate moment conditions needed for the validity of the Oracle Property of Adaptive LASSO in high dimensions. We consider the case where the dimension  $p$  satisfies the growth condition  $\log p = O(n^\alpha)$  for some  $\alpha \in (0, 1)$ . In contrast to the existing literature where exponential tail conditions are assumed on the error variables in such high dimensional problems, here we show that just finiteness of the error variance is enough to guarantee the Oracle property for any  $\alpha \in (0, 1/2)$ . For  $\alpha \in [1/2, 1)$ , we show that existence of suitably higher order moments of the error variables is sufficient to guarantee the oracle property. The key tool is a new maximal inequality in high dimensions that is valid under polynomially decaying tails of the error distribution function. Results from a moderately large simulation study confirms the theoretical findings in finite sample applications.

© Copyright 2015 by Peng Yang

All Rights Reserved

Penalized Regression and Model Selection in High Dimensions

by  
Peng Yang

A dissertation submitted to the Graduate Faculty of  
North Carolina State University  
in partial fulfillment of the  
requirements for the Degree of  
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2015

APPROVED BY:

---

Howard Bondell

---

Rui Song

---

Hua Zhou

---

Soumendra Lahiri  
Chair of Advisory Committee

# DEDICATION

To my loving family.

## BIOGRAPHY

The author was born in 1988 in Penglai, Shandong, China. In 2006, he was admitted by School of Gifted Young at University of Science and Technology of China (USTC). There he met his wife and many life-long friends, and discovered his interest in mathematics to solve real-world problems. After receiving a Bachelor degree in statistics in 2010, he attended North Carolina State University to pursue a PhD in statistics. With the valuable instruction and guidance from his adviser Dr. Soumendra Lahiri, he focuses on the research of penalized regression. He will complete his Ph.D in May 2015, and join WalmartLab as a data scientist.

## ACKNOWLEDGEMENTS

Throughout my life, I have always been a lucky boy who met so many nice people and received so much help from them.

I would like to take this opportunity to express my sincerest gratitude and appreciation to my adviser, Dr. Soumendra Lahiri, who provided me insightful instruction, inspiring ideas, and extraordinary kindness throughout my research. He has offered invaluable support for my study and career development. I have learned a lot from Dr. Lahiri, not only on how to conduct research work, but also the passion and strictness for truth. I feel very lucky to have him as my adviser, without whose guidance, I would not be able to accomplish this work.

I would also like to extend my appreciation to my committee members Dr. Hua Zhou, Dr. Howard Bondell, and Dr. Rui Song for their constant encouragement and wonderful advices. In particular, I want to thank Dr. Hua Zhou who kindly provided references for me and shared his working project. In addition, I owe special thanks to Dr. Shuva Gupta who I have worked with on a few topics. His knowledge and humor have been a great help for me.

I also owe my gratitude to all my supervisors at internships. In particular, Dr. Mandy Bergquist at GlaxoSmithKline is the best manager I have ever met, who taught me a lot on statistical knowledge and professional skills. Dr. Andrei Prudius at Bloomberg is one of the smartest and kindest people I have ever known, who lead me into quantitative finance and backed me during the challenging internship. I want to thank all the managers/coworkers I have met for your kind suggestion and encouragement, especially Mr. Igor Von Nyssen, Mr. Rudy Rodriguez, Dr. Satish Vedantam, and Dr. Steven Novick.

I am grateful to all the faculty members in the department of statistics for offering

a comprehensive collection of courses, which helps lay solid foundations for my skill set. I especially enjoy the course by Dr. Peter Bloomfield, Dr. Lexin Li and Dr. Leonard Stefanski. I am also grateful to Dr. Nagiza Samatova, Dr. Harry Perros, Dr. Steffen Heber for their interesting CS courses on machine learning and algorithms. I would also like to thank Dr. Charlie Smith for his warm greetings and snacks when I stay up late in SAS HALL. Also I would like to thank all the staff for their excellent service to the department.

Thanks to my fellow students and friends. It is your intelligence and diligence that stimulates me to surpass myself. And your supports and encouragements have made my life easier in the United States. Thanks to all USTC Alumni all over the world who treat me as families wherever I go.

Last but not least, I would like to express my deepest gratitude to my family. To my lovely wife, Shiyu Du, your love has been accompanying me every moment of my life. To my great parents Jianmin Yang and Qiurong Xu, your unconditional support and love is always the source of my strength. To my great grandma, Jingjuan Li, I miss you all the time.

# TABLE OF CONTENTS

<b>LIST OF TABLES</b> . . . . .	<b>viii</b>
<b>LIST OF FIGURES</b> . . . . .	<b>ix</b>
<b>Chapter 1 Introduction</b> . . . . .	<b>1</b>
1.1 Famous Penalty Functions . . . . .	3
1.2 Tuning Parameter Selection . . . . .	7
1.3 Our Work . . . . .	9
<b>Chapter 2 Oracle Family of Penalties</b> . . . . .	<b>10</b>
2.1 Introduction . . . . .	10
2.2 Oracle Family and Regularity Conditions . . . . .	12
2.2.1 Oracle Family . . . . .	12
2.2.2 Examples in $\mathcal{P}$ . . . . .	13
2.2.3 Regularity Conditions . . . . .	16
2.3 Existing Results and Comparisons . . . . .	18
2.4 Oracle Property . . . . .	21
2.4.1 General Results . . . . .	21
2.4.2 One-Step Estimator in High Dimensions . . . . .	24
2.5 Comparison of Popular Methods . . . . .	26
2.5.1 Oracle Family $\mathcal{P}$ . . . . .	26
2.5.2 Adaptive Lasso . . . . .	28
2.6 Proofs . . . . .	31
<b>Chapter 3 BIC's Failure in High Dimension</b> . . . . .	<b>48</b>
3.1 Introduction . . . . .	48
3.2 A Counter Example . . . . .	52
3.3 Proofs . . . . .	54
<b>Chapter 4 Partially Penalized Regression</b> . . . . .	<b>58</b>
4.1 Introduction . . . . .	58
4.2 Motivation . . . . .	60
4.2.1 The Limitation of Previous Results . . . . .	60
4.2.2 The Behavior of Risk Functions . . . . .	65
4.3 Partially Penalized Regression . . . . .	70
4.3.1 First Step . . . . .	71
4.3.2 Second Step . . . . .	73
4.4 Effects of Small Coefficients . . . . .	75
4.4.1 General Settings . . . . .	75

4.4.2	Amplification of Variance . . . . .	77
4.4.3	An Estimator for Large Coefficients . . . . .	80
4.4.4	A Dilemma in Fixed Dimensions . . . . .	82
4.5	The L-consistency of LIC . . . . .	83
4.6	Numerical Studies . . . . .	86
4.6.1	Simulations . . . . .	86
4.6.2	Real Data . . . . .	94
4.7	Proofs . . . . .	96
<b>Chapter 5 Finite Variance Oracle Property . . . . .</b>		<b>112</b>
5.1	Introduction . . . . .	112
5.2	Results on ALASSO . . . . .	118
5.2.1	The ALASSO Method . . . . .	118
5.2.2	Notation and Conditions . . . . .	119
5.2.3	Main Results . . . . .	122
5.3	A Maximal Inequality . . . . .	125
5.4	Simulation . . . . .	126
5.5	Proofs . . . . .	129
5.5.1	Lemmas . . . . .	129
5.5.2	Proof for ALASSO . . . . .	133
5.5.3	Proof of the Maximal Inequality . . . . .	139
<b>References . . . . .</b>		<b>142</b>

## LIST OF TABLES

Table 4.1	<i>Simple demonstration of magnitude assumption - with the comparison of LASSO and MCP. (<math>n = 100, p = 1000, p_0 = 10</math>)</i>	64
Table 4.2	Thresholding Rules ( $s(z, \lambda)$ is the thresholding rule of LASSO)	67
Table 4.3	<i>Comparison of methods when <math>n = 100, p = 1000, p_0 = 10, \beta</math> is type 1.(based on 500 replications)</i>	89
Table 4.4	<i>The standard error of statistics reported in Table 4.3</i>	90
Table 4.5	<i>Comparison of methods when <math>n = 100, p = 1000, p_0 = 10, \beta</math> is type 2.(based on 500 replications)</i>	91
Table 4.6	<i>Comparison of methods when <math>n = 200, p = 3000, p_0 = 20, \beta</math> is type 1.(based on 200 replications)</i>	92
Table 4.7	<i>Comparison of methods when <math>n = 200, p = 3000, p_0 = 20, \beta</math> is type 2.(based on 200 replications)</i>	93
Table 4.8	Comparison of methods - predicting 5 soil functional properties with Africa Soil Property Data	95
Table 4.9	Number of variables selected in the first step of PPR.	96
Table 5.1	<i>Probability of identifying the true model</i>	128
Table 5.2	<i>Correlation coefficient for Normal Q-Q plots for <math>\hat{\beta}_1</math></i>	128

## LIST OF FIGURES

Figure 1.1	The Comparison of TL, SCAD, MCP penalties. Without loss of generality, we take $a = 3.5$ and $\lambda = 1$ for all of them. . . . .	5
Figure 2.1	The Comparison of TL, SCAD, MCP penalties with the same $(a, \lambda)$ value. Without loss of generality, we take $a = 3.5$ and $\lambda = 1$ for all of them. . . . .	15
Figure 4.1	The comparison of risk function for the five penalty function list in table 4.2. . . . .	68
Figure 5.1	<i>Empirical probability of selecting the true model using MCP under error distributions (a)-(d)</i> . . . . .	114
Figure 5.2	<i>Normal Q-Q plots of MCP estimates of a nonzero component under error distributions (a) and (d) for sample sizes <math>n = 100, 1600</math>, based on 500 simulation runs.</i> . . . . .	115
Figure 5.3	<i>Empirical probability of selecting the true model using ALASSO under error distributions (a)-(d)</i> . . . . .	129
Figure 5.4	<i>Normal Q-Q plots of ALASSO estimates of a nonzero component under error distributions (a) and (d) for sample sizes <math>n = 100, 1600</math>, based on 500 simulation runs.</i> . . . . .	129

# Chapter 1

## Introduction

Due to the huge improvement in data gathering and data processing mechanisms, high-dimensional data arises in many areas. In contrast to the the conventional data where the number of observations is larger than the number of features, the high-dimensional data may have much more features than observations (although the related dimension could be much smaller). In this case, even the simplest traditional statistical methods fail to work.

A famous example of high-dimensional data is the DNA Microarray data. DNA microarrays are used to measure the expression levels of tens of thousands of genes simultaneously. However, the sample size of such data is much smaller (usually a few hundreds). Of all the gene expressions measured, only a few are related to the target we are interested (e.g. disease occurrence, survival time, drug response). Therefore, new techniques are in desperately need to identify important genes, to make precise estimations, or to perform classifications.

Another example is customer financial data. Every transaction is recorded, creating lots of new variables such as gas/housing/grocery payment within one week/month/season;

besides, an inquiry, a visit, a purchase can also be recorded into databases. All these forms a personal profile which may have thousands of features. People work on such data in order to perform fraud detection, advertisement, etc.

Linear models are the most popular and simple techniques to model the target variable on the input variables:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where  $\mathbf{Y}$  is an  $n \times 1$  vector of target variable,  $\mathbf{X}$  is an  $n \times p$  design matrix with  $p \gg n$ . The vector of coefficients  $\boldsymbol{\beta}$  is sparse, which means only a small number of entries are nonzero and all the others are zeroes.  $\boldsymbol{\varepsilon}_i$  are *i.i.d* random variables with mean 0 and variance  $\sigma^2$ . Let the true value of  $\boldsymbol{\beta}$  be

$$\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_p^*)^T = (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*),$$

without loss of generality, assume that  $\boldsymbol{\beta}_2^* = \mathbf{0}$ . We denote length of  $\boldsymbol{\beta}$  and  $\boldsymbol{\beta}_1$  by  $p$  and  $p_0$ , respectively. We do allow  $p$  and  $p_0$  goes to infinity as  $n \rightarrow \infty$ , but we don't add subscript  $n$  for the purpose of simplicity.

The ordinary least square(OLS) estimator  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$  does not apply here since  $\mathbf{X}^T \mathbf{X}$  is singular when  $p > n$ . Even when  $p < n$ , the OLS estimator tends to have a huge variance. Therefore, It is important to select a subset of input variables, and to conduct further estimation based on the subset.

The penalized regression is a class of methods which simultaneously select relevant variables and estimate parameters efficiently. The ordinary least squares (OLS) estimator is obtained by minimizing the squared errors  $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$ , while the penalized regression amounts to perform the least squares with some constraints on the coefficient, and the penalties are usually imposed upon the magnitude of coefficients. The penalized loss

function is usually as following:

$$Q_n(\boldsymbol{\beta}; \lambda) = \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n \sum_{j=1}^p P_\lambda(|\beta_j|) \quad (1.2)$$

where  $P_\lambda(\cdot)$  is a penalty function, and  $\lambda$  is a tuning parameter. The argument that minimizes (1.2) is the estimator of coefficients, i.e.  $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta}; \lambda)$ .

Compared to the traditional way of subset selection, penalized regression have advantages in three aspects: 1) The number of all possible subsets grows exponentially with  $p$ , which is computationally infeasible for subset selection. 2) The penalized regression has a more stable estimator. 3) The Penalized estimator achieves the selection and estimation simultaneously, while the subset selection is a two-step procedure, where the errors in the first step can be amplified in the second step.

## 1.1 Famous Penalty Functions

We start with the LASSO penalty (Tibshirani, 1994), which is one of the most famous and powerful penalties. It minimizes the following loss function:

$$Q_n(\boldsymbol{\beta}; \lambda) = \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n \sum_{j=1}^p \lambda |\beta_j| \quad (1.3)$$

It penalizes the  $L_1$  norm of the coefficients and results in a sparse estimator with only a few nonzero entries, therefore it achieves the model selection and coefficient estimation simultaneously. (Efron et al., 2004) proposed the LARS algorithm and applied it to calculate the solution path for LASSO. Given a few grid points of  $\lambda$ , LARS calculates the estimated coefficients at each point among the path. This makes the tuning process

much effective. (Zhao and Yu, 2006) gives the so-called “Irrepresentable Condition” under which LASSO selects the true model consistently.

Prior to LASSO, many penalties have been proposed, such as Ridge regression and Bridge regression (Frank and Friedman, 1993), but few of them draw attentions in recent years. People have a commonly recognized criteria to judge the goodness of a penalty function, which is called the “Oracle Property”.

Imagine that you luckily identify the true subset of nonzero coefficients and calculate the OLS estimator only based on the subset (call it “Oracel Estimator”  $\hat{\beta}_{1,oracle}$ ). This should give you the best linear unbiased estimator. If the penalized regression also returns such a nice estimator (asymptotically), we say that it enjoys the “Oracle property”. That is,

1. The probability of identifying the true model tends to 1. i.e.  $P(\hat{\beta}_2 = 0) \rightarrow 1$ , as  $n \rightarrow \infty$ .
2. The estimator has an asymptotical normal distribution as the oracle estimator does.  
i.e.  $\hat{\beta}_1 \rightarrow_d \hat{\beta}_{1,oracle}$

The Oracle Property has become the ultimate goal of the design of penalty functions. As we have discussed, LASSO enjoys the oracle properties only under certain conditions (Zhao and Yu (2006), Zou (2006)).

In the following years, many penalties have been designed to enjoy the Oracle Properties. Most of them can be divided into two classes: The first class relies on a delicately designed format. Usually, the function is nonconcave with a positive right-derivative at 0, and becomes constant beyond a certain point. Examples includes SCAD (Fan and Li, 2001), MCP (Zhang, 2010), and Truncated LASSO (Shen et al., 2012). The other class utilizes an initial estimator (e.g. OLS estimator or LASSO estimator) to adjust the

penalty weight on each entry, and the initial estimators are guaranteed to enjoy certain properties under general conditions. Famous examples include Adaptive LASSO (Zou, 2006) and one-step estimator (Zou and Li, 2008).

Here we present three examples in the first class: SCAD, MCP, and Truncated LASSO. Their penalty functions  $P_\lambda(t)$  are defined using the derivatives: ( $t > 0$ )

$$\begin{cases} P'_\lambda(t) = \lambda \left[ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right], & \text{for SCAD} \\ P'_\lambda(t) = \lambda \left( 1 - \frac{t}{a\lambda} \right)_+, & \text{for MCP} \\ P'_\lambda(t) = \lambda I(t \leq a\lambda), & \text{for truncated LASSO (TL)} \end{cases} \quad (1.4)$$

The following figure 1.1 gives the comparison of the penalty functions and their derivatives.

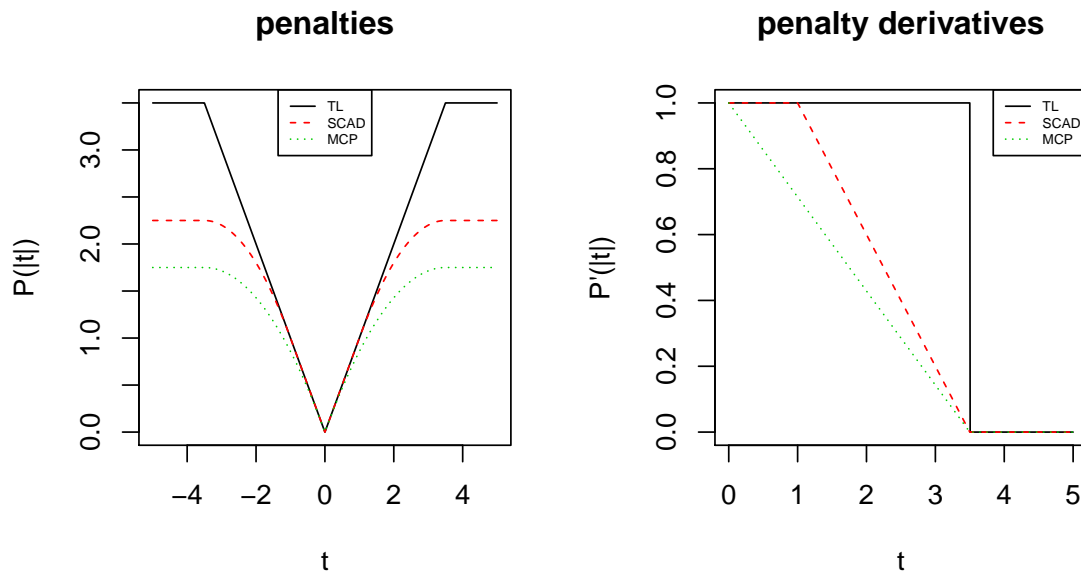


Figure 1.1: The Comparison of TL, SCAD, MCP penalties. Without loss of generality, we take  $a = 3.5$  and  $\lambda = 1$  for all of them.

As we can see from figure 1.1 and their formula, these three penalties have something in common:

1.  $P_\lambda(t)$  is a constant as long as  $t \geq a\lambda$ . This ensures that the nonzero coefficients are not penalized at all as long as they are sufficiently large (since the derivative is exactly zero).
2.  $P'_\lambda(0+) = \lambda$  is a positive number. This ensures that zero coefficients are shrunk to exactly 0.

Therefore, they all require the derivative function decreases from  $\lambda$  to 0, while the difference is the way they decrease. The first two decrease continuously, while the last has a jump from  $\lambda$  to 0. All of them have been proved to enjoy the oracle properties in high dimensions.

The second class aim to adjust the penalty weights based on each coefficient's initial estimator, in hope that the weight is small for large coefficients but large for zero coefficients. Some general assumptions are used to guarantee that the initial estimator and the true parameter are close enough, so that the behavior of weights given initial estimators is close to that given true parameters.

Here we give two examples in this class. Adaptive LASSO (Zou, 2006) is given by

$$Q_n(\boldsymbol{\beta}; \lambda) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^p \frac{|\beta_j|}{|\hat{\beta}_j|^r} \quad (1.5)$$

where  $\hat{\beta}_j$  is the  $j$ -th element taken from an initial estimator. The suggestion is ordinary least square estimator for low-dimensional data, and LASSO or Ridge estimator for high-dimensional data. Under general conditions and with certain rate of  $\lambda_n$ , the function penalizes heavily on zero coefficients as their initial estimators are close to zeroes; and

it penalizes slightly on the large nonzero coefficients as their initial estimators are also large. Thus it achieves the Oracle Properties and outperforms LASSO in most cases. (Chattergee and Lahiri, 2013) proved its oracle properties in high-dimensional settings.

Similar to the idea of ALASSO, (Zou and Li, 2008) proposed the one-step estimator, which minimizes

$$Q_n(\boldsymbol{\beta}; \lambda) = \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n \sum_{j=1}^p P'_\lambda(\hat{\beta}_j) |\beta_j| \quad (1.6)$$

where  $P'_\lambda(\hat{\beta}_j)$  denotes the derivative of SCAD penalty function taken value at an initial estimator of  $\beta_j$ . Since  $P'_\lambda(t)$  decreases as  $t$  increases and decays to 0 when  $t > a\lambda$ , the one-step is able to adjust the penalization weight according to the initial estimator. Comparing it with ALASSO, it has a different way to define the penalty weights as functions of initial estimators, and its penalization weight is exactly 0 for large coefficients (which is asymptotically 0 for ALASSO).

Although there are other penalties, such as Elastic Net (Zou and Hastie, 2005), Grouped LASSO (Yuan and Lin, 2006), Dantzig (Candes and Tao, 2007), etc. Our focus will be on the penalties which achieve Oracle Property.

## 1.2 Tuning Parameter Selection

Another issue is the tuning problem of regularization parameter ( $\lambda$ ).

Even though those aforementioned penalized estimators enjoy the oracle properties with certain choices of regularization parameter  $\lambda$ , it is not easy to find the right choice given a practical problem. It turns out the performance of penalized regression depends heavily on the tuning step, which makes this problem even more important.

Intuitively, people use k-fold cross-validation (CV) or generalized cross-validation

(GCV) to solve the tuning problem, which is conventional in practice and also suggested in the original paper (Fan and Li, 2001). However, Wang et al. (2007) showed that the commonly used GCV is similar to AIC (Akaike, 1973), and that both of them cannot select the tuning parameter satisfactorily, with a nonignorable overfitting effect in the resulting model. In addition, they proved that BIC (Schwarz, 1978) is able to identify the true model consistently when the feature dimension  $p$  is fixed.

As we know, for the penalized linear regression, the general information criterion(IC) is

$$\text{IC}(\lambda) = \log \hat{\sigma}_\lambda^2 + |A_\lambda| \frac{C_n}{n} \quad (1.7)$$

where  $A_\lambda = \{j : 1 \leq j \leq p, \hat{\beta}_j(\lambda) \neq 0\}$  is index set of nonzero coefficients identified by  $\lambda$ ,  $\hat{\beta}(\lambda)$  is the estimated value of coefficients under tuning parameter  $\lambda$ , and  $\hat{\sigma}_\lambda^2 = n^{-1} \|\mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda)\|^2$ .  $C_n$  is a number which determines the type of IC. For example, when  $C_n = 2$ , (1.7) denotes AIC; and when  $C_n = \log n$ , (1.7) denotes BIC.

When  $p$  diverges with  $n$ , it turns out BIC is not the right choice. In Chapter 3, we shall construct a general counter-example, by which we show that BIC fails to select the true model with a positive probability in high dimensions. Depending on the dimensionality,  $C_n$  is suggested to take value  $C_n/\log p \rightarrow \infty$  or  $C_n/\log n \rightarrow \infty$ . (Wang et al., 2013; Fan and Tang, 2013; Sun et al., 2013; Wang and Zhu, 2011).

The intuitive reason is as following: each candidate model contributes a bit to the probability of incorrect selection, even though each contribution is small, when adding up together, the whole probability of inconsistent selection is amplified, by the total number of candidate model. In fixed dimension, this total number is a fixed number since  $p$  is fixed, thus  $\log n$  in BIC is good enough. However, when  $p$  diverges, the number of candidate models can not be considered as fixed. Hence we need a larger  $C_n$  which is

greater than  $\log n$ .

## 1.3 Our Work

By studying the design of the popular penalty functions, we proposed a general class of nonconcave penalty function, and further proved that all the penalties in that class enjoy the Oracle Property. Furthermore, we compare the bias and variance of several popular methods, and demonstrate the effects that small coefficients have on the bias and covariance matrix.

Noting that selecting the tuning parameter  $\lambda$  is an important issue, we show a counter example in which BIC fails to select the true model consistently in high dimensions.

In addition, all the theoretical results on selection consistency depend on an assumption that the smallest coefficients must be large enough. We question the validity of this assumption and propose a partially penalized regression method which does not rely on this assumption. It picks the large nonzero coefficients first, then penalizes only on the remaining coefficients. We will show its advantage over other penalties.

In the end, we explore the cases when the error distribution does not meet the exponential tail conditions, and show that the Oracle Property can be achieved under weaker tail conditions.

# Chapter 2

## Oracle Family of Penalties

### 2.1 Introduction

Due to the huge improvement in data gathering and data processing mechanisms, high-dimensional data arises in many areas. In contrast to the the conventional data where the number of observations is larger than the number of features, the high-dimensional data may have much more features than observations. In this case, even the simplest traditional statistical methods fail to work.

Linear models are the most popular and simple techniques to model the target variable on the input variables:

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

where  $\mathbf{Y}$  is an  $n \times 1$  vector of target variable,  $\mathbf{X}$  is an  $n \times p$  design matrix with  $p \gg n$ . The vector of coefficients  $\boldsymbol{\beta}$  is sparse, which means only a small number of entries are nonzero and all the others are zeroes.  $\boldsymbol{\varepsilon}_i$  are *i.i.d* random variables with mean 0 and variance  $\sigma^2$ .

The penalized regression is a popular method for high dimensional data. It aims to

minimize the following penalized loss function:

$$Q_n(\boldsymbol{\beta}; \lambda) = \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n \sum_{j=1}^p P_\lambda(|\beta_j|) \quad (2.2)$$

where  $\lambda$  is a tuning parameter and  $P_\lambda(\cdot)$  is a penalty function. For instance,  $P_\lambda(|t|) = |t|$  for the LASSO (Tibshirani, 1994);  $P_\lambda(|t|) = t^2/2$  for the Ridge (Hoerl and Kennard, 1970). The argument that minimizes (2.2) is the estimator of coefficients, i.e.

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathcal{R}^p} Q_n(\boldsymbol{\beta}; \lambda) \quad (2.3)$$

Compared to the traditional way of subset selection, the advantages are mainly in three aspects: 1) The number of all possible subsets grows exponentially with  $p$ , which is computationally infeasible for subset selection. 2) The penalized regression has a more stable estimator. 3) The Penalized estimator achieves the selection and estimation simultaneously, while the subset selection is a two-step procedure, where the errors in the first step can be amplified in the second step.

There has been a rich literature in the format of  $P_\lambda(\cdot)$ , famous examples includes LASSO (Tibshirani, 1994), SCAD (Fan and Li, 2001), MCP (Zhang, 2010), Truncated LASSO (Shen et al., 2012), one-step estimator (Zou and Li, 2008), Hard Thresholding (Zheng et al., 2014). Although following different format, they enjoy similar properties, such as model selection consistency and estimation efficiency.

In this paper, we propose a family  $\mathcal{P}$  of penalty functions, and show the nice properties of  $\hat{\boldsymbol{\beta}}$  using functions from  $\mathcal{P}$ . Furthermore, based on the bias and variance obtained from section 2.4, we compare the performance of a few famous penalties. We show that most of them are more or less the same and that the existence of small coefficients not

only affects the selection consistency, but also increases the  $L_2$  loss of  $\hat{\beta}$ .

## 2.2 Oracle Family and Regularity Conditions

In this section, we propose a family of nonconcave functions, and give a few examples of this family. Besides, we state a set of regularity conditions to facilitate the theoretical results in later sections.

### 2.2.1 Oracle Family

We propose a family of penalty functions  $\mathcal{P}$ , which we refer as the ‘‘Oracle Family’’. We shall prove that any the penalty function belonging to  $\mathcal{P}$  leads to estimators that achieve the Oracle Property in high dimensions.

The Oracle Family  $\mathcal{P}$  includes any function  $P_\lambda(t) : [0, \infty) \rightarrow R$  that satisfies:

(p.1)  $P_\lambda(t)$  is differentiable in  $(0, \infty)$ ,

(p.2)  $P_\lambda(0) = 0$  and  $P'_\lambda(0+) = \lambda$ ,

(p.3)  $\exists a > 0$ , such that  $P'_\lambda(t) = 0$ , when  $t \geq a\lambda$ ,

(p.4)  $\exists c, 0 < c < 1$ , such that  $\forall t \in (0, a\lambda), P_\lambda(t) > c\lambda \cdot t$ .

In addition, we need a further requirement which depends on the design matrix, and we will state that in regularity condition (C.3).

These properties provide some hints on how to design a good penalty function with good properties:

- First property guarantees the continuity of solution.
- It has a positive right derivative at 0, thus it is able to shrink certain  $\beta$ 's to be zero.

- It is constant beyond certain points, so that estimation of large coefficients has no bias
- It is above a LASSO type penalty in order to avoid false positive selection.

These hints will be discussed further throughout the documents.

## 2.2.2 Examples in $\mathcal{P}$

The Oracle family  $\mathcal{P}$  includes a few famous examples, we shall discuss three of them in the this section.

- **SCAD** Fan and Li (2001) proposed the famous SCAD penalty function, and proved its oracle property under fixed dimensions. It turns out that SCAD also works well in high dimensions (Xie and Huang, 2009; Kim et al., 2008).

SCAD is defined using the derivative:

$$P'_\lambda(t) = \lambda \left[ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right] \quad (2.4)$$

which creates the penalty function as

$$P_\lambda(t) = \begin{cases} \lambda t, & 0 \leq t \leq \lambda \\ -\frac{t^2 - 2a\lambda t + \lambda^2}{2(a-1)}, & \lambda < t \leq a\lambda \\ \frac{(1+a)\lambda^2}{2}, & t > a\lambda \end{cases} \quad (2.5)$$

Obviously, it satisfies requirements (p.1) - (p.3); as for (p.4), we define

$$c_0 = \inf_{t \in (0, a\lambda)} \frac{P_\lambda(t)}{\lambda t}, \quad (2.6)$$

which is the largest constant  $c$  that makes (p.4) true. Take  $c_0 = \frac{a+1}{2a}$ , the (p.4) is true for SCAD as long as  $c \leq c_0$ .

- **MCP** The MCP penalty (Zhang, 2010) is defined using:

$$P'_\lambda(t) = \lambda \left(1 - \frac{t}{a\lambda}\right)_+ \quad (2.7)$$

which makes

$$P_\lambda(t) = \begin{cases} \lambda t - \frac{t^2}{2a}, & 0 \leq t \leq a\lambda \\ \frac{a\lambda^2}{2}, & t > a\lambda \end{cases} \quad (2.8)$$

In fact, the general MCP class includes all the quadratic splines, and SCAD is also a special case in MCP.

Similar to SCAD penalty, MCP satisfies (p.1) - (p.4) with  $c_0 = 1/2$ .

- **TL** Shen et al. (2012) suggested constrained  $L_0$  likelihood. Its computational surrogate Truncated LASSO(TL) is optimal in that it achieves feature selection consistency and sharp parameter estimation. TL is defined using the derivative

$$P'_\lambda(t) = \lambda I(t \leq a\lambda) \quad (2.9)$$

which makes

$$P_\lambda(t) = \begin{cases} \lambda t, & 0 \leq t \leq a\lambda \\ a\lambda^2, & t > a\lambda \end{cases} \quad (2.10)$$

This does not belong to  $\mathcal{P}$ , but we can treat it as an extreme case. We keep it here for knowledge and comparison purposes.

Figure 2.1 demonstrates the three examples. The left panel shows the shape of penalty function  $P_\lambda(|t|)$ , while the right panel shows its derivative  $P'_\lambda(|t|)$ . As we can see, for the

derivative function, they share the same part in the range  $\{0\} \cup [a\lambda, \infty)$ . In fact, functions in  $\mathcal{P}$  all share the same part in that range, but they are more flexible in  $(0, a\lambda)$ . They can take any format in  $(0, a\lambda)$ , as long as the  $P_\lambda(|t|)$  is differentiable and it is above  $c\lambda|t|$  for some constant  $c$ .

In other words, SCAD and MCP are two special functions in this family. Furthermore, we shall see that as  $n \rightarrow \infty$ , SCAD, MCP, and all other functions in  $\mathcal{P}$  have the same bias and variance under certain regularity conditions. And their minor difference in practical performance is a consequence of small coefficients and the selection of penalty parameter  $\lambda$ . We shall explain this further in following sections.

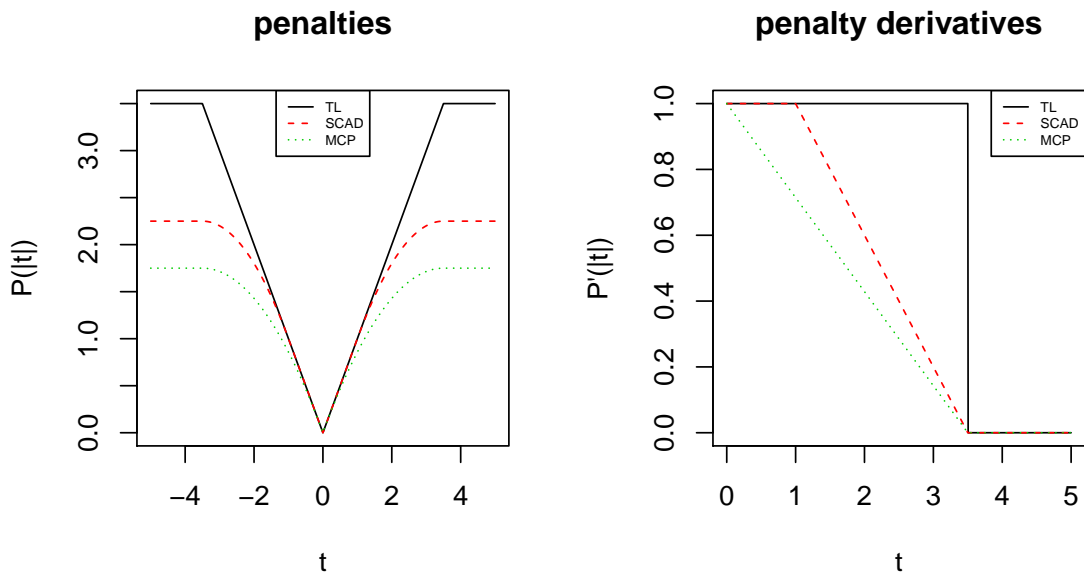


Figure 2.1: The Comparison of TL, SCAD, MCP penalties with the same  $(a, \lambda)$  value. Without loss of generality, we take  $a = 3.5$  and  $\lambda = 1$  for all of them.

### 2.2.3 Regularity Conditions

Let the true value of  $\boldsymbol{\beta}$  be

$$\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_p^*)^T = \begin{pmatrix} \boldsymbol{\beta}_1^* \\ \boldsymbol{\beta}_2^* \end{pmatrix},$$

without loss of generality, assume that  $\boldsymbol{\beta}_2^* = \mathbf{0}$ . We denote length of  $\boldsymbol{\beta}$  and  $\boldsymbol{\beta}_1$  by  $p$  and  $p_0$ , respectively. In high dimensions, we do allow  $\beta'_j$ 's ( $j = 1, \dots, p$ ) depend on  $n$ , and  $p$  and  $p_0$  goes to infinity as  $n \rightarrow \infty$ , but we do not add subscript  $n$  for the purpose of simplicity. Besides, assume the design matrix is standardized such that  $\|\mathbf{X}_{\cdot j}\|^2 = n$  for any  $j$ ,  $1 \leq j \leq p$ .

Denote  $\mathbf{X}_1$  be the  $n \times p_0$  submatrix of  $\mathbf{X}$  that corresponds to the nonzero coefficients  $\boldsymbol{\beta}_1$ , and  $\mathbf{X}_2$  is the submatrix of the remaining columns. Let  $\mathbf{C}_n = n^{-1}\mathbf{X}^T\mathbf{X}$ , and write  $\mathbf{C}_n$  as

$$\mathbf{C}_n = \begin{bmatrix} \mathbf{C}_{11,n} & \mathbf{C}_{12,n} \\ \mathbf{C}_{21,n} & \mathbf{C}_{22,n} \end{bmatrix},$$

where  $\mathbf{C}_{11,n}$  is the  $p_0 \times p_0$  submatrix that corresponds to the nonzero coefficients  $\boldsymbol{\beta}_1^*$ . i.e.  $\mathbf{C}_{11,n} = n^{-1}\mathbf{X}_1^T\mathbf{X}_1$ .

We shall make use of the following regularity conditions:

(C.1) There exists  $\delta \in (0, 1)$ , such that for all  $n > \delta^{-1}$

$$(\mathbf{X}^T\mathbf{C}_{12,n}\mathbf{Y})^2 \leq \delta^2(\mathbf{X}^T\mathbf{C}_{11,n}\mathbf{X})(\mathbf{Y}^T\mathbf{C}_{22,n}\mathbf{Y}) \quad \text{for all } \mathbf{X} \in \mathbb{R}^{p_0}, \mathbf{Y} \in \mathbb{R}^{p-p_0}.$$

(C.2) (i) The eigenvalues of  $\mathbf{C}_{11,n}$  is bounded away from 0 and  $\infty$  uniformly. That is,

$$\mathbf{C}_{11,n} \in \mathcal{U}(c_1, c_2) = \{\mathbf{A} : 0 < c_1 \leq \boldsymbol{\alpha}^T \mathbf{A} \boldsymbol{\alpha} \leq c_2 < \infty, \forall \|\boldsymbol{\alpha}\| = 1\} \quad (2.11)$$

(ii)

$$\max_{1 \leq j \leq p} \left\{ n^{-1} \sum_{i=1}^n |x_{ij}^r| \right\} = O(1),$$

where  $r \geq 3$  is an integer (to be specified in the statement of theorems). Depending on  $r$ , the dimensionality satisfies  $p = o(n^{(r-2)/2})$ .

(C.3) Define the maximum concavity of penalty  $P_\lambda$  as

$$\kappa(P_\lambda) = \sup_{t_1 < t_2 \in (0, \infty)} \frac{P'_\lambda(t_1) - P'_\lambda(t_2)}{t_2 - t_1},$$

$\kappa(P_\lambda)$  satisfies

$$\kappa(P_\lambda) < \frac{1}{2}(1 - \delta) \cdot c_1$$

where  $c_1$  is the lower limit of eigenvalues in (C.2).

(C.4) The nonzero coefficients satisfy

$$\min_{1 \leq j \leq p_0} |\beta_j^*| \cdot \frac{\sqrt{n}}{\sqrt{p_0 \log n}} \rightarrow \infty.$$

Condition (C.1) is a general condition on the multiple correlation between the relevant variables (corresponding to  $\boldsymbol{\beta}_1$ ) and the spurious variables (corresponding to  $\boldsymbol{\beta}_2$ ). Here we require the correlation to be strictly smaller than 1. This condition is much weaker than assuming zero correlation between the sets of variables. (e.g. (Huang et al., 2008))

Condition (C.2) requires that the eigenvalues of  $\mathbf{C}_{11,n}$  is strictly bounded away from 0 and  $\infty$ . Similar conditions can be found in Bickel and Levina (2008); Kim et al. (2008).

Condition (C.3) is similar to the sparse convexity condition in Zhang (2010), which requires that the maximum concavity of the penalty function is smaller than the smallest eigenvalue of  $\mathbf{C}_{11,n}$  multiplied by a constant. However, Zhang (2010) requires such inequality for a collection of sparse model  $A$ , while we only require it to be true for the true sparse model. A similar assumption is utilized to prove that the desired local minimizer of  $Q_n(\boldsymbol{\beta}; \lambda)$  also be the global minimizer (Fan and Lv (2011)).

Condition (C.4) imposes a lower bound of the growing rate for the smallest signals. In fix  $p$  case, people usually assume  $\boldsymbol{\beta}$  is independent of  $n$ , under which such condition is always satisfied. However, in high-dimensional settings, the signals are allowed to grow with  $n$ , we do need a lower bound on the smallest coefficients. Without such conditions, (to the best of our knowledge) all the current result on oracle properties fail (see Chatterjee and Lahiri (2013); Huang and Xie (2007); Kim et al. (2008)).

We shall have a further discussion on condition (C.4) in the next Chapter where we propose the partially penalized regression. To be brief, condition (C.4) is too ideal for a practical problem, and it is hard to validate. We shall allow the existence of small coefficients and propose a cure for that case. In this Chapter, we assume (C.4) is true temporarily.

## 2.3 Existing Results and Comparisons

This section summarizes the existing results on the Oracle properties, and compares those results with ours.

Kim et al. (2008) shows the Oracle Property of SCAD penalty in high dimensions. They establish the selection consistency and asymptotic normal distribution by showing that the oracle estimator is one of the local minimizer of the loss function. The selection

consistency is achieved for  $p = O(n^\alpha)$ ,  $\alpha > 0$  if the noises satisfy  $E(\varepsilon_i^{2k}) < \infty$ ; or for  $p = O(\exp(cn))$  if noises are Gaussian distributed. (See Theorem 1 & 2 in (Kim et al., 2008)).

Under one more condition:  $p \leq n$  and  $\mathbf{C}_n$  is nonsingular, with the smallest eigenvalue greater than 0 and the largest eigenvalue bounded by  $M$ . They are able to show that the oracle estimator is a global estimator.

In summary, their conditions are quite similar to ours. It is worth mentioning that they also require a lower bound on the smallest signal, that is

$$\min_{1 \leq j \leq p_0} |\beta_j^*| > cn^{-1/2+q/2} \quad (2.12)$$

where  $q$  satisfies  $p_0 = o(n^q)$ . Comparing to our requirement (C.4), (2.12) is slightly stronger.

Another theoretical asymptotic result on SCAD can be found in (Huang and Xie, 2007, Xie and Huang, 2009), where they have a set of condition on relationship among eigenvalues of gram matrix, smallest signal and  $\lambda$ . However, their result need very strong conditions. For example, they require

$$\begin{aligned} \lim_{n \rightarrow \infty} p^2 / (n \cdot \rho_{\min}(\mathbf{C}_n)) &= 0 \\ \lim_{n \rightarrow \infty} p_0 \lambda_n^2 / \rho_{\min}(\mathbf{C}_n) &= 0 \end{aligned}$$

These two conditions are possibly true only when  $p \leq n$ , and even when  $p \leq n$  and  $\rho_{\min}(\mathbf{C}_n) > 0$ , these two conditions are hardly satisfied. In addition, their requirement

on the smallest signal is

$$\min_{1 \leq j \leq p_0} |\beta_j^*| \gg \max \left( \sqrt{p/(n\rho_{\min}(\mathbf{C}_n))}, \lambda_n p_0 / \sqrt{\rho_{\min}(\mathbf{C}_n)} \right) \quad (2.13)$$

Zhang (2010) theoretically shows the selection consistency,  $L_2$  consistency of the estimated coefficients when  $p \gg n$ . Although the asymptotic distribution of estimated coefficients is not mentioned, it follows directly from the selection consistency. We list some of their conditions:

1. Let  $d^*$  be a positive integer. The loss function (2.2) is “sparse convex” with rank  $d^*$  if it is convex in all models with fewer than  $d^*$  nonzero coefficients. This sparse convexity condition holds if

$$\kappa(P_\lambda) < \min_{|A| < d^*} \rho_{\min}(\mathbf{C}_{AA,n}), \text{ where } \mathbf{C}_{AA,n} = n^{-1} \mathbf{X}_A^T \mathbf{X}_A \quad (2.14)$$

2. The sparse Riesz condition (SRC) on  $\mathbf{X}$  is true: for suitable  $0 < c_* \leq c^* < \infty$  and rank  $d^*$ ,

$$c_* \leq \min_{|A| \leq d^*} \rho_{\min}(\mathbf{C}_{AA,n}) \leq \max_{|A| \leq d^*} \rho_{\max}(\mathbf{C}_{AA,n}) \leq c^*.$$

3.  $p_0 \leq d^*/(c^*/c_* + 1/2)$

The first two conditions are similar to our condition (C.3) and (C.2), respectively. In their conditions, the required relationships (2.14)(2.15) must be true for all sparse models  $A$ ; while in ours, it only needs to be true for the single true model. They have no requirements on the growth rate of  $p$ , instead, they need a condition on  $p_0$ ; while we requires  $p = O(n^\alpha)$ . As for the requirement on the smallest signal, they require something close to  $\min_{1 \leq j \leq p_0} |\beta_j^*| > c\sqrt{\log p/n}$ , which is weaker than ours.

Zheng et al. (2014) have some results on hard thresholding in high dimensions. In their paper, they use the concept of robust spark  $M = \text{rspar}k_c(\mathbf{X})$ , defined as following

$$M = \max\{k : \min_{\|\alpha\|_0 < k, \|\alpha\|_2 = 1} n^{-1/2} \|\mathbf{X}\alpha\|_2 \geq c\}.$$

based on which, they have the following requirement on the smallest signal, which is similar to ours.

- $p_0 < M/2$  and  $p_0 = o(n)$ , the smallest signal satisfies

$$b_0 = \min_{1 \leq j \leq p_0} |\beta_j| > (\sqrt{16/c^2} \vee 1) c^{-1} c_2 \sqrt{(2p_0 + 1) \log(p \vee n)/n},$$

where  $M$  is the robust spark of  $\mathbf{X}$  with bound  $c$  and  $c_2 > \sigma\sqrt{10}$  is a constant.

## 2.4 Oracle Property

In this section, we show that for any penalty function  $P_\lambda(\cdot)$  that belongs to  $\mathcal{P}$ , the estimator of  $\boldsymbol{\beta}$  enjoys the oracle properties in high dimensions.

### 2.4.1 General Results

**Theorem 2.1.** *Under condition (C.1) to (C.4), using any function  $P_\lambda(\cdot)$  from  $\mathcal{P}$ , if  $\lambda_n$  is chosen properly such that*

$$\lambda_n < a^{-1} \min_{1 \leq j \leq p_0} |\beta_j^*| \quad \text{and} \quad \frac{\lambda_n \sqrt{n}}{\sqrt{p_0 \log n}} \rightarrow \infty, \quad (2.15)$$

*then in the set  $\mathbb{N}(\boldsymbol{\beta}) = \mathbb{R}^{p_0} \oplus \{\mathbf{v} \in \mathbb{R}^{p-p_0} : \|\mathbf{v}\| = O(\rho_{\max}^{-1/2}(\mathbf{C}_{22,n}) p_0 \sqrt{\log n})\}$ , there exists a local minimizer of  $Q_n(\boldsymbol{\beta}; \lambda)$ , denoted as  $\hat{\boldsymbol{\beta}}^T = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)$ .  $\hat{\boldsymbol{\beta}}$  achieves the model selection*

consistency. *i.e.*

$$P(\hat{\boldsymbol{\beta}}_2 = \mathbf{0}) \rightarrow 1, \quad \text{as } n \rightarrow \infty \quad (2.16)$$

Theorem 2.1 shows that there exists a local minimizer which achieves the selection consistency, and it points out the neighborhood of the local minimizer. Although it is a local minimizer, people are always able to control the starting point within this neighborhood. In fact, by setting the initial  $\lambda$  sufficiently large, all the coefficients are estimated as 0, and obviously  $\mathbf{0}_p \in \mathbb{N}(\boldsymbol{\beta})$ . Therefore, the local minimizer conclusion is good enough.

As we can see, if the regularization parameters  $(\lambda, a)$  are chosen properly (based on the data), then the model selection consistency holds for all the functions in the family  $\mathcal{P}$  in high dimensions.

Regarding the choice of  $\lambda_n$ , there are two types of undesired behavior of  $\hat{\boldsymbol{\beta}}$ : (1)  $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|$  is too large, (2)  $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|$  is small, but  $\hat{\boldsymbol{\beta}}_2 \neq \mathbf{0}$ . Larger  $\lambda_n$  avoids the second type but facilitates the first type; while smaller  $\lambda_n$  avoids the first type but facilitates the second type. The choice of  $\lambda_n$  balances these two errors. If there is a gap between the two regions, then oracle property is established by picking  $\lambda_n$  in that gap; if not, the selection consistency can not be achieved.

In this literature, there is always a condition on the threshold of the smallest signal. In our work, it is condition (C.4). This lower bound of signals is indispensable because that researchers want all the nonzero coefficients to locate in the constant part of the penalty function. It depends on the maximum of a set of linear combinations of errors, thus different authors have different requirements on its value. We shall have a further discussion on this condition in Chapter 4.

Once the selection consistency is true, the asymptotic normality of  $\hat{\boldsymbol{\beta}}_1$  follows immediately. The following theorem shows that the local minimizer obtained in Theorem 2.1

has an asymptotic normal distribution.

**Theorem 2.2.** *Let  $\mathbf{A}_n$  be a  $q \times p_0$  matrix, if  $\mathbf{A}_n \mathbf{C}_{11,n} \mathbf{A}_n^T \rightarrow \mathbf{Q}$  and  $\max_i \|(\mathbf{A}_n \mathbf{X}_1^T)_{\cdot i}\| = o(\sqrt{n})$ . Under condition (C.1) to (C.4), the estimator of nonzero coefficients  $\hat{\boldsymbol{\beta}}_1$  converges in distribution to a multivariate normal distribution:*

$$\sqrt{n} \mathbf{A}_n \left[ (\mathbf{C}_{11,n} + \Gamma_n) (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) + \mathbf{b}_n \right] \rightarrow_d N(0, \sigma^2 \mathbf{Q}) \quad (2.17)$$

where

$$\mathbf{b}_n^T = (P'_{\lambda_n}(|\beta_1^*|) \text{sgn}(\beta_1^*), \dots, P'_{\lambda_n}(|\beta_{p_0}^*|) \text{sgn}(\beta_{p_0}^*)) \quad (2.18)$$

$$\Gamma_n = \text{diag}\{P''_{\lambda_n}(|\beta_1^*|), \dots, P''_{\lambda_n}(|\beta_{p_0}^*|)\} \quad (2.19)$$

given that the derivatives exist.

Theorem 2.1 & 2.2 shows the oracle property is achieved using our family in high dimensions.

From this result, we see that the oracle property is not something mysterious, but can be possessed by a class of functions. Let us revisit the requirements (P.1) - (P.4), and list the idea behind those requirements

- $P_\lambda(t)$  has to be constant to the right of a certain point, otherwise the estimator has a lower-order bias term.
- To the left of that point, the penalty needs to be greater than a lasso type function.
- The larger  $c$  is (see (P.4)), the easier the penalty achieves the oracle property.

**Remark 1.** *If the loss function is defined using likelihood:*

$$Q_n(\boldsymbol{\beta}; \lambda) = \sum_{i=1}^n l(\mathbf{X}_i, \mathbf{Y}_i, \boldsymbol{\beta}) - n \sum_{j=1}^{p_0} P_{\lambda_n}(|\beta_j|),$$

*a similar result on the oracle property is also true. Simply replacing  $\mathbf{C}_{11,n}$  by  $\mathbf{I}_1(\boldsymbol{\beta}_1^*)$ , where  $\mathbf{I}_1(\boldsymbol{\beta}_1^*)$  is the information matrix relevant to nonzero coefficients, all the conclusions and proofs should follow immediately.*

## 2.4.2 One-Step Estimator in High Dimensions

The One-Step Estimator (Zou and Li, 2008) is defined as  $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} Q_n(\boldsymbol{\beta}; \lambda)$ , where

$$Q_n(\boldsymbol{\beta}; \lambda) = \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n \sum_{j=1}^p P'_\lambda(|\tilde{\beta}_j|) |\beta_j|, \quad (2.20)$$

$P'_\lambda(t)$  is a penalty function such as SCAD, and  $\tilde{\boldsymbol{\beta}}$  is an initial estimator.

This estimator originates from the local linear approximation (LLA), which is an algorithm for maximizing the penalized likelihood of concave penalty functions. LLA solves the optimization problem of (2.2) by iteratively optimizing (2.20). Therefore, (2.20) is merely a single step in LLA, that is where the name comes from.

Note (2.20) is a convex function, and it is a single step of LLA, therefore it is much easier to optimize than nonconcave penalties. In addition, Zou and Li (2008) showed that the one-step estimator enjoys oracle property in fixed dimension as long as  $\tilde{\boldsymbol{\beta}}$  and  $\lambda$  are properly selected.

This method is similar to the adaptive LASSO: both require an initial estimator, and adjust the penalization based on it. Although the one-step estimator has been successful in fixed dimensions, its performance in high dimensions has not been explored. Here we are

going to prove that the one-step estimator also enjoys oracle properties in high dimensions.

Following similar approach in the previous section, we have the following theorem on its selection consistency.

**Theorem 2.3.** *If the penalty function satisfies:*

- (a)  $\exists a > 0$ , such that  $P'_\lambda(t) = 0$ , when  $t \geq a\lambda$ ,
- (b)  $\exists \alpha > 0$ , such that  $P_\lambda(t) > \alpha\lambda$  when  $|t| = o(\lambda)$ .

*Under condition (C.1)(C.2)(C.4), if  $\lambda_n$  is chosen properly such that*

$$\frac{1}{a} \min_{1 \leq j \leq p_0} |\beta_j^*| > \lambda_n \gg \sqrt{p_0 \log n/n}. \quad (2.21)$$

*As long as the initial estimator  $\tilde{\boldsymbol{\beta}}$  satisfies (for some constant  $K$ )*

$$P\left(\|\sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*)\|_\infty > K\sqrt{\log n}\right) = o(1)$$

*Then the one-step estimator  $\hat{\boldsymbol{\beta}}$  by (2.20) achieves the model selection consistency. i.e.*

$$P(\hat{\boldsymbol{\beta}}_2 = \mathbf{0}) \rightarrow 1, \quad \text{as } n \rightarrow \infty \quad (2.22)$$

Following theorem 2.3, the asymptotic distribution of one-step estimator can be easily obtained (similar to theorem 2.2). Therefore, we conclude that the one-step estimator also enjoys Oracle Properties.

## 2.5 Comparison of Popular Methods

In high dimensions, we have proven the Oracle Property of the oracle family  $\mathcal{P}$  and the one-step estimator, and we also know that ALASSO enjoys such properties (Chattergee and Lahiri, 2013). Once the selection consistency is achieved, we are more interested in their asymptotic bias and variance, and would like to make a comparison of them. This comparison will also reveal the effects that small coefficients have on the bias and variance.

### 2.5.1 Oracle Family $\mathcal{P}$

With the result of theorem 2.2, it is straightforward to obtain the bias and variance of  $\hat{\beta}_1$ :

$$\text{bias} = -(\mathbf{C}_{11,n} + \Gamma_n)^{-1} \mathbf{b}_n, \quad (2.23)$$

$$\text{var} = n^{-1} \sigma^2 (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \mathbf{C}_{11,n} (\mathbf{C}_{11,n} + \Gamma_n)^{-1}, \quad (2.24)$$

where  $\Gamma_n$  and  $\mathbf{b}_n$  are both defined in theorem 2.2. These expressions are true for all the penalties belonging to the family  $\mathcal{P}$ , including MCP and SCAD.

Let  $\boldsymbol{\alpha}$  be a vector of length  $p_0$  with  $\|\boldsymbol{\alpha}\| = 1$ , then the mean square errors(MSE) of  $\boldsymbol{\alpha}^T \hat{\beta}_1$  is as follows:

$$MSE(\boldsymbol{\alpha}^T \hat{\beta}_1) = \frac{\sigma^2}{n} [\boldsymbol{\alpha}^T (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \mathbf{C}_{11,n} (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \boldsymbol{\alpha}] + [\boldsymbol{\alpha}^T (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \mathbf{b}_n]^2. \quad (2.25)$$

We are going to discuss how the minimum MSE is achieved. We will consider two cases, and argue that the MSE in case 1 is always smaller than that in case 2.

Case 1 When  $\min_{1 \leq j \leq p_0} (|\beta_j^*|) > a\lambda_n$ , i.e. all the coefficients are large. By definition, we

have  $\mathbf{b}_n = \mathbf{0}$  and  $\Gamma_n = \mathbf{0}$ . That means,

$$\text{bias}_1(\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1) = 0, \quad (2.26)$$

$$\text{var}_1(\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1) = n^{-1} \sigma^2 (\boldsymbol{\alpha}^T \mathbf{C}_{11,n}^{-1} \boldsymbol{\alpha}) \quad (2.27)$$

Therefore, the MSE is  $n^{-1} \sigma^2 (\boldsymbol{\alpha}^T \mathbf{C}_{11,n}^{-1} \boldsymbol{\alpha})$ , and it is identical to that of the oracle estimator.

Case 2 Otherwise, there exist some coefficients whose magnitude is below  $a\lambda_n$ . That means,  $\mathbf{b}_n \neq \mathbf{0}$  and  $\Gamma_n \neq \mathbf{0}$ . And we have

$$\begin{aligned} \text{bias}_2(\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1) &= -\boldsymbol{\alpha}^T (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \mathbf{b}_n \neq \mathbf{0}, \\ \text{var}_2(\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1) &= n^{-1} \sigma^2 [\boldsymbol{\alpha}^T (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \mathbf{C}_{11,n} (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \boldsymbol{\alpha}] \end{aligned} \quad (2.28)$$

If we can show that  $\text{var}_1(\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1) \leq \text{var}_2(\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1)$  is true for any unit vector  $\boldsymbol{\alpha}$ , then it follows immediately that MSE in case 1 is less than that in case 2.

In other words, it remains to show that

$$\boldsymbol{\alpha}^T \mathbf{C}_{11,n}^{-1} \boldsymbol{\alpha} \leq \boldsymbol{\alpha}^T (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \mathbf{C}_{11,n} (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \boldsymbol{\alpha} \quad \forall \boldsymbol{\alpha}, \text{ where } \|\boldsymbol{\alpha}\| = 1. \quad (2.29)$$

$\Gamma_n$  is a diagonal matrix with its  $j$ th diagonal element being  $P''_{\lambda_n}(|\beta_j^*|)$ , which is either 0 (if  $|\beta_j^*| \geq a\lambda_n$ ) or a negative number. Therefore,  $\Gamma_n$  is negative semidefinite, i.e.  $\Gamma_n \preceq \mathbf{0}$ . Furthermore, according to condition (C.2), all the eigenvalues of  $\mathbf{C}_{11,n}$  are greater than  $\max_{1 \leq j \leq p_0} P''_{\lambda_n}(|\beta_j^*|)$ , thus  $\mathbf{C}_{11,n} + \Gamma_n$  is positive definite. Remember that if two matrices  $\mathbf{A}$  and  $\mathbf{B}$  are both positive definite, and if  $\mathbf{A}^{-1} \succeq \mathbf{B}^{-1}$ , then  $\mathbf{B} \succeq \mathbf{A}$ . (because  $\mathbf{B} - \mathbf{A} = \mathbf{A}(\mathbf{A}^{-1} - \mathbf{B}^{-1})\mathbf{B} \succeq \mathbf{0}$ .) There-

fore, (2.29) is true iff  $(\mathbf{C}_{11,n} + \Gamma_n)\mathbf{C}_{11,n}^{-1}(\mathbf{C}_{11,n} + \Gamma_n) \preceq \mathbf{C}_{11,n}$ . In fact, we have

$$(\mathbf{C}_{11,n} + \Gamma_n)\mathbf{C}_{11,n}^{-1}(\mathbf{C}_{11,n} + \Gamma_n) - \mathbf{C}_{11,n} = 2\Gamma_n + (-\Gamma_n)\mathbf{C}_{11,n}^{-1}(-\Gamma_n) \preceq \Gamma_n \preceq 0 \quad (2.30)$$

Thus we have shown that (2.29) is true.

Comparing case 1 and case 2, the variance in case 2 is always greater than that in case 1, and the bias in case 2 is non-zero while that in case 1 is exactly zero. Therefore, the smallest MSE of  $\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1$  is obtained when  $\min_{1 \leq j \leq p_0} (|\beta_j^*|) > a\lambda_n$ , and the minimum is  $\boldsymbol{\alpha}^T \mathbf{C}_{11}^{-1} \boldsymbol{\alpha}$ .

To summarize, the optimal MSE of oracle family  $\mathcal{P}$  is equal to the MSE of oracle properties. To achieve the optimum, the selected model must be true, and the magnitude of any nonzero coefficients must be greater than  $a\lambda_n$ . If there exists some coefficients whose magnitude is below  $a\lambda_n$ , then the variance is amplified, and a small bias also occurs. Therefore, small coefficients causes two problems: the selection accuracy and the amplification of variance.

## 2.5.2 Adaptive Lasso

Zou (2006) suggested the so-called adaptive lasso method, in which the slope of LASSO is determined by the initial estimator, i.e.

$$P_{\lambda_n}(|\beta_j|) = \lambda_n |\tilde{\beta}_j|^{-r} |\beta_j|,$$

where  $\tilde{\boldsymbol{\beta}}$  is the initial estimator, and  $r$  is a constant that adjusts the power. Its oracle property in high dimensions is achieved by (Chatterjee and Lahiri, 2013). For the asymptotic MSE of  $\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1$  of ALASSO, we have the following conclusion: the minimum

asymptotic MSE of  $\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1$  is

$$\frac{\sigma^2}{n} \boldsymbol{\alpha}^T \mathbf{C}_{11}^{-1} \boldsymbol{\alpha} + \lambda_n^2 (\boldsymbol{\alpha}^T \mathbf{C}_{11}^{-1} \mathbf{b}_0)^2 \quad (2.31)$$

where  $\mathbf{b}_0 = (\text{sgn}(\beta_1^*) |\beta_1^*|^{-r}, \dots, \text{sgn}(\beta_{p_0}^*) |\beta_{p_0}^*|^{-r})^T$ .

Following the proof in Zou (2006), instead of treating  $\hat{\omega}_j = |\tilde{\beta}_j|^{-r} = O_p(n^{r/2})$ , we use Taylor expansion to include the higher order terms in ALASSO:

$$\hat{\omega}_j = |\beta_j^*|^{-r} + [(-r) \text{sgn}(\beta_j^*) |\beta_j^*|^{-r-1} + o(1)] (\hat{\beta}_j - \beta_j^*), \text{ if } \beta_j^* \neq 0.$$

then the estimator of nonzero coefficients satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) = \left[ \mathbf{C}_{11}^{-1} W_n + \sqrt{n} \lambda_n r \mathbf{C}_{11}^{-1} \mathbf{D}_0 (\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) - \sqrt{n} \lambda_n \mathbf{C}_{11}^{-1} \mathbf{b}_0 \right] (1 + o_p(1)) \quad (2.32)$$

where  $W_n = \frac{\mathbf{x}_1^T \boldsymbol{\varepsilon}}{\sqrt{n}} \rightarrow_d N(0, \sigma^2 \mathbf{C}_{11})$ ,  $\mathbf{D}_0 = \text{diag}(|\beta_1^*|^{-r-1}, \dots, |\beta_{p_0}^*|^{-r-1})$ ,

and  $\mathbf{b}_0 = (\text{sgn}(\beta_1^*) |\beta_1^*|^{-r}, \dots, \text{sgn}(\beta_{p_0}^*) |\beta_{p_0}^*|^{-r})$ .

As long as  $\tilde{\boldsymbol{\beta}}$  is unbiased (which is usually true), the bias comes from the third term of (2.32). That is

$$\text{bias} = -\lambda_n \mathbf{C}_{11}^{-1} \mathbf{b}_0. \quad (2.33)$$

let  $r$  be fixed, then the bias increases as  $\lambda_n$  increases. Note that the bias term is on order  $O_p(\lambda_n)$ , and it never vanishes.

Variance comes from the first two terms. As long as  $\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\| = o_p(1)$ , the first term dominates the second term, thus the asymptotic variance of  $\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}$  is

$$\frac{\sigma^2}{n} \boldsymbol{\alpha}^T \mathbf{C}_{11}^{-1} \boldsymbol{\alpha} \quad (2.34)$$

The minimum MSE is obtained from (2.33) and (2.34).

Another relevant issue is the choice of  $r$ , and we are going to show that it is better to use some  $r > 1$ , in order to obtain smaller MSE after achieving selection consistency.

Intuitively, the oracle estimator should be the best choice of the initial estimator, so we plug in the oracle estimator for  $\tilde{\boldsymbol{\beta}}$  and see the MSE of  $\hat{\boldsymbol{\beta}}$ . In (2.32), replacing  $\tilde{\boldsymbol{\beta}}_1$  by  $\tilde{\boldsymbol{\beta}}_1^0 = (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{Y}$ , we got

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) = [\mathbf{C}_{11}^{-1} (\mathbf{I} + \lambda_n r \mathbf{D}_0 \mathbf{C}_{11}^{-1}) W_n - \sqrt{n} \lambda_n \mathbf{C}_{11}^{-1} \mathbf{b}_0] (1 + o_p(1)) \quad (2.35)$$

Thus the asymptotic covariance matrix of  $\tilde{\boldsymbol{\beta}}$  is

$$\text{AVar}(\tilde{\boldsymbol{\beta}}; \lambda_n) = \frac{\sigma^2}{n} \{ \mathbf{C}_{11}^{-1} + 2\lambda_n r \mathbf{C}_{11}^{-1} \mathbf{D}_0 \mathbf{C}_{11}^{-1} + (\lambda_n r)^2 \mathbf{C}_{11}^{-1} \mathbf{D}_0 \mathbf{C}_{11}^{-1} \mathbf{D}_0 \mathbf{C}_{11}^{-1} \} \quad (2.36)$$

Thereafter, the asymptotic MSE of  $\boldsymbol{\alpha}^T \tilde{\boldsymbol{\beta}}$  is a quadratic function of  $\lambda_n$ :

$$MSE = m_2 \lambda_n^2 + m_1 \lambda_n + m_0, \quad (2.37)$$

where

$$\begin{aligned} m_2 &= n^{-1} \sigma^2 r^2 \boldsymbol{\alpha}^T \mathbf{C}_{11}^{-1} \mathbf{D}_0 \mathbf{C}_{11}^{-1} \mathbf{D}_0 \mathbf{C}_{11}^{-1} \boldsymbol{\alpha} + (\boldsymbol{\alpha}^T \mathbf{C}_{11}^{-1} \mathbf{b}_0)^2, \\ m_1 &= 2n^{-1} \sigma^2 r \boldsymbol{\alpha}^T \mathbf{C}_{11}^{-1} \mathbf{D}_0 \mathbf{C}_{11}^{-1} \boldsymbol{\alpha}, \\ m_0 &= n^{-1} \sigma^2 \boldsymbol{\alpha}^T \mathbf{C}_{11}^{-1} \boldsymbol{\alpha}. \end{aligned} \quad (2.38)$$

Since  $m_2$  is always positive, the minimum of MSE is obtained at  $\lambda_n = -\frac{m_1}{2m_2}$ , and it is on the order  $O_p(1/n)$ . Remember that the condition to achieve oracle property is  $n^{1/2+r} \lambda_n \rightarrow \infty$ , therefore, in order to achieve oracle property and obtain minimum

MSE simultaneously, it is necessary that  $r > 1$ .

## 2.6 Proofs

We start with a lemma on the lower bound of  $P_{\lambda_n}(\beta_j + \frac{u_j}{\sqrt{n}}) - P_{\lambda_n}(\beta_j)$  where  $\beta_j \neq 0$ .

**Lemma 2.1.** *Define two sets as following:*

$$I_1 = \{j : \beta_j \cdot (\beta_j + \frac{u_j}{\sqrt{n}}) \geq 0\} \quad \text{and} \quad I_2 = \{j : \beta_j \cdot (\beta_j + \frac{u_j}{\sqrt{n}}) < 0\},$$

then

$$\begin{aligned} & P_{\lambda_n}(|\beta_j + \frac{u_j}{\sqrt{n}}|) - P_{\lambda_n}(|\beta_j|) \\ & \geq -\kappa(P_{\lambda_n}) \cdot \frac{u_j^2}{n} + P'_{\lambda_n}(|\beta_j|) \operatorname{sgn}(\beta_j) \left[ \frac{u_j}{\sqrt{n}} I(j \in I_1) - (\frac{u_j}{\sqrt{n}} + 2\beta_j) I(j \in I_2) \right] \end{aligned} \quad (2.39)$$

It turns out the lower bound is a quadratic function of  $u_j$ . This bound is much better than  $-\max_{t \geq 0} P_{\lambda_n}(|t|)$ , thus allows us to construct the oracle properties of the whole family.

*Proof of Lemma 2.1.* Let  $A_{jn}$  be  $P_{\lambda_n}(|\beta_j + \frac{u_j}{\sqrt{n}}|) - P_{\lambda_n}(|\beta_j|)$ , we shall consider the lower bound of  $A_n$  in two sets  $I_1$  and  $I_2$ , respectively.

- (1) If  $j \in I_1$ , for a fixed  $n$ , consider the function  $f(t) = P_{\lambda_n}(|t|)$ :  $f(\cdot)$  is continuous on  $[0, \infty)$  and differentiable on  $(0, \infty)$ . (And  $f(\cdot)$  is continuous on  $(-\infty, 0]$  and differentiable on  $(-\infty, 0)$ .) Therefore, as long as  $t_1$  and  $t_2$  are on the same side of 0, we can apply the mean value theorem to  $f(t_1) - f(t_2)$ .

When  $j \in I_1$ , we apply the mean value theorem to  $A_{jn}$

$$A_{jn} = P_{\lambda_n}(|\beta_j + \frac{u_j}{\sqrt{n}}|) - P_{\lambda_n}(|\beta_j|) = P'_{\lambda_n}(|\beta_j^{(0)}|)\text{sgn}(\beta_j^{(0)})\frac{u_j}{\sqrt{n}}, \quad (2.40)$$

where  $\beta_j^{(0)}$  is between  $\beta_j$  and  $\beta_j + \frac{u_j}{\sqrt{n}}$ . Furthermore, write (2.40) as

$$\begin{aligned} A_{jn} &= P'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j)\frac{u_j}{\sqrt{n}} + \left[ P'_{\lambda_n}(|\beta_j^{(0)}|)\text{sgn}(\beta_j^{(0)}) - P'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) \right] \cdot \frac{u_j}{\sqrt{n}} \\ &= P'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j)\frac{u_j}{\sqrt{n}} + \frac{P'_{\lambda_n}(|\beta_j^{(0)}|) - P'_{\lambda_n}(|\beta_j|)}{\frac{u_j}{\sqrt{n}}} \cdot \text{sgn}(\beta_j)\frac{u_j^2}{n} \end{aligned}$$

Note that the coefficient of the second term satisfies

$$\left| \frac{P'_{\lambda_n}(|\beta_j^{(0)}|) - P'_{\lambda_n}(|\beta_j|)}{\frac{u_j}{\sqrt{n}}} \cdot \text{sgn}(\beta_j) \right| \leq \left| \frac{P'_{\lambda_n}(|\beta_j^{(0)}|) - P'_{\lambda_n}(|\beta_j|)}{|\beta_j^{(0)}| - |\beta_j|} \right| \leq \kappa(P_{\lambda_n}) \quad (2.41)$$

Therefore,

$$A_{jn} \geq -\kappa(P_{\lambda_n}) \cdot \frac{u_j^2}{n} + P'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j)\frac{u_j}{\sqrt{n}}, \quad \text{if } j \in I_1. \quad (2.42)$$

(2) If  $j \in I_2$ , then  $\beta_j + \frac{u_j}{\sqrt{n}}$  and  $-\beta_j$  are on the same side of 0. Using mean value theorem, we have

$$A_{jn} = P_{\lambda_n}(|\beta_j + \frac{u_j}{\sqrt{n}}|) - P_{\lambda_n}(|-\beta_j|) = P'_{\lambda_n}(|\beta_j^{(-0)}|)\text{sgn}(\beta_j^{(-0)}) \cdot \left[ \frac{u_j}{\sqrt{n}} + 2\beta_j, \right] \quad (2.43)$$

where  $\beta_j^{(-0)}$  is between  $-\beta_j$  and  $\beta_j + \frac{u_j}{\sqrt{n}}$ . Similar to case (1), write (2.43) as

$$\begin{aligned}
A_{jn} &= \left[ P'_{\lambda_n}(|\beta_j^{(-0)}|) \operatorname{sgn}(\beta_j^{(-0)}) - P'_{\lambda_n}(|\beta_j|) \operatorname{sgn}(-\beta_j) \right] \cdot \left( \frac{u_j}{\sqrt{n}} + 2\beta_j \right) \\
&\quad + P'_{\lambda_n}(|\beta_j|) \operatorname{sgn}(-\beta_j) \left( \frac{u_j}{\sqrt{n}} + 2\beta_j \right) \\
&= P'_{\lambda_n}(|\beta_j|) \operatorname{sgn}(-\beta_j) \left( \frac{u_j}{\sqrt{n}} + 2\beta_j \right) \\
&\quad + \frac{P'_{\lambda_n}(|\beta_j^{(-0)}|) - P'_{\lambda_n}(|\beta_j|)}{\left( \frac{u_j}{\sqrt{n}} + 2\beta_j \right)} \cdot \operatorname{sgn}(-\beta_j) \left( \frac{u_j}{\sqrt{n}} + 2\beta_j \right)^2
\end{aligned}$$

Following the same argument as (2.41), we have

$$\left| \frac{P'_{\lambda_n}(|\beta_j^{(-0)}|) - P'_{\lambda_n}(|\beta_j|)}{\left( \frac{u_j}{\sqrt{n}} + 2\beta_j \right)} \cdot \operatorname{sgn}(-\beta_j) \right| \leq \left| \frac{P'_{\lambda_n}(|\beta_j^{(-0)}|) - P'_{\lambda_n}(|\beta_j|)}{|\beta_j^{(-0)}| - |\beta_j|} \right| \leq \kappa(P_{\lambda_n}) \quad (2.44)$$

Thus,

$$\begin{aligned}
A_{jn} &\geq P'_{\lambda_n}(|\beta_j|) \operatorname{sgn}(-\beta_j) \left( \frac{u_j}{\sqrt{n}} + 2\beta_j \right) - \kappa(P_{\lambda_n}) \left( \frac{u_j}{\sqrt{n}} + 2\beta_j \right)^2 \\
&\geq P'_{\lambda_n}(|\beta_j|) \operatorname{sgn}(-\beta_j) \left( \frac{u_j}{\sqrt{n}} + 2\beta_j \right) - \kappa(P_{\lambda_n}) \frac{u_j^2}{n} - \kappa(P_{\lambda_n}) \left[ 4 \cdot \beta_j \cdot \left( \beta_j + \frac{u_j}{\sqrt{n}} \right) \right]
\end{aligned} \quad (2.45)$$

Since  $j \in I_2$ ,  $\beta_j \cdot \left( \beta_j + \frac{u_j}{\sqrt{n}} \right) < 0$ ; besides,  $\kappa(P_{\lambda_n}) \geq 0$ . Thus the last term in (2.45) is positive. We have

$$A_{jn} \geq P'_{\lambda_n}(|\beta_j|) \operatorname{sgn}(-\beta_j) \left( \frac{u_j}{\sqrt{n}} + 2\beta_j \right) - \kappa(P_{\lambda_n}) \frac{u_j^2}{n}, \quad \text{if } j \in I_2. \quad (2.46)$$

Putting (2.42) and (2.45) together, we conclude that

$$\begin{aligned}
& P_{\lambda_n}(|\beta_j + \frac{u_j}{\sqrt{n}}|) - P_{\lambda_n}(|\beta_j|) \\
& \geq -\kappa(P_{\lambda_n}) \cdot \frac{u_j^2}{n} + P'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) \left[ \frac{u_j}{\sqrt{n}}I(j \in I_1) - (\frac{u_j}{\sqrt{n}} + 2\beta_j)I(j \in I_2) \right] \quad (2.47)
\end{aligned}$$

□

**Lemma 2.2.** *Under conditions (C.2), let  $\boldsymbol{\varepsilon}_n$  be the  $n \times 1$  vector of errors,*

- (i)  $P\left(\|\frac{1}{\sqrt{n}}\mathbf{X}_1^T\boldsymbol{\varepsilon}_n\| > K\sqrt{p_0\log n}\right) = O(p_n \cdot n^{-(r-2)/2})$ .
- (ii)  $P\left(\|\frac{1}{\sqrt{n}}\mathbf{X}_l^T\boldsymbol{\varepsilon}_n\|_\infty > K\sqrt{\log n}\right) = O(p^{(l)} \cdot n^{-(r-2)/2})$ , for  $l = 0, 1, 2$ .
- (iii)  $P(\|\frac{1}{\sqrt{n}}\boldsymbol{\varepsilon}_n\| > \sqrt{\log n}) = O((\log n)^{-1})$ .

*Proof of Lemma 2.2.* The first two statements are obtained from (Chattergee and Lahiri, 2013), for the last one:

$$P(\|\boldsymbol{\varepsilon}_n/\sqrt{n}\| > \sqrt{\log n}) = P(n^{-1}\boldsymbol{\varepsilon}_n^T\boldsymbol{\varepsilon}_n > \log n) \leq \frac{n^{-1}\sum_{j=1}^n E(\varepsilon_{j,n}^2)}{\log n} \quad (2.48)$$

Following (2.48), by central limit theorem,

$$\exists M \in \mathbb{R}, \text{ such that } P(\|\boldsymbol{\varepsilon}_n/\sqrt{n}\| > \sqrt{\log n}) \leq M\sigma^2 \cdot (\log n)^{-1} = O((\log n)^{-1}).$$

□

*Proof of theorem 2.1.* First of all, for any fixed  $\mathbf{u}^T = (\mathbf{u}_1^T, \mathbf{u}_2^T) \in \mathbb{R}^p$ , define the set

$$\begin{aligned}
B_n = & \left\{ \|\frac{1}{\sqrt{n}}\boldsymbol{\varepsilon}_n\| \leq \sqrt{\log n} \right\} \cap \left\{ \frac{1}{\sqrt{n}}\mathbf{X}_1^T\boldsymbol{\varepsilon}_n \leq K\sqrt{p_0\log n} \right\} \\
& \cap \left\{ \|\frac{1}{\sqrt{n}}\mathbf{X}_2^T\boldsymbol{\varepsilon}_n\|_\infty \leq K\sqrt{\log n} \right\} \quad (2.49)
\end{aligned}$$

Directly from Lemma 2.2, we can show that  $P(B_n) \rightarrow 1$ , as  $n \rightarrow \infty$ .

In the following steps, we will show the model selection consistency always holds in  $B_n$ . As  $P(B_n) \rightarrow 1$ , it follows that the selection consistency holds with probability tending to 1.

Let  $\boldsymbol{\beta}^*$  be the true value of coefficients. Define

$$V_n(\mathbf{u}) = Q_n(\boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}; \lambda_n) - Q_n(\boldsymbol{\beta}^*; \lambda_n), \quad (2.50)$$

and define  $\mathbb{N}(\mathbf{u}) = \mathbb{R}_0^p \oplus \{\mathbf{v} \in \mathbb{R}^{p-p_0} : \|\mathbf{v}\| = O(\rho_{\max}^{-1/2}(\mathbf{C}_{22,n})p_0\sqrt{\log n/n})\}$ . The model selection consistency (2.16) is achieved if and only if

$$\operatorname{argmin}_{\mathbf{u} \in \mathbb{N}(\mathbf{u})} V_n(\mathbf{u}) = \operatorname{argmin}_{\mathbf{u}_1 \in \mathbb{R}^{p_0}} V_n(\mathbf{u}_1, \mathbf{0}) \quad (2.51)$$

To prove (2.51), we are going to show that with probability tending to 1,  $\mathbf{u}^T = (\mathbf{u}_1^T, \mathbf{u}_2^T) \in \mathbb{N}(\mathbf{u})$  cannot be the global minimizer if it belongs to any of the following two sets:

- (1)  $D_{1,n} = \{\mathbf{u} \in \mathbb{N}(\mathbf{u}) : \|\mathbf{u}_1\|_2 > M_n\}$ ,
- (2)  $D_{1,n} = \{\mathbf{u} \in \mathbb{N}(\mathbf{u}) : \|\mathbf{u}_1\|_2 \leq M_n \text{ and } \mathbf{u}_2 \neq \mathbf{0}\}$ ,

where  $M_n$  is the boundary of the two sets satisfying  $M_n/\sqrt{p_0 \log n} \rightarrow \infty$ . Define  $\Delta_n = \frac{1}{2}(1 - \delta)c_1 - \kappa(P_{\lambda_n})$ , which is a positive number by condition (C.3). Also define  $M_n = 2\sqrt{\log n} \cdot \max \left\{ \Delta_n^{-1} \rho_{\max}^{1/2}(\mathbf{C}_{11,n}), \sqrt{2\delta(1 - \delta)^{-2} \Delta_n^{-1}}, \sqrt{\delta^{-1} \rho_{\max}^{-1}(\mathbf{C}_{11,n})} \right\}$ , and  $\|\mathbf{u}_1\|_{\mathbf{C}_{11,n}} = \sqrt{\mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1}$ .

(1) With simple algebra, it can be seen that

$$\begin{aligned}
V_n(\mathbf{u}) &= \frac{1}{2} \mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1 + \frac{1}{2} \mathbf{u}_2^T \mathbf{C}_{22,n} \mathbf{u}_2 + \mathbf{u}_1^T \mathbf{C}_{12,n} \mathbf{u}_2 - \frac{1}{\sqrt{n}} \mathbf{u}_1^T \mathbf{X}_1^T \boldsymbol{\varepsilon}_n - \frac{1}{\sqrt{n}} \mathbf{u}_2^T \mathbf{X}_2^T \boldsymbol{\varepsilon}_n \\
&\quad + n \sum_{j=1}^{p_0} \left[ P_{\lambda_n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}|) - P_{\lambda_n}(|\beta_j^*|) \right] + n \sum_{j=p_0+1}^p P_{\lambda_n}(|\frac{u_j}{\sqrt{n}}|) \quad (2.52)
\end{aligned}$$

By condition (C.1),

$$|\mathbf{u}_1^T \mathbf{C}_{12,n} \mathbf{u}_2| \leq \delta \sqrt{(\mathbf{u}_2^T \mathbf{C}_{22,n} \mathbf{u}_2)(\mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1)} \leq \frac{1}{2} \delta (\mathbf{u}_2^T \mathbf{C}_{22,n} \mathbf{u}_2) + \frac{1}{2} \delta (\mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1)$$

For any  $\mathbf{u} \in D_{1,n}$ , consider the difference

$$\begin{aligned}
&V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{0}, \mathbf{u}_2) \\
&= \frac{1}{2} \mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1 + \mathbf{u}_1^T \mathbf{C}_{12,n} \mathbf{u}_2 - \frac{1}{\sqrt{n}} \mathbf{u}_1^T \mathbf{X}_1^T \boldsymbol{\varepsilon}_n + n \sum_{j=1}^{p_0} \left[ P_{\lambda_n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}|) - P_{\lambda_n}(|\beta_j^*|) \right] \\
&\geq \frac{1}{2} (1 - \delta) \mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1 - \frac{1}{\sqrt{n}} \mathbf{u}_1^T \mathbf{X}_1^T \boldsymbol{\varepsilon}_n + n \sum_{j=1}^{p_0} \left[ P_{\lambda_n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}|) - P_{\lambda_n}(|\beta_j^*|) \right] \\
&\quad - \frac{1}{2} \delta (\mathbf{u}_2^T \mathbf{C}_{22,n} \mathbf{u}_2) \quad (2.53)
\end{aligned}$$

Define

$$S_1 = \frac{1}{2} (1 - \delta) \mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1 - \frac{1}{\sqrt{n}} \mathbf{u}_1^T \mathbf{X}_1^T \boldsymbol{\varepsilon}_n + n \sum_{j=1}^{p_0} \left[ P_{\lambda_n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}|) - P_{\lambda_n}(|\beta_j^*|) \right] \quad (2.54)$$

From lemma 2.1, we have

$$\begin{aligned}
& n \sum_{j=1}^{p_0} \left[ P_{\lambda_n} \left( \left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| \right) - P_{\lambda_n} (|\beta_j^*|) \right] \\
& \geq -\kappa(P_{\lambda_n}) \cdot \sum_{j=1}^{p_0} u_j^2 + n \sum_{j=1}^{p_0} P'_{\lambda_n} (|\beta_j^*|) \operatorname{sgn}(\beta_j^*) \left[ \frac{u_j}{\sqrt{n}} I(j \in I_1) - \left( \frac{u_j}{\sqrt{n}} + 2\beta_j \right) I(j \in I_2) \right] \\
& \geq -\kappa(P_{\lambda_n}) \cdot \sum_{j=1}^{p_0} u_j^2 \quad (\text{b/c } \lambda_n < a^{-1} \min_{1 \leq j \leq p_0} |\beta_j^*| \text{ by assumption (2.15)})
\end{aligned}$$

Thus, in the set  $B_n$ , we have

$$\begin{aligned}
S_1 & \geq \frac{1}{2}(1 - \delta) \mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1 - \frac{1}{\sqrt{n}} \mathbf{u}_1^T \mathbf{X}_1^T \boldsymbol{\varepsilon}_n - \kappa(P_{\lambda_n}) \cdot \sum_{j=1}^{p_0} u_j^2 \\
& \geq \mathbf{u}_1^T \left[ \frac{1}{2}(1 - \delta) \mathbf{C}_{11,n} - \kappa(P_{\lambda_n}) \mathbf{I} \right] \mathbf{u}_1 - \|\mathbf{u}_1\|_2 \cdot K \sqrt{p_0 \log n} \\
& \geq \Delta_n \|\mathbf{u}_1\|_2^2 - \|\mathbf{u}_1\|_2 \cdot K \sqrt{p_0 \log n}. \tag{2.55}
\end{aligned}$$

Combining (2.53) (2.55), for any fixed  $\mathbf{u} \in D_{1,n}$ , on set  $B_n$ , we have

$$\begin{aligned}
& V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{0}, \mathbf{u}_2) \\
& \geq S_1 - \frac{1}{2} \delta (\mathbf{u}_2^T \mathbf{C}_{22,n} \mathbf{u}_2) \\
& \geq \Delta_n \|\mathbf{u}_1\|_2^2 - \|\mathbf{u}_1\|_2 \cdot K \sqrt{p_0 \log n} - \frac{1}{2} \delta \|\mathbf{u}_2\|_2^2 \cdot \rho_{\max}(\mathbf{C}_{22,n}) \\
& \geq \|\mathbf{u}_1\|_2 (\Delta_n \|\mathbf{u}_1\|_2 - K \sqrt{p_0 \log n}) - O(p_0^2 \log n/n) \\
& \geq M_n (\Delta_n M_n - K \sqrt{p_0 \log n}) - O(p_0^2 \log n/n) \\
& > 0 \quad (\text{by the definition of } M_n) \tag{2.56}
\end{aligned}$$

By (2.56), we have shown that for any  $\mathbf{u} \in D_{1,n}$ ,

$$P [ V_n(\mathbf{u}_1, \mathbf{u}_2) > V_n(\mathbf{0}, \mathbf{u}_2) ] \geq P(B_{n,\mathbf{u}}) \rightarrow 1 \quad (2.57)$$

That is, with probability tending to 1,  $\mathbf{u} \in D_{1,n}$  cannot be the local minimizer of  $V_n(\mathbf{u})$ .

(2) for any  $\mathbf{u} \in D_{2,n}$ , on set  $B_n$ , consider the following difference

$$\begin{aligned} & V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}) \\ &= \frac{1}{2} \mathbf{u}_2^T \mathbf{C}_{22,n} \mathbf{u}_2 + \mathbf{u}_1^T \mathbf{C}_{12,n} \mathbf{u}_2 - \frac{1}{\sqrt{n}} \mathbf{u}_2^T \mathbf{X}_2^T \boldsymbol{\varepsilon}_n + n \sum_{j=p_0+1}^p P_{\lambda_n}(|\frac{u_j}{\sqrt{n}}|). \end{aligned} \quad (2.58)$$

We shall discuss two cases:

(2.1) There exists  $i \geq p_0 + 1$ , such that  $u_i/\sqrt{n} > a\lambda_n$ . Which means that Following (2.58), we have

$$\begin{aligned} & V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}) \\ &\geq \frac{1}{2}(1 - \delta) \mathbf{u}_2^T \mathbf{C}_{22,n} \mathbf{u}_2 - \frac{1}{2} \delta \mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1 - \left\| \frac{1}{\sqrt{n}} \mathbf{X}_2 \mathbf{u}_2 \right\| \|\boldsymbol{\varepsilon}\| + n P_{\lambda_n}(|\frac{u_i}{\sqrt{n}}|) \\ &\geq -\frac{1}{2} \delta \rho_{\max}(\mathbf{C}_{11,n}) \|\mathbf{u}_1\|_2^2 - \rho_{\max}^{1/2}(\mathbf{C}_{22,n}) \|\mathbf{u}_2\|_2 \sqrt{n \log n} + n P_{\lambda_n}(|\frac{u_i}{\sqrt{n}}|) \\ &\geq -\frac{1}{2} \delta \rho_{\max}(\mathbf{C}_{11,n}) M_n^2 - O(p_0 \log n) + n P_{\lambda_n}(|\frac{u_i}{\sqrt{n}}|) \\ &\geq n P_{\lambda_n}(|\frac{u_i}{\sqrt{n}}|) - \frac{1}{2} \delta \rho_{\max}(\mathbf{C}_{11,n}) M_n^2. \end{aligned} \quad (2.59)$$

The last inequality is by the definition of  $M_n/\sqrt{p_0 \log n} \rightarrow \infty$ .

Note that by the assumption (p.3), the penalty function  $P_{\lambda_n}(\cdot)$  must satisfy

$$\begin{cases} P_{\lambda_n}(|t|) > c\lambda_n|t|, & \text{if } |t| \leq a\lambda_n \\ P_{\lambda_n}(|t|) > a \cdot c\lambda_n^2, & \text{if } |t| > a\lambda_n \end{cases} \quad (2.60)$$

Combining (2.59) and (2.60), we have

$$V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}) > a \cdot c \cdot n\lambda_n^2 - \frac{1}{2}\delta\rho_{\max}(\mathbf{C}_{11,n})M_n^2 > 0, \quad (2.61)$$

The last inequality is true because the assumption on  $\lambda_n$  (2.15) leads to

$$\lambda_n > \frac{M_n}{\sqrt{n}} \cdot \sqrt{\frac{1}{2ac}\delta\rho_{\max}(\mathbf{C}_{11,n})} \text{ for sufficiently large } n.$$

(2.2) Otherwise,  $\forall j \geq p_0 + 1$ ,  $u_j/\sqrt{n} \leq a\lambda_n$ . We bound the difference (2.58) in another way.

Define  $\mathbf{e}_j$  as the basis vector with the  $j$ -th element 1 and all others 0, on the set  $B_n$ , we have

$$\begin{aligned} & V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}) \\ & \geq \mathbf{u}_1^T \mathbf{C}_{12,n} \mathbf{u}_2 - \frac{1}{\sqrt{n}} \mathbf{u}_2^T \mathbf{X}_2^T \boldsymbol{\varepsilon}_n + n \sum_{j=p_0+1}^p P_{\lambda_n}(|\frac{u_j}{\sqrt{n}}|) \\ & \geq - \sum_{j=p_0+1}^p |\mathbf{u}_1^T \mathbf{C}_{12,n} \mathbf{e}_j| |u_j| - \sum_{j=p_0+1}^p \left\| \frac{1}{\sqrt{n}} \mathbf{X}_2^T \boldsymbol{\varepsilon}_n \right\|_{\infty} |u_j| + \sum_{j=p_0+1}^p \sqrt{n} \cdot c\lambda_n |u_j| \\ & \geq \sum_{j=p_0+1}^p |u_j| \left[ \sqrt{n} \cdot c\lambda_n - \left\| \frac{\mathbf{x}_1 \mathbf{u}_1}{\sqrt{n}} \right\|_2 \left\| \frac{\mathbf{x}_2 \mathbf{e}_j}{\sqrt{n}} \right\|_2 - K\sqrt{\log n} \right] \\ & \geq \sum_{j=p_0+1}^p |u_j| \left[ \sqrt{n} \cdot c\lambda_n - \rho_{\max}^{1/2}(\mathbf{C}_{11,n}) M_n \cdot \left\| \frac{\mathbf{x}_j}{\sqrt{n}} \right\|_2 - K\sqrt{\log n} \right] \\ & \geq \sum_{j=p_0+1}^p |u_j| \left[ \sqrt{n} \cdot c\lambda_n - \rho_{\max}^{1/2}(\mathbf{C}_{11,n}) M_n - K\sqrt{\log n} \right] \end{aligned} \quad (2.62)$$

Similarly, by the assumption on  $\lambda_n$  (2.15),  $\lambda_n > \frac{M_n}{\sqrt{n}} \cdot \sqrt{\frac{1}{c} \delta \rho_{\max}^{1/2}(\mathbf{C}_{11,n})}$  for sufficiently large  $n$ . Thus (2.62) is greater than 0 on the set  $B_n$ .

Case (2.1) and (2.2) cover all the possibilities in  $D_{2,n}$ , therefore we conclude that

$$P(V_n(\mathbf{u}_1, \mathbf{u}_2) > V_n(\mathbf{u}_1, \mathbf{0})) \geq P(B_n) \rightarrow 1$$

So far, we have shown that the local minimizer of  $V_n(\mathbf{u})$  cannot be in  $\cup_{i=1}^2 D_{i,n}$ . Therefore, we conclude that

$$\operatorname{argmin}_{\mathbf{u} \in \mathbb{N}(\mathbf{u})} V_n(\mathbf{u}) \in \mathbb{N}(\mathbf{u}) / (\cup_{i=1}^2 D_{i,n}) = \{\mathbf{u} \in \mathbb{R}^p : \mathbf{u}_2 = \mathbf{0}\}, \quad (2.63)$$

Note that the local minimizer  $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{N}(\boldsymbol{\beta})} Q_n(\boldsymbol{\beta}; \lambda_n) = \sqrt{n} \operatorname{argmin}_{\mathbf{u} \in \mathbb{N}(\mathbf{u})} V_n(\mathbf{u}) + \boldsymbol{\beta}^*$ , therefore (2.63) is equivalent to

$$P(\hat{\boldsymbol{\beta}}_2 = \mathbf{0}) \rightarrow 1, \quad \text{as } n \rightarrow \infty$$

□

*Proof of Theorem 2.* By the conclusion of Theorem 1, with probability tending to one, the global minimizer  $\hat{\boldsymbol{\beta}}$  is of form  $(\hat{\boldsymbol{\beta}}_1, \mathbf{0})^T$ , hence we take the derivative wrt  $\hat{\boldsymbol{\beta}}_1$

$$\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} \Big|_{(\hat{\boldsymbol{\beta}}_1^T, \mathbf{0})^T} = 0.$$

Since

$$\begin{aligned} \frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_j} \Big|_{(\hat{\boldsymbol{\beta}}_1, \mathbf{0})^T} &= -\mathbf{X}_{\cdot j}^T(\mathbf{Y}_1 - \mathbf{X}_1 \boldsymbol{\beta}_1^*) + \mathbf{X}_{\cdot j}^T \mathbf{X}_1 (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) + n \left\{ P'_{\lambda_n}(|\beta_j^*|) \text{sgn}(\beta_j^*) \right. \\ &\quad \left. + \left[ P''_{\lambda_n}(|\beta_j^*|) + o_p(1) \right] (\hat{\beta}_j - \beta_j^*) \right\}, \quad j = 1, \dots, p_0 \end{aligned} \quad (2.64)$$

write the equation above into a matrix format, we have

$$\frac{1}{\sqrt{n}} \mathbf{X}_1^T (\mathbf{Y}_1 - \mathbf{X}_1 \boldsymbol{\beta}_1^*) + o_p(1) = \sqrt{n} \left[ (\mathbf{C}_{11,n} + \Gamma_n) (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) + \mathbf{b}_n \right].$$

It follows that

$$\sqrt{n} \mathbf{A}_n \left[ (\mathbf{C}_{11,n} + \Gamma_n) (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) + \mathbf{b}_n \right] = \frac{1}{\sqrt{n}} \mathbf{A}_n \mathbf{X}_1^T \boldsymbol{\varepsilon} + o_p(1)$$

Define  $\mathbf{Y}_{ni} = \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_i) \varepsilon_i$ , where  $\mathbf{e}_i$  denotes the basis vector with the  $i$ -th element being 1 and all the others being 0. The LHS is equal to  $\sum_{i=1}^n \mathbf{Y}_{ni} + o_p(1)$ , we want to show LHS has an asymptotic distribution using multivariate Lindeberg central limit theorem.

As we can see,  $\{\mathbf{Y}_{ni} : 1 \leq i \leq n\}$  is a collection of independent  $q$ -dimensional random vectors with  $E(\mathbf{Y}_{ni}) = \mathbf{0}$  and

$$\begin{aligned} \lim_{n \rightarrow \infty} \sum_{i=1}^n E(\mathbf{Y}_{ni} \mathbf{Y}_{ni}^T) &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{1}{n} \sigma^2 \mathbf{A}_n \mathbf{X}_1^T (\mathbf{e}_i \mathbf{e}_i^T) \mathbf{X}_1 \mathbf{A}_n^T = \lim_{n \rightarrow \infty} \frac{1}{n} \sigma^2 \mathbf{A}_n \mathbf{X}_1^T \left( \sum_{i=1}^n \mathbf{e}_i \mathbf{e}_i^T \right) \mathbf{X}_1 \mathbf{A}_n^T \\ &= \lim_{n \rightarrow \infty} \sigma^2 \mathbf{A}_n \mathbf{C}_{11,n} \mathbf{A}_n^T = \sigma^2 \mathbf{Q} \end{aligned}$$

by the multivariate Lindeberg central limit theorem and Slutsky's theorem, the multi-

variate normality is true if we can show for all  $\epsilon > 0$ ,

$$\sum_{i=1}^n E \|\mathbf{Y}_{ni}\|^2 \mathbf{I}(\|\mathbf{Y}_{ni}\| > \epsilon) \rightarrow 0.$$

In fact,

$$\begin{aligned} & \sum_{i=1}^n E \|\mathbf{Y}_{ni}\|^2 \mathbf{I}(\|\mathbf{Y}_{ni}\| > \epsilon) \\ & \leq \sum_{i=1}^n \left\| \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_i) \right\|^2 E \varepsilon_i^2 \mathbf{I} \left( \left\| \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_i) \right\| \cdot |\varepsilon_i| > \epsilon \right) \\ & \leq \left( \sum_{i=1}^n \left\| \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_i) \right\|^2 \right) \cdot \max_{1 \leq j \leq n} E \varepsilon_j^2 \mathbf{I} \left( \left\| \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_i) \right\| \cdot |\varepsilon_i| > \epsilon \right) \end{aligned} \quad (2.65)$$

By condition, we have  $\max_i \left\| \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_i) \right\| = o(1)$ , thus

$$\max_{1 \leq j \leq n} E \varepsilon_j^2 \mathbf{I} \left( \left\| \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_i) \right\| \cdot |\varepsilon_i| > \epsilon \right) \rightarrow 0, \quad \text{by Dominated Convergence Theorem} \quad (2.66)$$

Besides,

$$\sum_{i=1}^n \left\| \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_i) \right\|^2 = \text{trace} \left( \frac{1}{n} \mathbf{X}_1 \mathbf{A}_n^T \mathbf{A}_n \mathbf{X}_1^T \right) = \text{trace} (\mathbf{A}_n \mathbf{C}_{11,n} \mathbf{A}_n^T) \rightarrow \text{trace}(\mathbf{Q}) \quad (2.67)$$

Combining (2.65)(2.66)(2.67), we have  $\sum_{i=1}^n E \|\mathbf{Y}_{ni}\|^2 \mathbf{I}(\|\mathbf{Y}_{ni}\| > \epsilon) \rightarrow 0$ .

In conclusion,

$$\sqrt{n} \mathbf{A}_n \left[ (\mathbf{C}_{11,n} + \Gamma_n)(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) + \mathbf{b}_n \right] \rightarrow_d N(0, \sigma^2 \mathbf{Q}) \quad (2.68)$$

□

*proof of theorem 2.3.* First of all, define the set

$$B_n = \left\{ \left\| \frac{1}{\sqrt{n}} \boldsymbol{\varepsilon}_n \right\| \leq \sqrt{\log n} \right\} \cap \left\{ \frac{1}{\sqrt{n}} \mathbf{X}_1^T \boldsymbol{\varepsilon}_n \leq K \sqrt{p_0 \log n} \right\} \\ \cap \left\{ \left\| \frac{1}{\sqrt{n}} \mathbf{X}_l^T \boldsymbol{\varepsilon}_n \right\|_\infty \leq K \sqrt{\log n} \right\} \cap \left\{ \left\| \sqrt{n} (\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*) \right\|_\infty \leq K \sqrt{\log n} \right\} \quad (2.69)$$

Directly from Lemma 2.2, we can show that  $P(B_n) \rightarrow 1$ , as  $n \rightarrow \infty$ .

In the following steps, we will show that the model selection consistency always holds in  $B_n$ . Since  $P(B_n) \rightarrow 1$ , it follows immediately that the selection consistency holds with probability tending to 1.

Let  $\boldsymbol{\beta}^*$  be the true value of coefficients. Define

$$V_n(\mathbf{u}) = Q_n(\boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}; \lambda_n) - Q_n(\boldsymbol{\beta}^*; \lambda_n) \quad (2.70)$$

The model selection consistency (2.22) is achieved if and only if

$$\operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^p} V_n(\mathbf{u}) = \operatorname{argmin}_{\mathbf{u}_1 \in \mathbb{R}^{p_0}} V_n(\mathbf{u}_1, \mathbf{0}) \quad (2.71)$$

To prove (2.71), we are going to show that with probability tending to 1,  $\mathbf{u}^T = (\mathbf{u}_1^T, \mathbf{u}_2^T) \in \mathbb{R}^p$  cannot be the global minimizer if  $\mathbf{u} \in \{(\mathbf{u}_1, \mathbf{u}_2) : \mathbf{u}_2 = \mathbf{0}\}^c$ .

Define  $M_n = 2(1 - \delta)^{-1} \rho_{\min}^{-1}(\mathbf{C}_{11,n}) K \sqrt{p_0 \log n}$ , we split  $\{(\mathbf{u}_1, \mathbf{u}_2) : \mathbf{u}_2 = \mathbf{0}\}^c$  into the following two sets:

- (1)  $D_{1,n} = \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}_1\|_2 > M_n\}$
- (2)  $D_{2,n} = \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}_1\|_2 \leq M_n\}$

We will show that the global minimizer can not belong to either  $D_{1,n}$  or  $D_{2,n}$ .

(1) We start with  $D_{1,n}$ : with simple algebra, it can be seen that

$$\begin{aligned}
V_n(\mathbf{u}) &= \frac{1}{2} \mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1 + \frac{1}{2} \mathbf{u}_2^T \mathbf{C}_{22,n} \mathbf{u}_2 + \mathbf{u}_1^T \mathbf{C}_{12,n} \mathbf{u}_2 - \frac{1}{\sqrt{n}} \mathbf{u}_1^T \mathbf{X}_1^T \boldsymbol{\varepsilon}_n - \frac{1}{\sqrt{n}} \mathbf{u}_2^T \mathbf{X}_2^T \boldsymbol{\varepsilon}_n \\
&\quad + n \sum_{j=1}^{p_0} P'_\lambda(|\tilde{\beta}_j|) \left[ (|\beta_j^* + \frac{u_j}{\sqrt{n}}|) - (|\beta_j^*|) \right] + n \sum_{j=p_0+1}^p P'_\lambda(|\tilde{\beta}_j|) \left( \frac{|u_j|}{\sqrt{n}} \right)
\end{aligned} \tag{2.72}$$

By condition (C.1),

$$|\mathbf{u}_1^T \mathbf{C}_{12,n} \mathbf{u}_2| \leq \delta \sqrt{(\mathbf{u}_2^T \mathbf{C}_{22,n} \mathbf{u}_2)(\mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1)} \leq \frac{1}{2} \delta (\mathbf{u}_2^T \mathbf{C}_{22,n} \mathbf{u}_2) + \frac{1}{2} \delta (\mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1)$$

Thus,

$$\begin{aligned}
V_n(\mathbf{u}) &\geq \frac{1}{2} (1 - \delta) \mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1 - \frac{1}{\sqrt{n}} \mathbf{u}_1^T \mathbf{X}_1^T \boldsymbol{\varepsilon}_n + n \sum_{j=1}^{p_0} P'_\lambda(|\tilde{\beta}_j|) \left[ (|\beta_j^* + \frac{u_j}{\sqrt{n}}|) - (|\beta_j^*|) \right] \\
&\quad + \frac{1}{2} (1 - \delta) \mathbf{u}_2^T \mathbf{C}_{22,n} \mathbf{u}_2 - \frac{1}{\sqrt{n}} \mathbf{u}_2^T \mathbf{X}_2^T \boldsymbol{\varepsilon}_n + n \sum_{j=p_0+1}^p P'_\lambda(|\tilde{\beta}_j|) \left( \frac{|u_j|}{\sqrt{n}} \right) \\
&= S_1 + S_2.
\end{aligned} \tag{2.73}$$

By condition, we have

$$\min_{1 \leq j \leq p_0} |\beta_j^*| \succ \lambda_n \succ \sqrt{\log n/n} \tag{2.74}$$

In  $B_n$ , we have  $\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*\|_\infty \leq K \sqrt{\log n/n}$ . Therefore, for sufficiently large  $n$ , we have

$$\begin{cases} \min_{1 \leq j \leq p_0} |\tilde{\beta}_j| > a \lambda_n \\ \max_{p_0+1 \leq j \leq p} |\tilde{\beta}_j| = o(\lambda_n) \end{cases} \tag{2.75}$$

which leads to

$$\begin{cases} P'_\lambda(|\tilde{\beta}_j|) = 0, & j = 1, \dots, p_0 \\ P'_\lambda(|\tilde{\beta}_j|) > \alpha\lambda, & j = p_0 + 1, \dots, p \end{cases} \quad (2.76)$$

Thus, we have

$$\begin{aligned} S_1 &= \frac{1}{2}(1 - \delta) \mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1 - \frac{1}{\sqrt{n}} \mathbf{u}_1^T \mathbf{X}_1^T \boldsymbol{\varepsilon}_n + n \sum_{j=1}^{p_0} P'_\lambda(|\tilde{\beta}_j|) \left[ (|\beta_j^* + \frac{u_j}{\sqrt{n}}|) - (|\beta_j^*|) \right] \\ &= \frac{1}{2}(1 - \delta) \mathbf{u}_1^T \mathbf{C}_{11,n} \mathbf{u}_1 - \frac{1}{\sqrt{n}} \mathbf{u}_1^T \mathbf{X}_1^T \boldsymbol{\varepsilon}_n \\ &\geq \|\mathbf{u}_1\|_2 \left[ \frac{1}{2}(1 - \delta) \rho_{\min}(\mathbf{C}_{11,n}) \|\mathbf{u}_1\|_2 - K \sqrt{p_0 \log n} \right] \\ &> 0 \quad (\text{because } \|\mathbf{u}_1\|_2 > M_n = 2(1 - \delta)^{-1} \rho_{\min}^{-1}(\mathbf{C}_{11,n}) K \sqrt{p_0 \log n}) \end{aligned} \quad (2.77)$$

For  $S_2$ , we have

$$\begin{aligned} S_2 &= \frac{1}{2}(1 - \delta) \mathbf{u}_2^T \mathbf{C}_{22,n} \mathbf{u}_2 - \frac{1}{\sqrt{n}} \mathbf{u}_2^T \mathbf{X}_2^T \boldsymbol{\varepsilon}_n + n \sum_{j=p_0+1}^p P'_\lambda(|\tilde{\beta}_j|) \left( \frac{|u_j|}{\sqrt{n}} \right) \\ &> - \sum_{j=p_0+1}^p |u_j| \cdot \left\| \frac{1}{\sqrt{n}} \mathbf{X}_2^T \boldsymbol{\varepsilon}_n \right\|_\infty + \alpha \lambda_n \sqrt{n} \sum_{j=p_0+1}^p |u_j| \\ &= \sum_{j=p_0+1}^p |u_j| \left[ \alpha \lambda_n \sqrt{n} - K \sqrt{\log n} \right] \\ &> 0 \quad (\text{for sufficiently large } n) \end{aligned} \quad (2.78)$$

Combining (2.73) (2.77)(2.78), for sufficiently large  $n$ , on set  $B_n$ , we have

$$V_n(\mathbf{u}) > 0 \quad (2.79)$$

Note that  $V_n(\mathbf{0}) = 0$ , thus, by (2.79), we have shown that for any  $\mathbf{u} \in D_{1,n}$ ,

$$P(V_n(\mathbf{u}) > V_n(\mathbf{0})) \geq P(B_n) \rightarrow 1 \quad (2.80)$$

That is, with probability tending to 1,  $\mathbf{u} \in D_{1,n}$  cannot be the global minimizer of  $V_n(\mathbf{u})$ .

(2) for any  $\mathbf{u} \in D_{2,n}$ , on set  $B_n$ , consider the difference  $V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0})$ :

$$\begin{aligned} & V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}) \\ &= \frac{1}{2} \mathbf{u}_2^T \mathbf{C}_{22,n} \mathbf{u}_2 + \mathbf{u}_1^T \mathbf{C}_{12,n} \mathbf{u}_2 - \frac{1}{\sqrt{n}} \mathbf{u}_2^T \mathbf{X}_2^T \boldsymbol{\varepsilon}_n + n \sum_{j=p_0+1}^p P'_{\lambda_n}(|\tilde{\beta}_j|) \frac{|u_j|}{\sqrt{n}} \\ &\geq \mathbf{u}_1^T \mathbf{C}_{12,n} \mathbf{u}_2 - \frac{1}{\sqrt{n}} \mathbf{u}_2^T \mathbf{X}_2^T \boldsymbol{\varepsilon}_n + n \sum_{j=p_0+1}^p P'_{\lambda_n}(|\tilde{\beta}_j|) \frac{|u_j|}{\sqrt{n}} \\ &\geq - \sum_{j=p_0+1}^p |\mathbf{u}_1^T \mathbf{C}_{12,n} \mathbf{e}_j| |u_j| - \sum_{j=p_0+1}^p \left\| \frac{1}{\sqrt{n}} \mathbf{X}_2^T \boldsymbol{\varepsilon}_n \right\|_{\infty} |u_j| + \sum_{j=p_0+1}^p \sqrt{n} \cdot \alpha \lambda_n |u_j| \\ &\geq \sum_{j=p_0+1}^p |u_j| \left[ \sqrt{n} \cdot \alpha \lambda_n - \left\| \frac{\mathbf{x}_1 \mathbf{u}_1}{\sqrt{n}} \right\|_2 \left\| \frac{\mathbf{x}_2 \mathbf{e}_j}{\sqrt{n}} \right\|_2 - K \sqrt{\log n} \right] \\ &\geq \sum_{j=p_0+1}^p |u_j| \left[ \sqrt{n} \cdot c \lambda_n - \rho_{\max}^{1/2}(\mathbf{C}_{11,n}) M_n \cdot \left\| \frac{\mathbf{x}_j}{\sqrt{n}} \right\|_2 - K \sqrt{\log n} \right] \\ &\geq \sum_{j=p_0+1}^p |u_j| \left[ \sqrt{n} \cdot c \lambda_n - \rho_{\max}^{1/2}(\mathbf{C}_{11,n}) M_n - K \sqrt{\log n} \right] \\ &> 0 \end{aligned} \quad (2.81)$$

since  $\lambda_n \succ 2(1-\delta)^{-1} \rho_{\max}^{1/2}(\mathbf{C}_{11,n}) \rho_{\min}^{-1}(\mathbf{C}_{11,n}) K \sqrt{p_0 \log n/n}$ .

By (1) and (2), we have shown that the global minimizer of  $V_n(\mathbf{u})$  cannot be in

$\{(\mathbf{u}_1, \mathbf{u}_2) : \mathbf{u}_2 = \mathbf{0}\}^c$ . Therefore, we conclude that

$$\operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^p} V_n(\mathbf{u}) \in \{\mathbf{u} \in \mathbb{R}^p : \mathbf{u}_2 = \mathbf{0}\}, \quad (2.82)$$

Note that  $\operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^p} V_n(\mathbf{u}) = \sqrt{n} \left( \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right)$ , therefore, (2.82) leads to the conclusion

$$P(\hat{\boldsymbol{\beta}}_2 = \mathbf{0}) \rightarrow 1, \quad \text{as } n \rightarrow \infty$$

□

# Chapter 3

## BIC's Failure in High Dimension

### 3.1 Introduction

Choosing a suitable model is central to all statistical work with data. Selecting the variables for use in a regression model is one important example. These tasks have become more and more important with the emerge of large-volume and high-dimensional data.

Most of model selection criteria are based on likelihood functions or mean squared errors, such as the well-known AIC (Akaike, 1973), BIC Schwarz (1978), and cross validation. In addition to them, FIC (Claeskens et al., 2003) is a criterion which aims to select different models based on different parameters of interest. Model averaging (Hoeting et al., 2000) is another method which makes estimation based on weighted average from a group of models. These methods provide good insights into the problem. It is an important issue to determine what method to use under a particular situation, and this choice is usually made by the properties of the criteria, such as consistency and efficiency.

The most famous information criterion is AIC (Akaike's information criterion), which

is defined as

$$\text{AIC}(M) = 2l_n(\hat{\theta}) - 2\text{length}(\theta), \quad (3.1)$$

where  $\hat{\theta}$  denotes the maximum likelihood estimator,  $\text{length}(\theta)$  denotes the number of estimated parameters, and  $l_n(\hat{\theta})$  denotes the log-likelihood. The AIC method has intuitive appeal in penalizing the loglikelihood maximum for complexity, but it is not clear why the penalty factor should take the particular form  $-2\text{length}(\theta)$ . The Theory behind is to estimate the expected value of Kullback-Leibler (KL) distance from the unknown true density  $g(\cdot)$  to the parametric model  $f(\cdot, \theta)$ :

$$\text{KL}(g, f(\cdot, \theta)) = \int g(y) \log \frac{g(y)}{f(y, \theta)} \mathbf{d}y, \quad (3.2)$$

With MLE  $\hat{\theta}$ , we study the actually attained KL distance

$$\text{KL}(g, f(\cdot, \hat{\theta})) = \int g \log g \mathbf{d}y - \int g(y) \log f(y, \hat{\theta}) \mathbf{d}y, \quad (3.3)$$

Assuming that the approximating model is correct, i.e.  $g(y) = f(y, \theta_0)$ , then (3.3) leads to the definition of AIC (3.1).

However, such assumption ( $g(y) = f(y, \theta_0)$ ) is not true in general, thus there have been many modified versions of AIC to adjust for various situations. These modified versions have better approximation on small terms of (3.3), and consequently have more complicated format.  $\text{AIC}_C$  (Burnham and Anderson, 2002) is a famous corrected version of AIC, which is constructed as:

$$\text{AIC}_C = 2l_n(\hat{\theta}) - 2\text{length}(\theta) \frac{n}{n - \text{length}(n) - 1}. \quad (3.4)$$

Note that  $AIC_C$  only works in linear regression and autoregressive model, and there is no proof that it's applicable for general likelihood models.

Another approach is to pick the model with highest posterior probability, which is known as BIC:

$$BIC(M) = 2l_n(\hat{\theta}) - \log(n) \text{length}(\theta). \quad (3.5)$$

Usually, three properties are used to evaluate a criterion: consistency, efficiency, and parsimony. Under the assumption that the true model is one of the candidate models, weak consistency means the selection method is able to select the true model with probability tending to one; while strong consistency means that it picks the true model almost surely. Efficiency means that the method selects the model such that the ratio of the expected loss function at the selected model and loss at its theoretical minimizer converges to one in probability. Parsimony means that the selected model has the fewest parameters if there are more than one model that minimize the KL distance.

There are some crucial differences between AIC and BIC: (Claeskens and Hjort, 2008)

(i) The BIC behaves very well from the consistency point of view; with large  $n$  it gives a precise indication of which model is correct. (ii) But it pays a price for its null model consistency property: the risk function for the estimator exhibits unpleasant behaviour near the null model, and its maximal risk is unbounded with increasing sample size. (iii) AIC,  $AIC_C$  and Mallows's  $C_p$  (Mallows, 1973; Gilmour, 1996) are all asymptotically efficient; while BIC is not. In summary, AIC is efficient and BIC is consistent, but their strength can not be shared. In particular, any consistent model selection model cannot be minimax rate optimal. (Yang, 2005)

In the context of penalized regression, the aforementioned methods are applied to find the proper choice of  $\lambda$ . As we have discussed, the model selection consistency is usually

achieved with a proper choice of  $\lambda$ , which usually depends on  $p, n$  and the design matrix. In practice, the performance of penalization depends heavily on the selection of  $\lambda$ . For the linear model, typically, people chose  $\lambda$  which minimizes a certain type of information criterion (IC)

$$\text{IC}(\lambda) = \log \hat{\sigma}_\lambda^2 + |A_\lambda| \frac{C_n}{n} \quad (3.6)$$

where  $A_\lambda = \{j : 1 \leq j \leq p, \hat{\beta}_j(\lambda) \neq 0\}$  is index set of nonzero coefficients identified by  $\lambda$ ,  $\hat{\beta}(\lambda)$  is the estimated coefficients, and  $\hat{\sigma}_\lambda^2 = n^{-1} \|\mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda)\|^2$ .  $C_n$  is a number which determines the type of IC. For example, (3.6) is AIC when  $C_n = 2$ , or BIC when  $C_n = \log n$ .

Under the setting of fixed dimensions, Wang et al. (2007) showed that GCV/AIC has an inevitable overfitting effect in model selection, while BIC achieves the selection consistency. This result coincides with the previous discussion on AIC and BIC.

However, under the setting of high dimensions, things are quite different. The number of candidate models is no longer finite - it explodes with  $n$ . Therefore, it seems necessary that the penalty part  $|A_\lambda|C_n/n$  should be larger in order to conquer the exploration of possible subsets. Many works have been done on the modified BIC. Usually people increase  $C_n$  to a higher order (e.g.  $C_n/\log n \rightarrow \infty$ ,  $C_n/\log p \rightarrow \infty$ ), in order to adjust for the increased number of possible subsets. With such  $C_n$ , they are able to prove that the model selection consistency is achieved. (Wang et al., 2013; Wang and Zhu, 2011; Fan and Tang, 2013)

However, their results only imply  $C_n/\log n \rightarrow \infty$  is sufficient for consistency, while it remains unknown whether such  $C_n$  is necessary. Here we provide a general counter example which shows BIC fails to select the true model consistently in high dimensions. In fact, BIC tends to select an overfitted model in high dimension. That is, there is a

positive probability that some overfitted model is preferred to the true model using BIC.

## 3.2 A Counter Example

In this section, we show that BIC has an overfitting effect in high-dimension setting, just like GCV/AIC does in fixed-dimension setting. Therefore, it is necessary to use  $C_n/\log n \rightarrow \infty$  for high dimensional problems. We construct a counter example, and shall prove BIC's failure for that example. The reason why we did not generally prove BIC's failure will be discussed later.

We will use the following notations in this section: Let the index set of true model be  $A_0 = \{j : 1 \leq j \leq p, \beta_j^* \neq 0\}$ , and the index set of the selected model by  $\lambda$  be  $A_\lambda = \{j : 1 \leq j \leq p, \hat{\beta}_j(\lambda) \neq 0\}$ . We define two sets of models as following:

$$\Lambda_{n-} = \{A : A_0 \not\subset A\},$$

$$\Lambda_{n+} = \{A : A_0 \subset A, A_0 \neq A\}.$$

By the definition, we see that  $\Lambda_{n-}, \Lambda_{n+}$  are the sets of underfitted and overfitted models, respectively.

**Counter Example** Consider the linear regression problem (1.1), let  $p_0 < cn$  for some  $0 < c < 1$  and let  $\varepsilon_i$ 's be i.i.d random variables with Gaussian distribution  $N(0, \sigma^2)$ . Suppose there exists a subset of spurious variables, denoted as

$$\mathbb{S}_n = \{\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_{s_n} : \mathbf{Z}_i \text{ is a column of } \mathbf{X}_2, 1 \leq i \leq s_n\}$$

$\mathbb{S}_n$  satisfies

$$\begin{cases} |\mathbb{S}_n| > k \cdot n & \text{for some } k > 0, \\ \mathbf{X}_1^T \mathbf{Z}_i = \mathbf{0}, & \forall 1 \leq i \leq s_n, \\ \mathbf{Z}_i^T \mathbf{Z}_j = 0, & \forall 1 \leq i, j \leq s_n \text{ and } i \neq j \end{cases} \quad (3.7)$$

There exists a constant  $\alpha > 0$ , such that

$$P \left\{ \inf_{A \in \Lambda_{n+}} BIC(A) - BIC(A_0) < 0 \right\} \geq \alpha. \quad (3.8)$$

This counter example shows that there is a positive probability  $\alpha$  that the true model cannot be selected by BIC, since some overfitted model is preferred.

The requirement on dimensionality  $p$  is not strong:  $p$  does not have to be on the exponential or polynomial order of  $n$ . As long as  $p > (c+k)n$  for some  $k > 0$ , the counter example can be established. That means BIC fails as soon as  $p/n > k_0$  for some  $k_0 > 0$ .

This counter example is for the general purpose of subset selection. As for the selection of penalized regression, similar results can be established as well. Firstly, one needs to show that the penalized regression does select the true model with a proper  $\lambda_n$ ; secondly, one can show that the proper  $\lambda_n$  is not preferred by BIC. The second step is quite similar to the proof in section 3.3.

To prove BIC fails in high dimensions is much harder than to prove AIC fails in fixed dimensions. Recall that for AIC, one can pick a particular model, say the full model  $A_F = \{i : i = 1, 2, \dots, p\}$ , and show that  $P \{AIC(A_F) - AIC(A_0) < 0\} \geq \alpha$ . However, for BIC in high dimensions, for any particular model  $A$ , we actually can show that  $P \{BIC(A) - BIC(A_0) < 0\} \rightarrow 0$ . That is why we have to use  $\inf_{A \in \Lambda_{n+}} BIC(A)$  instead of  $BIC(A)$  in (3.8).

Once we use the inferior, we need to consider the distribution of the maximum of

correlated random variables, which is hard to quantify and relies on additional regularity conditions. For better understanding, we set up the counter example so that those random variables are independent, thus their maximum can be easily obtained. This is the intuition of the setting, and the reason why we did not strictly prove BIC's failure as well.

### 3.3 Proofs

*Proof of Counter Example.* Define  $A_i = A_0 \cup \{\mathbf{Z}_i\}$ , i.e. an overfitted model with  $p_0 + 1$  variables:  $A_0$  and one additional variable from  $\mathbb{S}_n$ . And define  $\mathbb{S}_A = \{A_1, A_2, \dots, A_{s_n}\}$ , then obviously  $\mathbb{S}_A$  is a subset of all overfitted models, i.e.  $\mathbb{S}_A \subset \Lambda_{n+}$ . Hence, by the definition of BIC,

$$\begin{aligned}
& P \left\{ \inf_{A \in \Lambda_{n+}} \text{BIC}(A) - \text{BIC}(A_0) < 0 \right\} \\
&= P \left\{ \inf_{A \in \Lambda_{n+}} \left[ -\log \left( \frac{\hat{\sigma}_{A_0}^2}{\hat{\sigma}_A^2} \right) + (|A| - p_0) \frac{\log n}{n} \right] < 0 \right\} \\
&\geq P \left\{ \inf_{A \in \mathbb{S}_A} \left[ -\log \left( \frac{\hat{\sigma}_{A_0}^2}{\hat{\sigma}_A^2} \right) + (|A| - p_0) \frac{\log n}{n} \right] < 0 \right\} \\
&= P \left\{ \frac{\log n}{n} - \sup_{A \in \mathbb{S}_A} \log \left( \frac{\hat{\sigma}_{A_0}^2}{\hat{\sigma}_A^2} \right) < 0 \right\} \tag{3.9}
\end{aligned}$$

For any overfitted model  $A$ , i.e.  $A_0 \subset A$ , we have  $n(\hat{\sigma}_{A_0}^2 - \hat{\sigma}_A^2) = \boldsymbol{\varepsilon}^T(\mathbf{P}_A - \mathbf{P}_{A_0})\boldsymbol{\varepsilon}$ .

Define  $x_A = \frac{\boldsymbol{\varepsilon}^T(\mathbf{P}_A - \mathbf{P}_{A_0})\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{A_0})\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T(\mathbf{P}_A - \mathbf{P}_{A_0})\boldsymbol{\varepsilon}}$ , we write  $\frac{\hat{\sigma}_{A_0}^2}{\hat{\sigma}_A^2}$  as

$$\frac{\hat{\sigma}_{A_0}^2}{\hat{\sigma}_A^2} = 1 + \frac{\boldsymbol{\varepsilon}^T(\mathbf{P}_A - \mathbf{P}_{A_0})\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_A)\boldsymbol{\varepsilon}} = 1 + \frac{\boldsymbol{\varepsilon}^T(\mathbf{P}_A - \mathbf{P}_{A_0})\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T(\mathbf{I}_n - \mathbf{P}_{A_0})\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T(\mathbf{P}_A - \mathbf{P}_{A_0})\boldsymbol{\varepsilon}} = 1 + x_A \tag{3.10}$$

It follows from (3.9) that

$$\begin{aligned}
& P \left\{ \inf_{A \in \Lambda_{n+}} \text{BIC}(A) - \text{BIC}(A_0) < 0 \right\} \\
& \geq P \left\{ \frac{\log n}{n} - \sup_{A \in \mathbb{S}_A} \log \left( \frac{\hat{\sigma}_{A_0}^2}{\hat{\sigma}_A^2} \right) < 0, \sup_{A \in \mathbb{S}_A} x_A < \frac{1}{2} \right\} \\
& + P \left\{ \frac{\log n}{n} - \sup_{A \in \mathbb{S}_A} \log \left( \frac{\hat{\sigma}_{A_0}^2}{\hat{\sigma}_A^2} \right) < 0, \sup_{A \in \mathbb{S}_A} x_A \geq \frac{1}{2} \right\} \tag{3.11}
\end{aligned}$$

We are going to show that  $P(\sup_{A \in \mathbb{S}_A} x_A \geq \frac{1}{2}) \rightarrow 0$ , so that only the first term matters. We shall use Theorem 1 in Daniel Hsu and Zhang (2012): for the subgaussian errors  $\boldsymbol{\varepsilon}$  with mean  $\mathbf{0}$  and variance  $\sigma^2 \mathbf{I}_n$ , the quadratic form satisfies ( $\|\cdot\|$  denotes the spectral norm)

$$P \left\{ \boldsymbol{\varepsilon}^T \Sigma \boldsymbol{\varepsilon} > \sigma^2 \left( \text{tr}(\Sigma) + 2\sqrt{\text{tr}(\Sigma^2)t} + 2\|\Sigma\|t \right) \right\} \leq e^{-t} \tag{3.12}$$

Furthermore, define  $K = \text{tr}(\Sigma)$ , if  $\Sigma$  is symmetric and idempotent, we have for all  $t > K$ ,

$$P \left\{ \boldsymbol{\varepsilon}^T \Sigma \boldsymbol{\varepsilon} > 3\sigma^2 K^{1/2}t \right\} < P \left\{ \boldsymbol{\varepsilon}^T \Sigma \boldsymbol{\varepsilon} > \sigma^2 \left( \text{tr}(\Sigma) + 2\sqrt{\text{tr}(\Sigma^2)t} + 2\|\Sigma\|t \right) \right\} \leq e^{-t} \tag{3.13}$$

For an overfitted model  $A$  with  $|A| = p_0 + 1$ , considering  $\Sigma = (\mathbf{P}_A - \mathbf{P}_{A_0})$ , we see that

$$K = \text{tr}(\Sigma) = \text{tr}(\mathbf{P}_A) - \text{tr}(\mathbf{P}_{A_0}) = \text{tr}((\mathbf{X}_A^T \mathbf{X}_A)^{-1} \mathbf{X}_A^T \mathbf{X}_A) - \text{tr}((\mathbf{X}_{A_0}^T \mathbf{X}_{A_0})^{-1} \mathbf{X}_{A_0}^T \mathbf{X}_{A_0}) = |A| - |A_0| = 1,$$

applying (3.13) to  $\Sigma = (\mathbf{P}_A - \mathbf{P}_{A_0})$ , let  $t = kn$ , we have

$$P \left\{ \boldsymbol{\varepsilon}^T (\mathbf{P}_A - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon} > 3\sigma^2 kn \right\} \leq e^{-kn},$$

it follows that (let  $t = kn$ )

$$\begin{aligned}
& P \left\{ \sup_{A \in \mathbb{S}_A} \boldsymbol{\varepsilon}^T (\mathbf{P}_A - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon} > 3\sigma^2 kn \right\} \\
& \leq |\mathbb{S}_n| \cdot P \left\{ \boldsymbol{\varepsilon}^T (\mathbf{P}_A - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon} > 3\sigma^2 kn \right\} \\
& < kn \cdot e^{-kn} \rightarrow 0
\end{aligned} \tag{3.14}$$

Therefore, we have

$$\begin{cases} \sup_{A \in \mathbb{S}_A} \boldsymbol{\varepsilon}^T (\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon} = o(n) \\ \boldsymbol{\varepsilon}^T (\mathbf{I}_n - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon} - \sup_{A \in \mathbb{S}_A} \boldsymbol{\varepsilon}^T (\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon} \rightarrow (n - p_0) \sigma^2 \cdot (1 + o_p(1)) \end{cases}$$

Thus

$$\sup_{A \in \mathbb{S}_A} x_A = \frac{\sup_{A \in \mathbb{S}_A} \boldsymbol{\varepsilon}^T (\mathbf{P}_A - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T (\mathbf{I}_n - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon} - \sup_{A \in \mathbb{S}_A} \boldsymbol{\varepsilon}^T (\mathbf{P}_A - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}} = o_p(1) \tag{3.15}$$

With (3.15), we can see that the second term in (3.11) goes to 0, as  $n \rightarrow \infty$ . Therefore, (3.11) mainly depends on the first term. Note that  $\frac{\hat{\sigma}_{A_0}^2}{\hat{\sigma}_A^2} = 1 + x_A$ , we apply  $\log(1 + x) > \frac{2}{3}x$ ,  $\forall x < \frac{1}{2}$  to the first term and get

$$\begin{aligned}
& P \left\{ \inf_{A \in \Lambda_{n+}} \text{BIC}(A) - \text{BIC}(A_0) < 0 \right\} \\
& \geq P \left\{ \frac{\log n}{n} - \sup_{A \in \mathbb{S}_A} \frac{2}{3} x_A < 0 \right\} + o_p(1) \\
& \geq P \left\{ \frac{\log n}{n} - \sup_{A \in \mathbb{S}_A} \frac{2}{3} x_A < 0 \right\} + o_p(1) \\
& \geq P \left\{ \frac{\log n}{n} - \frac{2}{3} \cdot \frac{\sup_{A \in \mathbb{S}_A} \boldsymbol{\varepsilon}^T (\mathbf{P}_A - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}}{(n - p_0) \sigma^2 - \sup_{A \in \mathbb{S}_A} \boldsymbol{\varepsilon}^T (\mathbf{P}_A - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}} < 0 \right\} + o_p(1) \\
& \geq P \left\{ \sup_{A \in \mathbb{S}_A} \boldsymbol{\varepsilon}^T (\mathbf{P}_A - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon} > \frac{3}{2} \sigma^2 \log n \right\} + o_p(1)
\end{aligned} \tag{3.16}$$

By the definition (3.7) of the counter example, consider a particular  $A = A_i \in \mathbb{S}_A$ , then

$$\boldsymbol{\varepsilon}^T(\mathbf{P}_{A_i} - \mathbf{P}_{A_0})\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^T \mathbf{Z}_i (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \boldsymbol{\varepsilon} = \frac{1}{n} (\mathbf{Z}_i^T \boldsymbol{\varepsilon})^2.$$

That means,  $T_i \stackrel{def}{=} \boldsymbol{\varepsilon}^T(\mathbf{P}_{A_i} - \mathbf{P}_{A_0})\boldsymbol{\varepsilon}$ ,  $i = 1, 2, \dots, s_n$  are *i.i.d* random variables with  $\chi_1^2$  distribution, which allows us to calculate the probability of (3.16)

$$\begin{aligned} & P \left\{ \inf_{A \in \Lambda_{n+}} \text{BIC}(A) - \text{BIC}(A_0) < 0 \right\} \\ & \geq P \left\{ \sup_{i=1,2,\dots,s_n} T_i > \frac{3}{2} \sigma^2 \log n \right\} + o_p(1) \\ & = 1 - \left( P \left\{ T_1 \leq \frac{3}{2} \sigma^2 \log n \right\} \right)^{s_n} \end{aligned} \quad (3.17)$$

As we know, the CDF of standard normal distribution  $\Phi(\cdot)$  satisfies

$$\Phi(t) = P(Z \leq t) \leq 1 - \frac{1}{\sqrt{2\pi}} \frac{t}{t^2 + 1} e^{-\frac{t^2}{2}}$$

Therefore, for sufficiently large  $n$ ,

$$P \left\{ T_1 \leq \frac{3}{2} \sigma^2 \log n \right\} \leq 1 - \frac{2}{\sqrt{2\pi}} \frac{\sqrt{\frac{3}{2} \log n}}{\frac{3}{2} \log n + 1} e^{-\frac{3}{4} \log n} < 1 - \frac{1}{kn} \quad (3.18)$$

Combine (3.17)(3.18) and  $s_n > kn$ , we have

$$P \left\{ \inf_{A \in \Lambda_{n+}} \text{BIC}(A) - \text{BIC}(A_0) < 0 \right\} > 1 - \left( 1 - \frac{1}{kn} \right)^{kn} \rightarrow 1 - e^{-1}. \quad (3.19)$$

That means the probability of selecting an overfitted model using BIC is at least  $1 - e^{-1}$ .  $\square$

# Chapter 4

## Partially Penalized Regression

### 4.1 Introduction

With the advance of data collecting techniques, increasing attention has been paid to the tools for analyzing large-volume high-dimensional data. One of the popular tools is the penalized regression, which is usually defined as:

$$Q_n(\boldsymbol{\beta}; \lambda) = \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n \sum_{j=1}^p P_\lambda(|\beta_j|) \quad (4.1)$$

where  $\lambda$  is a tuning parameter and  $P_\lambda(\cdot)$  is a penalty function. For instance,  $P_\lambda(|t|) = |t|$  for LASSO(Tibshirani, 1994);  $P_\lambda(|t|) = t^2/2$  for Ridge regression (Hoerl and Kennard, 1970). The argument that minimizes (4.1) is the estimator of coefficients, i.e.

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathcal{R}^p} Q_n(\boldsymbol{\beta}; \lambda) \quad (4.2)$$

There has been a rich literature in penalized regression, and there are lots of famous

examples of  $P_\lambda(\cdot)$ , such as LASSO (Tibshirani, 1994), SCAD (Fan and Li, 2001), MCP (Zhang, 2010), Adaptive LASSO (Zou, 2006), Truncated LASSO (Shen et al., 2012), one-step estimator (Zou and Li, 2008), Elastic Net Zou and Hastie (2005), Grouped LASSO Yuan and Lin (2006), Dantzig (Candes and Tao, 2007), etc. Although they follow different format, they enjoy nice properties under certain conditions. The Oracle Property is one the most desired properties for a penalized method, which refers to: (1)the true model is identified with probability tending to 1; (2) the estimation for nonzero coefficients is as efficient as if we know them in advance.

It is not easy to prove the Oracle Property in high dimensions. (Zhang, 2010; Kim et al., 2008; Huang and Xie, 2007; Zheng et al., 2014; (Chatterjee and Lahiri, 2013); Huang et al., 2008) Even with a successful theory, there is a commons issue, which is, almost all the theoretical results on model selection consistency reply on a common assumption: the magnitude of smallest coefficients must be large enough. However, this assumption is not necessarily true and is not easy to validate. In this paper, we will discuss the problems associated with small coefficients, and we aim to find a solution which does not depend on such assumption.

The rest of the article is organized as follows. In section 4.2, we discuss two motivations of our approach, which are the limitation of existing results and the merit of combining LASSO and nonconcave penalties. We introduce partially penalized regression in section 4.3, explaining what it is and why it is designed in that way. In section 4.4, we talk about the problems associated with small nonzero coefficient, and argue that it is very dangerous to ignore them. In section 4.5, we show that our suggested LIC criteria has the power to identify large coefficients. The numerical results are presented in section 4.6, including simulated data and two real dataset. Theoretical proofs are in section 4.7.

## 4.2 Motivation

Our motivation for the partially penalized regression comes from two aspects: one is from the limitation of a widely adopted assumption, which requires the smallest coefficient be greater than a threshold; the other is from the comparison of risk functions using different methods, which indicates the benefit of combining LASSO and non-concave penalties.

### 4.2.1 The Limitation of Previous Results

#### Magnitude Assumption

All the existing theoretical results rely on a similar assumption, which we refer as the Magnitude Assumption. It says:

*The smallest signal  $b^* = \min_{j:\beta_j \neq 0} |\beta_j|$  must be sufficiently large.*

The threshold is different from paper to paper, but in general it depends on  $n$ ,  $p$ ,  $p_0$ , and the design matrix. We list a few examples of the assumption as follows:

- **Fixed dimension** It requires  $b^*/n^{-1/2} \rightarrow \infty$ .
- **MCP (Zhang, 2010)** It requires  $b^* > c\sqrt{\log p/n}$ , where  $c$  depends on the design matrix, noises and  $a$ .
- **SCAD (Kim et al., 2008)** It requires  $b^* > cn^{-1/2+q/2}$ , where  $q$  satisfies  $n^q/p_0 \rightarrow \infty$  and  $c$  is a constant.
- **SCAD (Huang and Xie, 2007)** It requires  $b^* > \frac{\lambda_n \sqrt{p_0}}{\sqrt{\rho_{n,1}}}$ , where  $\rho_{n,1}$  is the smallest eigenvalue of the gram matrix  $n^{-1}\mathbf{X}^T\mathbf{X}$ .

- **Hard Thresholding (Zheng et al., 2014)** It requires  $b^* > c\sqrt{p_0 \log p/n}$ , where  $c$  depends on the eigenvalues of design matrix.

Many penalties have been designed in a certain way so that they enjoy the Oracle Property. The idea is that different coefficients receive different penalization based on its scale. Most of those penalty functions can be classified into two classes.

One class relies on an initial estimator, and the penalty on each coefficient depends on its own initial estimator. Famous examples include Adaptive LASSO (Zou, 2006) and one-step estimator Zou and Li (2008). The initial estimator can be OLS, Ridge or LASSO estimator, and some conditions guarantee that the initial estimator and the true parameter are close enough. This ensures small weights for large coefficients and large weights for small/zero coefficients.

The other class of functions does not rely on any kind of initial estimator, but they usually have a delicately designed format. In general, the function  $P_\lambda(t), t \geq 0$  is non-concave with a positive right-derivative at 0 and a constant part beyond a certain point  $a\lambda$ , see Chapter 2 for more details. Examples includes SCAD (Fan and Li, 2001), MCP (Zhang, 2010), Hard Thresholding (Zheng et al., 2014), Truncated LASSO (Shen et al., 2012), etc. Ideally, when  $n$  is sufficient large, zero and small coefficients are heavily penalized, and the large coefficients are not penalized at all.

Have studied these theoretical results, we found out the secret why all researchers need such a condition. That is, in order to achieve the selection consistency, people need two key relationships:

$$\begin{cases} \sup_i \boldsymbol{\alpha}_i^T \boldsymbol{\varepsilon} = o_p(\lambda_n), & \text{where } \boldsymbol{\alpha}_i \text{ s are vectors which depend on } \mathbf{X} \\ b^* > a\lambda_n. \end{cases} \quad (4.3)$$

The first defines the lower bound of  $\lambda_n$ , while the second defines the upper bound. Only when such  $\lambda_n$  exists, i.e. only when the upper bound is greater than the lower bound, the selection consistency is true. Therefore, people need  $\sup_i \boldsymbol{\alpha}_i^T \boldsymbol{\varepsilon} = o_p(b^*)$ . Depending their approach, different researchers came up with the aforementioned different quantities.

The reason that both conditions are necessary is as follows:

- All the nonzero coefficients need to locate in the constant part of the penalty function. If so, in the gradient of loss function  $Q_n(\cdot)$ , the terms related to nonzero coefficients will be exactly 0, and be dominated by the other terms. This leads to selection consistency, and furthermore, the Oracle Property. If not, the terms related to nonzero coefficients are either too large or depend on  $\lambda_n$ . If they are too large, they are large enough to affect the behavior of  $Q_n(\cdot)$ ; if they depend on  $\lambda_n$ , then some unrealistic inequality on  $\lambda_n$  needs to be true.
- $\lambda_n$  needs to be large enough to dominate the supremum of a set of linear combinations of random errors, so that no false positive occurs.

## What is Wrong

In practice, the Magnitude Assumption raises three issues.

First, the Magnitude Assumption is too ideal. The assumption indicates that some variables are quite useful while all the others are not useful at all, no variables are in between. This does not make sense for a real problem. A more reasonable assumption seems to be that the target variable is affected by many predictors, some of them are strong while some are weak. In other words, there are many nonzero coefficients, some are small, some are large. There is no clear gap between nonzero coefficients and zero.

In addition, the Magnitude Assumption is too hard to justify: usually, only one pair of

$(n, p)$  is observed, while the threshold depends on the growth rate of  $p$  as a function of  $n$ . One can not simply use the single pair to estimate the growth rate. In addition, we have no idea how large  $b^*$  is. We can approximate  $b^*$  using some estimators, but the estimators themselves are subject to the validity of the assumption and the model. Furthermore, the constant  $c$  depends on lots of factors (such as  $\mathbf{X}_1, p_0, \sigma$ , etc). It could be very large and it is hard to estimate.

Last but not least, the penalty parameter selection could be an issue. The performance of penalized regression depends heavily on the choice of  $\lambda_n$ , and in practice, this choice is made by minimizing a selection criterion. With the Magnitude Assumption, the selected  $\lambda_n$  aims to find the cut point between nonzero and zero coefficients; without the assumption, it actually is a compromise of large nonzero coefficients and small ones. The selection should be preciser if all the nonzero coefficients are on the same scale.

## Numerical Examples

To better understand the problem associated with the magnitude assumption, we conduct a small simulation with two methods - LASSO and MCP. We tweak the magnitude of  $\beta$  and the design matrix, so that the magnitude assumption is not always true in all scenarios. And we will see the performance of MCP changes dramatically.

We briefly explain the simulation setup, please refer to section 4.6 for more details. The true  $\beta$  has 10 nonzero coefficients on its first 10 entries, and they are randomly generated from uniform(1,2) distribution. We have used both  $\beta$  and  $\beta/2$ , so that the second setting violates the magnitude assumption more easily. We also tweak the correlation matrix of the design matrix from  $\Sigma = \mathbf{I}_n$  to  $\Sigma = \{\Sigma_{ij}\} = \rho^{|i-j|}$ ,  $\rho = 0.5, 0.7$ . Six metrics are reported, including the prediction MSE, probability of selecting true models, truth positive rate, false positive rate, selected model size, and  $l_2$  loss  $\|\hat{\beta} - \beta^*\|$ .

Table 4.1: *Simple demonstration of magnitude assumption - with the comparison of LASSO and MCP. ( $n = 100, p = 1000, p_0 = 10$ )*

Beta	$\Sigma$	Method	metrics					
			MSE	TM	TP	FP	Size	$\ \hat{\beta} - \beta^*\ $
$\beta$	Ind	MCP	1.1259	0.934	10.000	0.258	10.258	0.3921
		LASSO	2.8136	0.000	10.000	82.212	92.212	1.3058
	AR1(0.5)	MCP	1.1599	0.876	9.978	0.116	10.094	0.5113
		LASSO	1.6180	0.142	10.000	6.028	16.028	0.6464
	AR1(0.7)	MCP	2.6583	0.052	7.986	0.356	8.342	2.2892
		LASSO	1.4550	0.170	10.000	3.100	13.100	0.6796
$\beta/2$	Ind	MCP	2.2664	0.092	9.874	17.982	27.856	1.0767
		LASSO	2.6254	0.000	9.936	85.880	95.816	1.1677
	AR1(0.5)	MCP	2.1231	0.008	8.070	3.874	11.944	1.3642
		LASSO	1.8790	0.016	9.998	64.272	74.270	0.8929
	AR1(0.7)	MCP	2.4434	0.000	5.652	1.184	6.836	2.1937
		LASSO	1.4274	0.286	9.992	4.150	14.142	0.6396

In Table 4.1 we compare the performance of LASSO and MCP. As we can see, MCP outperforms LASSO in setting 1, 2 and 4, while LASSO outperforms MCP in setting 3, 5 and 6. By “outperform”, we mean smaller MSE, FP,  $l_2$  loss and larger TM, TP. Note that it is easier to violate the magnitude assumption when  $\beta$  is smaller and  $\rho$  is larger, therefore, the comparison indicates that LASSO is preferred to MCP when the magnitude assumption is violated.

It is recommended to use ROC or PR curve for comparison of methods, since they avoid the problem of tuning issues. However, in this case, the number of nonzero and zero coefficients are extremely unbalanced, which makes ROC/PR curves unsuitable for comparison.

Recall that, under certain conditions, MCP is claimed to enjoy Oracle Property and should have better selection accuracy and prediction accuracy than LASSO does, hence

in practice, people usually apply MCP for model selection without any examination on its pre-requirements. Here we show how dangerous it is. In fact, LASSO has a dominating advantage over MCP in section 3,5 and 6.

This table will also justify our arguments in section 4.2.2.

### 4.2.2 The Behavior of Risk Functions

Fan and Li (2001) studied the thresholding rule of LASSO, SCAD, and hard thresholding under a univariate scenario (or orthogonal design matrix), and derives the risk function  $R(\hat{\theta}, \theta) = E(|\hat{\theta} - \theta|^2)$ . They argued that the SCAD penalty outperforms the other two penalties by comparing their risk functions.

Following similar approach, we also derived the risk functions of a few methods, and will demonstrate why partially penalized regression is the right choice. In the comparison, we include five penalties but exclude the others because: (a)MCP, SCAD, HT, TL are the only ones which have been proved to enjoy selection consistency in high dimensions. (See Zhang, 2010, Kim et al., 2008, Zheng et al., 2014, Shen et al., 2012.) (b)LASSO has an important advantage over the others.

Let us assume that the design matrix is orthogonal and standardized, which means

$n^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{I}_n$ . We can write (4.1) as following:

$$\begin{aligned}
Q_n(\boldsymbol{\beta}; \lambda) &= \frac{1}{2}\|\mathbf{Y} - n^{-1}\mathbf{X}\mathbf{X}^T\mathbf{Y}\|^2 + \frac{1}{2}\|n^{-1}\mathbf{X}\mathbf{X}^T\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n\sum_{j=1}^p P_\lambda(|\beta_j|) \\
&= \frac{1}{2}\|\mathbf{Y} - n^{-1}\mathbf{X}\mathbf{X}^T\mathbf{Y}\|^2 + \frac{1}{2}\|n^{-1}\mathbf{X}\mathbf{X}^T\mathbf{Y}\|^2 + \frac{1}{2}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\
&\quad - n^{-1}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{Y} + n\sum_{j=1}^p P_\lambda(|\beta_j|) \\
&= C(\mathbf{X}, \mathbf{Y}) + \frac{n}{2}[\boldsymbol{\beta}^T\boldsymbol{\beta} - \boldsymbol{\beta}^T(n^{-1}\mathbf{X}^T\mathbf{Y})] + n\sum_{j=1}^p P_\lambda(|\beta_j|) \tag{4.4}
\end{aligned}$$

where  $C(\mathbf{X}, \mathbf{Y}) = \frac{1}{2}\|\mathbf{Y} - n^{-1}\mathbf{X}\mathbf{X}^T\mathbf{Y}\|^2 + \frac{1}{2}\|n^{-1}\mathbf{X}\mathbf{X}^T\mathbf{Y}\|^2$  only depends  $\mathbf{X}$  and  $\mathbf{Y}$ . Therefore, to minimize  $Q_n(\boldsymbol{\beta}; \lambda)$  is equivalent to minimize the remaining terms. Define  $z_j = n^{-1}\mathbf{X}_j^T\mathbf{Y}$ , we have

$$\operatorname{argmin}_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta}; \lambda) = \operatorname{argmin}_{\boldsymbol{\beta}} \sum_{j=1}^p \left[ \frac{1}{2}(\beta_j - z_j)^2 + P_\lambda(|\beta_j|) \right] \tag{4.5}$$

Equation (4.5) can be minimized componentwisely, and each component is the same:

$$q_n(\theta; \lambda) = \frac{1}{2}(\theta - z)^2 + P_\lambda(|\theta|) \tag{4.6}$$

where  $z = n^{-1}\mathbf{X}^T\mathbf{Y}$ . This optimization is not hard, in fact, for a given penalty function, one can easily get the closed form of  $\hat{\theta} = \operatorname{argmin}_{\theta} q_n(\theta; \lambda)$ . In table 4.2 we summarize the minimizers for a few famous penalty functions, which is also referred as the thresholding rule.

Table 4.2: Thresholding Rules (  $s(z, \lambda)$  is the thresholding rule of LASSO)

Penalty Name	Penalty Function ( $t \geq 0$ )	Thresholding Rule
LASSO	$P_\lambda(t) = \lambda t$	$\hat{\theta} = s(z, \lambda) = \begin{cases} z - \lambda, & \text{if } z > \lambda \\ 0, & \text{if }  z  \leq \lambda \\ z + \lambda, & \text{if } z < -\lambda \end{cases}$
MCP	$P_\lambda(t) = \begin{cases} \lambda t - \frac{t^2}{2a}, & \text{if } t \leq a\lambda \\ \frac{1}{2}a\lambda^2, & \text{if } t > a\lambda \end{cases}$	$\hat{\theta} = \begin{cases} \frac{s(z, \lambda)}{1-1/a}, & \text{if }  z  \leq a\lambda \\ z, & \text{if }  z  > a\lambda \end{cases}$
SCAD	$P_\lambda(t) = \begin{cases} \lambda t, & \text{if } t \leq \lambda \\ \frac{a\lambda t - (t+\lambda)^2/2}{a-1}, & \text{if } \lambda < t \leq a\lambda \\ \frac{(a+1)\lambda^2}{2}, & \text{if } t > a\lambda \end{cases}$	$\hat{\theta} = \begin{cases} s(z, \lambda), & \text{if }  z  \leq 2\lambda \\ \frac{s(z, a\lambda)/(a-1)}{1-1/(a-1)}, & \text{if } 2\lambda <  z  \leq a\lambda \\ z, & \text{if }  z  > a\lambda \end{cases}$
Truncated LASSO	$P_\lambda(t) = \lambda t I(t \leq a\lambda)$	$\hat{\theta} = \begin{cases} s(z, \lambda), & \text{if }  z  \leq a\lambda \\ z, & \text{if }  z  > a\lambda \end{cases}$
Hard Thresholding	$P_\lambda(t) = \frac{1}{2} [\lambda^2 - (\lambda - t)_+^2]$	$\hat{\theta} = z I( z  > \lambda)$

Using the thresholding rule in table 4.2, and assuming  $\mathbf{z} \sim N(\theta, 1)$ , we are able to derive the risk function  $R(\hat{\theta}, \theta) = E(|\hat{\theta} - \theta|^2)$ . A tricky part is that  $R(\hat{\theta}, \theta)$  depends heavily on the choice of  $\lambda$ . To make the scale of five risk functions roughly comparable, we fix  $\lambda = 2$  for the hard thresholding and adjusted the  $\lambda$  value for the other penalties.

In Figure 4.1, we compare the risk functions of the aforementioned five penalties. For the upper plot,  $\lambda$  is adjusted so that the risks are equal at  $\theta = 3$ ; while for the lower plot,  $\lambda$  is adjusted so that risks are equal at  $\theta = 0$ . In addition to the demonstrated two plots, we have also tried a few other sets of  $\lambda$ . Their shapes and comparison relationship are similar, therefore we only report these two representatives.

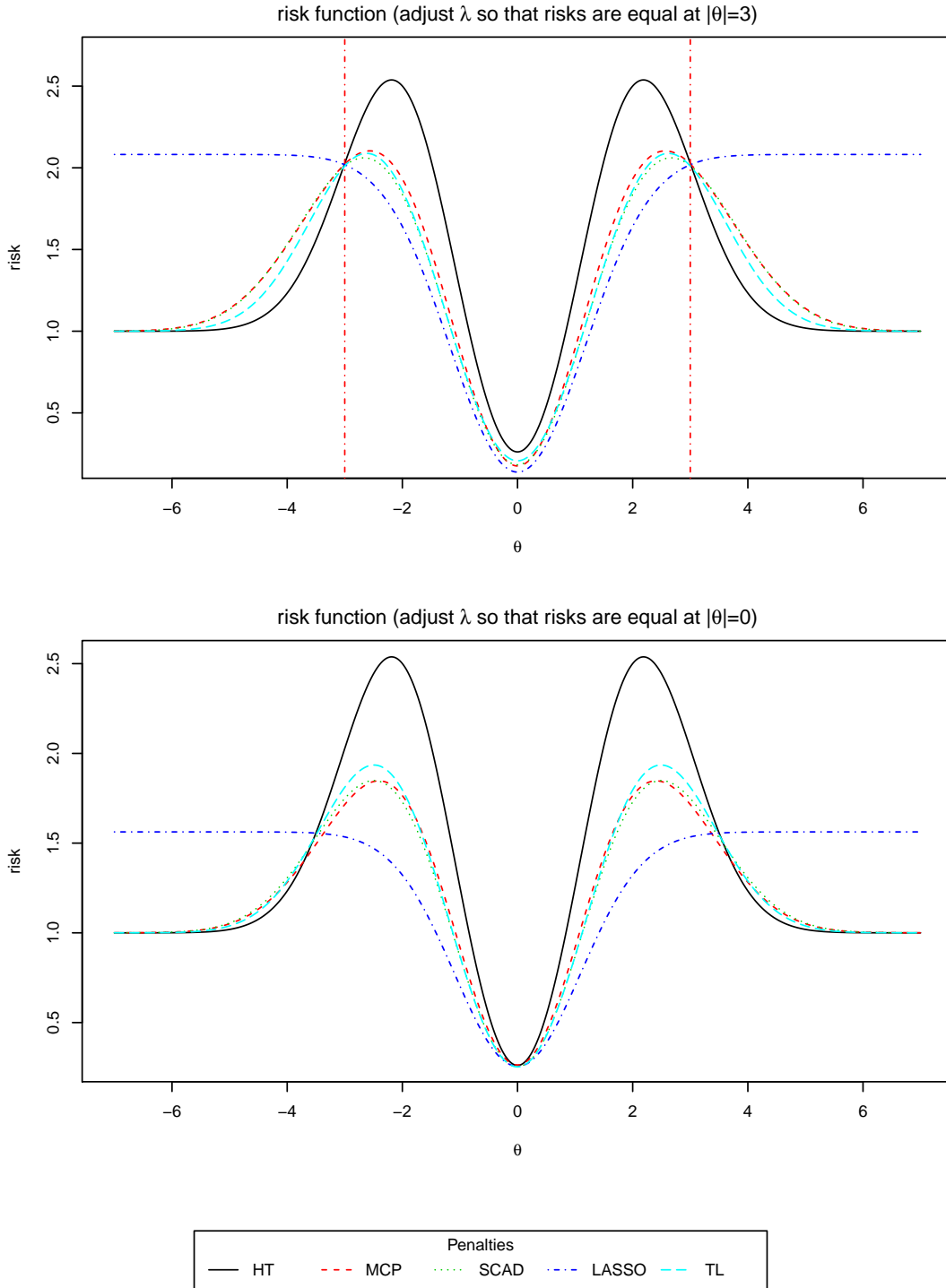


Figure 4.1: The comparison of risk function for the five penalty function list in table 4.2.

As we can see from Figure 4.1,

1. When  $|\theta|$  is large, LASSO has a much greater risk than all others; while all others have the same risk when  $\theta$  is beyond certain points. This is expected since the LASSO estimator has a large bias (which is equal to  $\lambda$  in this case), while all the other estimators have no bias for large  $\theta$ .
2. When  $|\theta|$  is small, the risk of LASSO is the smallest.
3. No matter what  $\lambda$  we choose, none of the penalties has a uniformly smaller risk than the others.

When there are  $p$  coefficients, some large, some small, and some zeroes, the choice of  $\lambda$  must be a compromise between risks on large coefficients and risks on smaller ones. We should avoid this compromise by controlling the coefficients to a similar scale. Since the majority of coefficients are zero, we want to focus on the small or zero coefficients, and conduct the tuning parameter selection without the “noises” from large coefficients.

Based on the comparison of risk functions, here comes our second motivation for the partially penalized regression:

We should take the advantages of both LASSO and nonconcave penalties. LASSO is good for small coefficients because of smaller risk, while nonconcave penalties are good for large coefficients because of no bias. What if we can combine the smaller risks of LASSO for smaller coefficients and the smaller risks of nonconcave penalties for larger coefficients? By doing so, we are able to create an approach which has an uniformly smaller risk than all these popular approaches.

Besides, the selection of regularization parameter makes more sense using PPR. When assuming the Magnitude Assumption, the purpose of tuning  $\lambda_n$  is to find the separation

so that no penalization is imposed on nonzero coefficient. When that assumption is violated, the selection consistency is no longer achievable, then the purpose of tuning should focus more on small coefficients.

Following the discussion above, we have come up with the partially penalized regression. Its detail is provided in the next section.

### 4.3 Partially Penalized Regression

Now let us ignore the Magnitude Assumption, i.e. we allow that some nonzero coefficients be smaller than the threshold. We propose the partially penalized regression, and we believe it to be a cure for this situation.

Due to the existence of small nonzero coefficients, there is no theoretical guarantee that the existing methods are capable of separating zero and nonzero coefficients. In fact, the randomness of errors are large enough to blind the identification of small signals. Therefore, the goal has to be adjusted, from picking the right model to finding a good model with accurate estimations. Hence our goal is to separate between important variables and less-important variables. (we shall show what “important” means later.) After that, we can focus on the less-important variables, and make a compromise of selection consistency and estimation accuracy.

Here we propose the following partially penalized regression. It has two steps:

1. Denote  $A_0$  as the index set of all nonzero coefficients. In the first step, the procedure selects a subset  $L \subset A_0$ , such that  $L$  contains all the large coefficients and no zero coefficients. Some small coefficients may also belong to  $L$ , which is allowed, and this indicates that they are too important to be ignored. In other words, PPR finds an index set  $L$  which has all the important coefficients.

2. Let the  $S = \{j : j \notin L, 1 \leq j \leq p\}$  be the index set of remaining coefficients. Define  $\mathbf{X}_L$  as the sub-matrix of  $\mathbf{X}$  which corresponds to  $L$ , and the remaining matrix as  $\mathbf{X}_S$ . In the second step, we only penalize on the less-important coefficients  $\{\beta_j : j \in S\}$ . That is, we minimize

$$Q_n(\boldsymbol{\beta}_L, \boldsymbol{\beta}_S; \lambda_n) = \frac{1}{2} \|\mathbf{Y} - \mathbf{X}_L \boldsymbol{\beta}_L - \mathbf{X}_S \boldsymbol{\beta}_S\|^2 + n \sum_{j \in S} P_{\lambda_n}(\beta_j) \quad (4.7)$$

Let  $\hat{\boldsymbol{\beta}}^T = (\hat{\boldsymbol{\beta}}_L^T, \hat{\boldsymbol{\beta}}_S^T) = \operatorname{argmin} Q_n(\boldsymbol{\beta}_L, \boldsymbol{\beta}_S; \lambda_n)$ , then  $\hat{\boldsymbol{\beta}}$  is the final estimator of the coefficients.

We are going to discuss our solution to either step in the following subsections.

### 4.3.1 First Step

The first step is to find the desired index set  $L$ . Our solution is MCP penalized regression combined with a modified version of information criterion (IC).

As we know, for linear regression, IC has a general format:

$$\text{IC}(\lambda) = \log \hat{\sigma}_\lambda^2 + |A_\lambda| \frac{C_n}{n} \quad (4.8)$$

where  $A_\lambda = \{j : 1 \leq j \leq p, \hat{\beta}_j(\lambda) \neq 0\}$  is index set of nonzero coefficients identified by  $\lambda$ ,  $\hat{\boldsymbol{\beta}}(\lambda)$  is the estimated value of coefficients under tuning parameter  $\lambda$ , and  $\hat{\sigma}_\lambda^2 = n^{-1} \|\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}(\lambda)\|^2$ .  $C_n$  is a number which determines the type of IC. For example, AIC is the case when  $C_n = 2$  and BIC is when  $C_n = \log n$ . The selected  $\lambda$  is  $\lambda = \operatorname{argmax}_{\lambda \geq 0} \text{IC}(\lambda)$ .

In the setting of high dimensions,  $C_n$  is suggested to take a larger value, for example,  $C_n / \log p \rightarrow \infty$  or  $C_n / \log n \rightarrow \infty$ . There has been theoretical proof which shows such IC combined with SCAD/MCP consistently selects the true model. (Fan and Tang 2013,

Wang and Zhu 2011, Wang et al. 2013) However, those results are also based on some type of Magnitude Assumption.

The idea of IC is to penalize on the model complexity, each nonzero coefficient in the selected model contributes equal weight to  $A_\lambda$ . However, in addition to distinguishing between zero and nonzero coefficients, it is also important to distinguish between large and small nonzero coefficients. Our suggestion is to assign different weights to coefficients according to their estimated values, and we set the boundary between large and small to be  $a\lambda$ , which is the point beyond which the penalty function becomes constant. We do have a reason to set the boundary at  $a\lambda$ , and we shall discuss it in section 4.4. Hence, we suggest a modified version of (4.8) which aims to select large nonzero coefficients. It is defined as follows:

$$\text{LIC}(\lambda; a) = \log \hat{\sigma}_\lambda^2 + \frac{C_n}{n} \sum_{j=1}^p L(\hat{\beta}_j(\lambda); a, \lambda) \quad (4.9)$$

where  $L(\cdot)$  is a function that aims to penalize more on small nonzero coefficients.

$$L(t; a, \lambda) = \begin{cases} 0, & \text{if } t = 0, \\ 1, & \text{if } |t| > a\lambda, \\ a\lambda/|t|, & \text{otherwise.} \end{cases} \quad (4.10)$$

Compare LIC with IC, the only difference is that we replace  $|A_\lambda|$  by  $\sum_{j=1}^p L(\hat{\beta}_j(\lambda); a, \lambda)$ . If all the estimated coefficients  $\hat{\beta}_j(\lambda)$ ,  $j = 1, \dots, p$  are greater than  $a\lambda$ , then LIC and IC are equal. However, if there exists some nonzero coefficient whose absolute value is below  $a\lambda$ , then LIC can be much greater than IC.

LIC assigns weight 1 to all large nonzero coefficients, but larger weight to small nonzero coefficient. Therefore, LIC selects the set of coefficients which are either large,

or small but important. Applying MCP and select  $\lambda$  using LIC, we are able to select the large coefficient set  $L \subset A_0$ . The theoretical results can be found in section 4.4 and 4.5.

### 4.3.2 Second Step

Following the first step, we have been able to identify the subset  $L$  which contains large coefficients, the next step we only want to penalize the remaining coefficients. i.e. we minimize

$$Q_n(\boldsymbol{\beta}_L, \boldsymbol{\beta}_S; \lambda_n) = \frac{1}{2} \|\mathbf{Y} - \mathbf{X}_L \boldsymbol{\beta}_L - \mathbf{X}_S \boldsymbol{\beta}_S\|^2 + n \sum_{j \in S} P_{\lambda_n}(\beta_j) \quad (4.11)$$

The first problem is what penalty to use for this step, and our suggestion is LASSO or ALASSO. The reason is as following:  $S$  only contains small or zero coefficients, and LASSO works the best for them (see figure 4.1), and it is faster than nonconcave penalties. In fact, we have tried a few choices in the second step, and LASSO/ALASSO does perform much better than the others. The comparison can be found in numerical studies.

Another problem is the optimization of  $Q_n(\boldsymbol{\beta}_L, \boldsymbol{\beta}_S; \lambda_n)$ . Let the estimator be  $\hat{\boldsymbol{\beta}}^T = (\hat{\boldsymbol{\beta}}_L^T, \hat{\boldsymbol{\beta}}_S^T) = \operatorname{argmin} Q_n(\boldsymbol{\beta}_L, \boldsymbol{\beta}_S; \lambda_n)$ . As we can see, for any particular  $\boldsymbol{\beta}_S$ , the  $\boldsymbol{\beta}_L$  that minimizes  $Q_n(\boldsymbol{\beta}_L, \boldsymbol{\beta}_S; \lambda_n)$  must satisfy

$$\mathbf{X}_L^T \mathbf{X}_L \boldsymbol{\beta}_L = \mathbf{X}_L^T (\mathbf{Y} - \mathbf{X}_S \boldsymbol{\beta}_S)$$

as long as  $\mathbf{X}_L^T \mathbf{X}_L$  is not singular (which is usually true since  $|L| \leq p_0 \leq n$ ), we can substitute  $\boldsymbol{\beta}_L$  by  $(\mathbf{X}_L^T \mathbf{X}_L)^{-1} \mathbf{X}_L^T (\mathbf{Y} - \mathbf{X}_S \boldsymbol{\beta}_S)$  in (4.11), and transform the problem into minimizing

$$\tilde{Q}_n(\boldsymbol{\beta}_S; \lambda_n) = \frac{1}{2} \|(\mathbf{I} - \mathbf{P}_L) \mathbf{Y} - (\mathbf{I} - \mathbf{P}_L) \mathbf{X}_S \boldsymbol{\beta}_S\|^2 + \sum_{j \in S} P_{\lambda}(\beta_j) \quad (4.12)$$

where  $\mathbf{P}_L = \mathbf{X}_L(\mathbf{X}_L^T\mathbf{X}_L)^{-1}\mathbf{X}_L^T$  is the projection matrix of  $\mathbf{X}_L$ .

It follows directly that

$$\begin{cases} \hat{\boldsymbol{\beta}}_S = \operatorname{argmin} \tilde{Q}_n(\boldsymbol{\beta}_S; \lambda_n) \\ \hat{\boldsymbol{\beta}}_L = (\mathbf{X}_L^T\mathbf{X}_L)^{-1}\mathbf{X}_L^T\mathbf{Y} - (\mathbf{X}_L^T\mathbf{X}_L)^{-1}\mathbf{X}_L^T\mathbf{X}_S\hat{\boldsymbol{\beta}}_S \end{cases} \quad (4.13)$$

The only remaining problem is how to minimize (4.12).

One intuitive way is to let  $\tilde{\mathbf{Y}} = (\mathbf{I} - \mathbf{P}_L)\mathbf{Y}$  and  $\tilde{\mathbf{X}}_S = (\mathbf{I} - \mathbf{P}_L)\mathbf{X}_S$ . We can directly apply existing algorithms (glmnet, MCP, etc.) to  $(\tilde{\mathbf{X}}_S, \tilde{\mathbf{Y}})$ . That is, we treat  $(\tilde{\mathbf{X}}_S, \tilde{\mathbf{Y}})$  as the new pair of independent and dependent variables and obtain the corresponding  $\hat{\boldsymbol{\beta}}_S$ . One small problem is that the errors corresponding to the new pair have the covariance matrix  $\sigma^2(\mathbf{I} - \mathbf{P}_L)$ , which is slightly different from  $\sigma^2\mathbf{I}$ . Below is a more accurate way to minimize  $\tilde{Q}_n(\boldsymbol{\beta}_S; \lambda_n)$ .

Let  $P_\lambda(\beta_j) = \lambda|\beta_j|$ , we can optimize  $\tilde{Q}_n(\boldsymbol{\beta}_S; \lambda_n)$  in a more accurate way: (let  $r$  be the size of  $L$ )

1. Do eigen-decomposition to  $\mathbf{I} - \mathbf{P}_L = \mathbf{Q} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q}^T = (\mathbf{Q}_1, \mathbf{Q}_2) \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{Q}_1^T \\ \mathbf{Q}_2^T \end{pmatrix}$ .  
 $\mathbf{Q}_1$  is the  $n \times (n - r)$  submatrix of  $\mathbf{Q}$ , and  $\mathbf{Q}_1^T\mathbf{Q}_1 = \mathbf{I}_{n-r}$ .

2. Let

$$\hat{Q}_n(\boldsymbol{\beta}_S; \lambda_n) = \frac{1}{2} \|\mathbf{Q}_1^T(\mathbf{I} - \mathbf{P}_L)\mathbf{Y} - \mathbf{Q}_1^T(\mathbf{I} - \mathbf{P}_L)\mathbf{X}_S\boldsymbol{\beta}_S\|^2 + n\lambda \sum_{j \in S} |\beta_j| \quad (4.14)$$

Since  $\mathbf{Q}_1^T$  is full-rank, thus  $\hat{\boldsymbol{\beta}} = \operatorname{argmin} \tilde{Q}_n(\boldsymbol{\beta}_S; \lambda_n) = \operatorname{argmin} \hat{Q}_n(\boldsymbol{\beta}_S; \lambda_n)$ .

3. Let  $\tilde{\mathbf{Y}} = \mathbf{Q}_1^T(\mathbf{I} - \mathbf{P}_L)\mathbf{Y}$  and  $\tilde{\mathbf{X}}_S = \mathbf{Q}_1^T(\mathbf{I} - \mathbf{P}_L)\mathbf{X}_S$ , then the transformed error term is  $\tilde{\boldsymbol{\varepsilon}} = \mathbf{Q}_1^T(\mathbf{I} - \mathbf{P}_L)\boldsymbol{\varepsilon}$  with covariance matrix  $\mathbf{I}_{n-r}$ . Therefore,  $\hat{Q}_n(\boldsymbol{\beta}_S; \lambda_n)$  is

equivalent to

$$\hat{Q}_n(\boldsymbol{\beta}_S; \lambda_n) = \frac{1}{2} \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}_S \boldsymbol{\beta}_S\|^2 + (n-r)\lambda \sum_{j \in S} \frac{|\beta_j|}{\frac{n-r}{n}}. \quad (4.15)$$

4. Minimize (4.15) to get  $\hat{\boldsymbol{\beta}}_S$ : (4.15) is similar to Adaptive LASSO problem, whose algorithm is described in Zou (2006).

The aforementioned procedure also works when the adaptive LASSO penalty is used in the second step. However, in practice, the different between  $\sigma^2(\mathbf{I} - \mathbf{P}_L)$  and  $\sigma^2\mathbf{I}$  is small, and the second algorithm has no advantage in terms of performance and computation cost, thus we adopt first algorithm in the numerical studies.

## 4.4 Effects of Small Coefficients

In this section, we show that the existence of small coefficients causes a few problems in estimation: For any linear combination of estimated coefficient  $\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}$ , the existence of small coefficients add bias, and it also amplifies the asymptotic variance. In addition, with small coefficients, it is much harder or even impossible to achieve selection consistency.

### 4.4.1 General Settings

In high dimensions,  $\beta'_j$  ( $j = 1, \dots, p$ ) are allowed to depend on  $n$ . However, for the purpose of simplicity, we do not add  $n$  to the subscript.

Let the true value of  $\boldsymbol{\beta}$  be

$$\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_p^*)^T = \begin{pmatrix} \beta_1^* \\ \beta_2^* \end{pmatrix}, \quad (4.16)$$

without loss of generality, assume that  $\beta_2^* = \mathbf{0}$ . We denote length of  $\beta$  and  $\beta_1$  by  $p$  and  $p_0$ , respectively. We do allow  $p$  and  $p_0$  goes to infinity as  $n \rightarrow \infty$ , but we don't add subscript  $n$  for the purpose of simplicity. Besides, assume the design matrix is standardized such that  $\|\mathbf{X}_{\cdot j}\|^2 = n$  for any  $j$ ,  $1 \leq j \leq p$ .

Denote  $\mathbf{X}_1$  as the  $n \times p_0$  submatrix of  $\mathbf{X}$  that corresponds to the nonzero coefficients  $\beta_1$ , and  $\mathbf{X}_2$  as the submatrix of the remaining columns. Let  $\mathbf{C}_n = n^{-1}\mathbf{X}^T\mathbf{X}$ , and write  $\mathbf{C}_n$  as

$$\mathbf{C}_n = \begin{bmatrix} \mathbf{C}_{11,n} & \mathbf{C}_{12,n} \\ \mathbf{C}_{21,n} & \mathbf{C}_{22,n} \end{bmatrix},$$

where  $\mathbf{C}_{11,n}$  is the  $p_0 \times p_0$  submatrix that corresponds to the nonzero coefficients  $\beta_1^*$ . i.e.  $\mathbf{C}_{11,n} = n^{-1}\mathbf{X}_1^T\mathbf{X}_1$ .

We shall make use of the following regularity conditions:

(C.1) The eigenvalues of  $\mathbf{C}_{11,n}$  is bounded away from 0 and  $\infty$  uniformly. That is,

$$\mathbf{C}_{11,n} \in \mathcal{U}(c_1, c_2) = \{\mathbf{A} : 0 < c_1 \leq \boldsymbol{\alpha}^T \mathbf{A} \boldsymbol{\alpha} \leq c_2 < \infty, \forall \|\boldsymbol{\alpha}\| = 1\} \quad (4.17)$$

(C.2) Define the maximum concavity of penalty  $P_\lambda$  as

$$\kappa(P_\lambda) = \sup_{t_1 < t_2 \in (0, \infty)} \frac{P'_\lambda(t_1) - P'_\lambda(t_2)}{t_2 - t_1},$$

$\kappa(P_\lambda)$  satisfies

$$\kappa(P_\lambda) < \frac{1}{2}(1 - \delta) \cdot c_1$$

Condition (C.1) requires that the eigenvalues of  $\mathbf{C}_{11,n}$  is strictly bounded away from 0 and  $\infty$ . Similar conditions can be found in Bickel and Levina (2008); Kim et al. (2008).

Condition (C.2) is similar to the sparse convexity condition in Zhang (2010), which requires that the maximum concavity of the penalty function is smaller than the product of the smallest eigenvalue of  $\mathbf{C}_{11,n}$  and a constant. A similar assumption is utilized to prove that the desired local minimizer of  $Q_n(\boldsymbol{\beta}; \lambda)$  also be the global minimizer (Fan and Lv (2011)).

#### 4.4.2 Amplification of Variance

We have discussed the limitation of Magnitude assumption, which indicates that the small coefficients deteriorate the selection accuracy. Furthermore, we will show that they also amplify the variance of estimators.

Recall that in Chapter 2, we have shown that the model selection consistency can be achieved by a general family of nonconcave penalty functions. Here we simply start from the point that the selected model is identical to the true model, the following theorem shows the asymptotic distribution of estimators. Furthermore, we can derive the asymptotic bias and variance based on it.

**Theorem 4.1.** *Let  $\mathbf{A}_n$  be a  $q \times p_0$  matrix, if  $\mathbf{A}_n \mathbf{C}_{11,n} \mathbf{A}_n^T \rightarrow \mathbf{Q}$  and  $\max_i \|(\mathbf{A}_n \mathbf{X}_1^T)_{\cdot i}\| = o(\sqrt{n})$ . If the selected model is the same as the truth, then the estimator of nonzero coefficients  $\hat{\boldsymbol{\beta}}_1$  converges in distribution to a multivariate normal distribution:*

$$\sqrt{n} \mathbf{A}_n \left[ (\mathbf{C}_{11,n} + \Gamma_n)(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) + \mathbf{b}_n \right] \rightarrow_d N(0, \sigma^2 \mathbf{Q}) \quad (4.18)$$

where

$$\mathbf{b}_n^T = (P'_{\lambda_n}(|\beta_1^*|)sgn(\beta_1^*), \dots, P'_{\lambda_n}(|\beta_{p_0}^*|)sgn(\beta_{p_0}^*)) \quad (4.19)$$

$$\Gamma_n = \text{diag}\{P''_{\lambda_n}(|\beta_1^*|), \dots, P''_{\lambda_n}(|\beta_{p_0}^*|)\} \quad (4.20)$$

**Remark 2.** If the loss function is defined using likelihood:  $Q_n(\boldsymbol{\beta}; \lambda) = \sum_{i=1}^n l(\mathbf{X}_i, \mathbf{Y}_i, \boldsymbol{\beta}) - n \sum_{j=1}^{p_0} P_{\lambda_n}(\beta_j)$ , a similar result based on likelihood is also true. Simply replacing  $\mathbf{C}_{11,n}$  by  $\mathbf{I}_1(\boldsymbol{\beta}_1^*)$ , where  $\mathbf{I}_1(\boldsymbol{\beta}_1^*)$  is the information matrix relevant to nonzero coefficients, all the proofs and results should follow immediately.

With the result of theorem 4.1, we obtain the asymptotic bias and variance of  $\hat{\boldsymbol{\beta}}_1$ :

$$\text{bias} = -(\mathbf{C}_{11,n} + \Gamma_n)^{-1} \mathbf{b}_n \quad (4.21)$$

$$\text{var} = n^{-1} \sigma^2 (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \mathbf{C}_{11,n} (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \quad (4.22)$$

These expressions are true for a general family of nonconcave penalties, including MCP and SCAD.

Let  $\boldsymbol{\alpha}$  be a vector of length  $p_0$  with  $\|\boldsymbol{\alpha}\| = 1$ , then the mean square errors(MSE) of  $\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1$  is as follows:

$$MSE(\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1) = \frac{\sigma^2}{n} [\boldsymbol{\alpha}^T (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \mathbf{C}_{11,n} (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \boldsymbol{\alpha}] + [\boldsymbol{\alpha}^T (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \mathbf{b}_n]^2. \quad (4.23)$$

We are going to discuss how MSE is affected by small coefficients. We consider two cases, and argue that the MSE in case 1 is always smaller than that in case 2.

Case 1 When  $\min_{1 \leq j \leq p_0} (|\beta_j^*|) > a\lambda_n$ , i.e. all the coefficients are large. By definition, we

have  $\mathbf{b}_n = \mathbf{0}$  and  $\Gamma_n = \mathbf{0}$ . That means,

$$\text{bias}_1(\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1) = 0, \quad (4.24)$$

$$\text{var}_1(\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1) = n^{-1} \sigma^2 (\boldsymbol{\alpha}^T \mathbf{C}_{11,n}^{-1} \boldsymbol{\alpha}) \quad (4.25)$$

Therefore, the MSE is  $n^{-1} \sigma^2 (\boldsymbol{\alpha}^T \mathbf{C}_{11,n}^{-1} \boldsymbol{\alpha})$ , and it is identical to that of the oracle estimator.

Case 2 Otherwise, there exist some coefficients whose magnitude is below  $a\lambda_n$ . That means,  $\mathbf{b}_n \neq \mathbf{0}$  and  $\Gamma_n \neq \mathbf{0}$ . And we have

$$\begin{aligned} \text{bias}_2(\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1) &= -\boldsymbol{\alpha}^T (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \mathbf{b}_n \neq \mathbf{0}, \\ \text{var}_2(\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1) &= n^{-1} \sigma^2 [\boldsymbol{\alpha}^T (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \mathbf{C}_{11,n} (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \boldsymbol{\alpha}] \end{aligned} \quad (4.26)$$

If we can show that  $\text{var}_1(\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1) \leq \text{var}_2(\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1)$  is true for any unit vector  $\boldsymbol{\alpha}$ , then it follows immediately that MSE in case 1 is less than that in case 2.

In other words, it remains to show that

$$\boldsymbol{\alpha}^T \mathbf{C}_{11,n}^{-1} \boldsymbol{\alpha} \leq \boldsymbol{\alpha}^T (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \mathbf{C}_{11,n} (\mathbf{C}_{11,n} + \Gamma_n)^{-1} \boldsymbol{\alpha} \quad \forall \boldsymbol{\alpha}, \text{ where } \|\boldsymbol{\alpha}\| = 1. \quad (4.27)$$

$\Gamma_n$  is a diagonal matrix with its  $j$ th diagonal element being  $P''_{\lambda_n}(|\beta_j^*|)$ , which is either 0 (if  $|\beta_j^*| \geq a\lambda_n$ ) or a negative number. Therefore,  $\Gamma_n$  is negative semidefinite, i.e.  $\Gamma_n \preceq \mathbf{0}$ . Furthermore, according to condition (C.2), all the eigenvalues of  $\mathbf{C}_{11,n}$  are greater than  $\max_{1 \leq j \leq p_0} P''_{\lambda_n}(|\beta_j^*|)$ , thus  $\mathbf{C}_{11,n} + \Gamma_n$  is positive definite. Remember that if two matrices  $\mathbf{A}$  and  $\mathbf{B}$  are both positive definite, and if  $\mathbf{A}^{-1} \succeq \mathbf{B}^{-1}$ , then  $\mathbf{B} \succeq \mathbf{A}$ . (because  $\mathbf{B} - \mathbf{A} = \mathbf{A}(\mathbf{A}^{-1} - \mathbf{B}^{-1})\mathbf{B} \succeq \mathbf{0}$ .) There-

fore, (4.27) is true iff  $(\mathbf{C}_{11,n} + \Gamma_n)\mathbf{C}_{11,n}^{-1}(\mathbf{C}_{11,n} + \Gamma_n) \preceq \mathbf{C}_{11,n}$ . In fact, we have

$$(\mathbf{C}_{11,n} + \Gamma_n)\mathbf{C}_{11,n}^{-1}(\mathbf{C}_{11,n} + \Gamma_n) - \mathbf{C}_{11,n} = 2\Gamma_n + (-\Gamma_n)\mathbf{C}_{11,n}^{-1}(-\Gamma_n) \preceq \Gamma_n \preceq 0 \quad (4.28)$$

Thus we have shown that (4.27) is true.

comparing case 1 and case 2, the variance in case 2 is always greater than that in case 1, and the bias in case 2 is non-zero while that in case 1 is exactly zero. Therefore, the smallest MSE of  $\boldsymbol{\alpha}^T \hat{\boldsymbol{\beta}}_1$  is obtained when  $\min_{1 \leq j \leq p_0} (|\beta_j^*|) > a\lambda_n$ , and the minimum is  $\boldsymbol{\alpha}^T \mathbf{C}_{11}^{-1} \boldsymbol{\alpha}$ .

To summarize, the optimal MSE of nonconcave penalties is equal to the MSE of Oracle Properties. To achieve the optimum, the selected model must be true, and the magnitude of any nonzero coefficients must be great than  $a\lambda_n$ . If there exists some coefficients whose magnitude is below  $a\lambda_n$ , then the variance is amplified, and a small bias also occurs.

Therefore, small coefficients causes two problems: the failure of selection accuracy and the amplification of variance.

### 4.4.3 An Estimator for Large Coefficients

In this literature, the choice of  $\lambda_n$  is an important issue. By shifting  $\lambda_n$ , different estimators can be found. A popular goal is to find the  $\lambda_n$  whose corresponding estimator is close to the oracle estimator. However, when small coefficients exist, this goal become too hard to be achieved. In this case, we found an estimator which is meaningful, and we shall also use it for PPR.

We assume the following setting in this section:

$$(C.3) \quad p_0 = O(n^d), \text{ where } 0 \leq d < 1. \text{ Without loss of generality, all the coefficient } \beta_1, \dots, \beta_p$$

can be partitioned into three sets:

- (1) Large Coefficients  $\beta_1, \beta_2, \dots, \beta_{q_0}$ :  $\min_{1 \leq j \leq q_0} |\beta_j| > M \cdot n^{-1/2+d}$  for some  $M > 0$ .
- (2) Small Coefficients  $\beta_{q_0+1}, \beta_{q_0+2}, \dots, \beta_{p_0}$ :  $|\beta_j| = o(n^{-1/2+d/2})$  and  $\beta_j \neq 0, \forall j, q_0 + 1 \leq j \leq p_0$ .
- (3) Zero Coefficients  $\beta_{p_0+1}, \beta_{p_0+2}, \dots, \beta_p$ :  $\beta_j = 0, \forall j, p_0 + 1 \leq j \leq p$ .

Furthermore, we split the design matrix into three parts:

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{X}_2)$$

where  $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{X}_2$  are the corresponding design matrices for large, small, and zero coefficients, with dimension  $n \times q_0, n \times (p_0 - q_0), n \times (p - p_0)$ .

The following theorem shows that we can find an interesting estimator which only identifies the set of large coefficients using a proper  $\lambda_n$ .

**Theorem 4.2.** *Assume that  $E(\varepsilon_1^{2k}) < \infty$  for some  $k > 0$ , under condition (C.1) and (C.3), suppose  $\lambda_n$  satisfies (1)  $\lambda_n/(n^{-1/2+d}) \rightarrow \infty$  (2)  $\lambda_n \leq a^{-1} \min_{1 \leq j \leq q_0} |\beta_j|$  (3)  $p =$*

*$o((n\lambda_n^2)^k)$ . With probability going to 1,  $\tilde{\beta}^{(0)} = \begin{pmatrix} (\mathbf{Z}_1^T \mathbf{Z}_1)^{-1} \mathbf{Z}_1^T \mathbf{Y} \\ \mathbf{0}_{p_0 - q_0} \\ \mathbf{0}_{p - p_0} \end{pmatrix}$  is a local solution of*

*MCP penalized regression.*

Note that we do not require anything similar to Magnitude Assumption for this theorem. In fact, the smallest signal can be extremely small. With a properly chosen  $\lambda_n$ , both small coefficients and zero coefficients are estimated as zeroes with  $p \rightarrow 1$ . And the sub-vector for large coefficient estimator is the same as ordinary least square estimator of  $\mathbf{Y}$  onto  $\mathbf{Z}_1$ . This estimator treats only the large coefficients instead of all nonzero

coefficients as the true model, and achieve a pseudo “Oracle Property”.

This also leads to an interesting thought: if some coefficients are so small that they are blinded by the errors, is it really important to identify them? Isn't it more reasonable to identify the larger ones and to target for better MSE?

This theorem tells us that we are able to consistently identify the large coefficients, and the only problem in practice is to find such  $\lambda_n$ . This issue will be discussed in section 4.5.

#### 4.4.4 A Dilemma in Fixed Dimensions

To further understand the trouble that small coefficients make, we present a dilemma in fixed dimensions. When  $p$  is fixed, if there exists some small nonzero coefficient  $\beta_j$  such that  $\beta_j = O(n^{-1/2})$ , we show that no choice of  $\lambda$  leads to the Oracle Property.

Similar to condition (C.3), we have the following condition in fixed dimensions.

(C.3')  $p$  is fixed, without loss of generality, all the coefficient  $\beta_1, \dots, \beta_p$  can be partitioned into three sets: (none of them is empty)

- (1) There exists  $\delta > 0$  and  $M > 0$ , such that large coefficients  $\beta_1, \beta_2, \dots, \beta_{q_0}$  satisfy  $\min_{1 \leq j \leq q_0} |\beta_j| > M \cdot n^{-1/2+\delta}$ ;
- (2) Small Coefficients  $\beta_{q_0+1}, \beta_{q_0+2}, \dots, \beta_{p_0}$  satisfy  $|\beta_j| = O(1/\sqrt{n})$ ,  $j = q_0 + 1, \dots, p_0$ .
- (3) Zero Coefficients  $\beta_{p_0+1}, \beta_{p_0+2}, \dots, \beta_p$ :  $\beta_j = 0$ ,  $\forall j, p_0 + 1 \leq j \leq p$

Consider three options of  $\lambda_n$ : (1)  $\lambda_n = O(n^{-1/2})$ , (2)  $\lambda_n = o(n^{-1/2+\delta})$  and  $\sqrt{n}\lambda_n \rightarrow \infty$ , (3) otherwise. We shall show that none of these choices lead to selection consistency.

For the first option, in lemma 4.1, we claim that there is a positive probability that

some zero coefficients will be estimated as nonzeros, even if the sample size tends to infinity.

**Lemma 4.1.** *If  $\lambda_n = O(n^{-1/2})$ , then there exists  $\alpha$ ,  $0 \leq \alpha < 1$ , such that*

$$\limsup_{n \rightarrow \infty} P(\hat{\beta}_j = 0, \forall j \geq p_0 + 1) \leq \alpha$$

For the second option, in lemma 4.2, generalizing theorem 4.2 to the fixed dimensions, we claim that both small and zero coefficients are estimated as zeroes.

**Lemma 4.2.** *Under the setting of case (C.3'), if  $\lambda_n = o(n^{-1/2+\delta})$  and  $\sqrt{n}\lambda_n \rightarrow \infty$ , with probability going to 1,  $\tilde{\beta}^{(0)}$  is a local minimizer of  $Q(\beta; \lambda_n)$ . ( $\tilde{\beta}^{(0)}$  is defined in Theorem 4.2)*

For the third option, things are even worse than the second case. That is, even less variables are selected.

Hence, there is a dilemma in this situation: in order to estimate all the zero coefficients as zeroes (with  $p \rightarrow 1$ ), we must choose  $\sqrt{n}\lambda_n \rightarrow \infty$ ; but once we do that, the small coefficients are estimated as zeroes at the same time. There is no intermediate choices, and we can never achieve selection consistency in this situation. This dilemma demonstrates the importance of the Magnitude Assumption in fixed dimensions.

## 4.5 The L-consistency of LIC

Recall that we have defined LIC as

$$\text{LIC}(\lambda; a) = \log \hat{\sigma}_\lambda^2 + \frac{C_n}{n} \sum_{j=1}^p L(\hat{\beta}_j(\lambda); a, \lambda) \quad (4.29)$$

where  $L(\cdot)$  is a function that aims to penalize more on small nonzero coefficients.

$$L(t; a, \lambda) = \begin{cases} 0, & \text{if } t = 0, \\ 1, & \text{if } |t| > a\lambda, \\ a\lambda/|t|, & \text{otherwise.} \end{cases} \quad (4.30)$$

We will show that, with probability going to 1, LIC combined with MCP selects a set of coefficients  $L$ , such that  $L$  contains all the large coefficients and  $L$  contains no zero coefficients. We name it as the L-consistency.

We shall use the following notations in this section: without loss of generality, let the index set of true model be  $A_0 = \{j : 1 \leq j \leq p_0\}$ , and the index of large coefficients be  $A_1 = \{j : 1 \leq j \leq q_0\}$ , and the index set of the selected model by  $\lambda$  be  $A_\lambda = \{j : 1 \leq j \leq p, \hat{\beta}_j(\lambda) \neq 0\}$ , where  $\hat{\beta}_j(t)$  is the estimated coefficient when  $\lambda = t$ . We define two sets of models as following:

$$\begin{aligned} \Lambda_{1-} &= \{\lambda : \lambda \geq 0, A_1 \not\subset A_\lambda\}, \\ \Lambda_{1+} &= \{\lambda : \lambda \geq 0, A_1 \subset A_\lambda, A_\lambda \neq A_1\}. \end{aligned}$$

By the definition, we see that  $\Lambda_{1-}$  is the set of  $\lambda$  which leads to a model that does not cover all large coefficients, while  $\Lambda_{1+}$  is the set of  $\lambda$  which leads to a model that does cover all large coefficients, but cover more than that. Therefore, if we can show that any model from either  $\Lambda_{1-}$  or  $\Lambda_{1+}$  has a greater LIC value than  $A_1$  does, then we can claim that  $A_1$  will be chosen by LIC.

We shall need the following asymptotic identifiability condition, which guarantees the identification of the true model.

(C.4) We only consider the models which contains no more than  $K_n$  attributes. Assume that there exists a positive constant  $\delta$ , such that

$$\liminf_{n \rightarrow \infty} \min_{A \notin A_0, |A| \leq K_n} \left\{ n^{-1} \|(\mathbf{I}_n - \mathbf{P}_A)\boldsymbol{\mu}\|^2 \right\} \geq \delta \quad (4.31)$$

where  $\mathbf{I}_n$  is the  $n \times n$  identity matrix,  $\mathbf{P}_A$  is the projection matrix corresponding to the column space of  $\mathbf{X}_A$ .

Condition (C.4) is a popular assumption. (Chen and Chen, 2008; Wang and Zhu, 2011) It means that no model of comparable size other than the true submodel can predict the response almost equally well. And it is used to guarantee that the underfitted model and the true model are separated by a small gap.

**Theorem 4.3.** *Under the condition (C.1)(C.3)(C.4), if  $d \leq 1/2$  and  $\boldsymbol{\varepsilon}$  are i.i.d sub-gaussian errors;  $K_n \log p \log n = o(n)$ ,  $K_n^{3/2} \log p = O(n)$  and  $C_n p_0 = o(n)$ . Furthermore, if  $\|\boldsymbol{\beta}_{(2)}\| = O(\sqrt{\frac{\log p}{n}})$ , and by picking  $C_n / \log p \rightarrow \infty$ , then  $\lambda_n = \operatorname{argmin}_{\lambda \geq 0} LIC(\lambda; a)$  selects a set  $L$ , which satisfies*

$$P(L = A_1) \rightarrow 1 \quad (4.32)$$

The theorem above means that LIC does select the desired subset of true model -  $L$ , which contains all the large coefficients, and contains no zero coefficients.

Both properties are important:

- $L$  contains all the large coefficients, thus all the large coefficients are not penalized in the second step, and they all remain in the final model. If some large coefficients are missing, then the second step needs to handle both large and small coefficients simultaneously, which is hard for any known penalty function and bad for tuning parameter selection.

- $L$  contains no zero coefficients. If some zero coefficients are included, then they will not be penalized in the second step, and will eventually remain in the final model. This increases the false positive rate.

## 4.6 Numerical Studies

### 4.6.1 Simulations

We conduct comprehensive simulations to compare the performances of our methods to existing methods. And our method does outperform in most cases.

First of all, we summarize our suggested method as following: in the first stage, apply MCP combined with LIC (4.9) to identify the indexes of large nonzero coefficients (only the indexes, we do not use the coefficient estimates); in the second stage, apply LASSO/ALASSO combined with HIC, and minimize (4.7) to get the estimates of each coefficients.

In the simulation, we compare our proposed methods PPR to the famous MCP, LASSO, ALASSO penalties. Depending on the penalty used in the second step, our PPR is referred as “P-MCP”, “P-LASSO”, “P-ALASSO”. Note that we only recommend P-LASSO and P-ALASSO, and we keep P-MCP only for comparison purpose. All the six methods are compared based on six metrics which measures their selection accuracy and estimation accuracy.

Simulations are conducted under four settings. Different settings differ in the dimensionality or the true parameter. The dimensions are either  $(n = 100, p = 1000, p_0 = 10)$  or  $(n = 200, p = 3000, p_0 = 20)$ . As for the true parameters  $\beta$ , we have two different setups:

- **Type 1  $\beta$ .** In the first setup, nonzero coefficients are well bounded and clustered together. Define  $\mathbf{b}_{10}^T = (1, -2, 3, 0, 0, 0, -1.5, 2, 0, 0)$ , we let

$$\begin{aligned}\beta^T &= (\mathbf{b}_{10}^T, \mathbf{0}_{490}^T, \mathbf{b}_{10}^T, \mathbf{0}_{490}^T) && \text{when } p = 1000 \\ \beta^T &= (\mathbf{b}_{10}^T, \mathbf{0}_{740}^T, \mathbf{b}_{10}^T, \mathbf{0}_{740}^T, \mathbf{b}_{10}^T, \mathbf{0}_{740}^T, \mathbf{b}_{10}^T, \mathbf{0}_{740}^T) && \text{when } p = 3000\end{aligned}\tag{4.33}$$

- **Type 2  $\beta$ .** In the second setup,  $p_0$  nonzero coefficients are random generated from  $Uniform(-3, 3)$  distribution, and their locations are randomly selected from all  $p$  possible locations. Therefore, the nonzero coefficients are not bounded below and not clustered.

We create four settings based on the combination of dimensionality and  $\beta$  types.

In each replication, we generate random samples  $\{y_i, \mathbf{X}_i\}$ ,  $i = 1, \dots, n$  with linear model  $y_i = \mathbf{X}_i\beta + \varepsilon_i$ . The length- $p$  vector  $\mathbf{X}_i$  is generated from  $N(\mathbf{0}, \Sigma)$ , where  $\Sigma = \mathbf{I}$  (denoted as “Ind”) or  $\Sigma = \{\Sigma_{i,j}\} = \rho^{|i-j|}$  (denoted as “AR1( $\rho$ )”). The error terms  $\varepsilon_i$ ’s are *i.i.d* distributed with  $N(0, 1)$  distribution. The number of replications is 500 when  $p = 1000$  or 200 when  $p = 3000$ .

As has been discussed in previous sections, the performance of penalized regression not only depends on the method itself, but also relies on the tuning parameter selection associated with the method. As suggested in the literature, we use IC(4.8) with high  $C_n$  value where  $C_n$  depends on dimensionality. We tried three tuning parameter selection criteria: BIC (when  $C_n = \log n$ ), HBIC (when  $C_n = \log \log n \log n$ ), and HIC (when  $C_n = \log \log n \log p$ ). It turns out, under our four settings, HBIC is the best choice for LASSO/ALASSO; while HIC is the best for the other methods. Thus the reported statistics in table 4.3-4.7 are based on the best combination of method-criterion. That is, we did not use a single criterion which favors our method, and we rule out the possibility

that our advantage is due to a better tuning procedure.

In each replication, we fit the model and tune the parameter based on the training set. With the obtained estimators, we predict the target variable in the testing set. This rule out the possibility of overfitting.

To compare their performance, we report the following six metrics. They are based on the average of all replications.

- MSE: the mean squared errors of predictions on  $Y$  in the testing set.
- TM: the proportion of replications that the true model is exactly identified.
- TP: average number of nonzero coefficients correctly estimated to be nonzero. (i.e. true positive)
- FP: average number of zero coefficients incorrectly estimated to be nonzero. (i.e. false positive)
- Size: the average number of coefficients estimated to be nonzero.
- $l_2$  loss: the average Euclidean norm of the difference between estimated coefficients and the truth. (i.e.  $\|\hat{\beta} - \beta\|^2$ )

A method is considered better than others if it has: smaller MSE, FP, and  $l_2$  loss; and larger TP, TM. And we hope the size to be smaller than or close to  $p_0$ . MSE and  $l_2$  loss are metrics for the estimation accuracy, while the other four are for selection consistency.

We reported the comparison of methods in Table 4.3 4.54.64.7, and the standard errors of statistics in Table 4.3 is reported in Table 4.4. As we can see from those tables, compared to the existing methods, our suggested methods (P-LASSO and P-ALASSO)

Table 4.3: Comparison of methods when  $n = 100$ ,  $p = 1000$ ,  $p_0 = 10$ ,  $\beta$  is type 1. (based on 500 replications)

Beta	$\Sigma$	Method	metrics					
			MSE	TM	TP	FP	Size	$\ \hat{\beta} - \beta^*\ $
$\beta$	Ind	MCP	1.1505	0.920	10.000	0.172	10.172	0.4315
		LASSO	3.8811	0.000	9.978	13.926	23.904	1.6246
		ALASSO	1.2245	0.832	10.000	0.202	10.202	0.4913
		P-MCP	2.8429	0.000	10.000	38.178	48.178	1.3499
		<b>P-LASSO</b>	<b>1.1333</b>	<b>0.978</b>	<b>10.000</b>	<b>0.024</b>	<b>10.024</b>	<b>0.4129</b>
		<b>P-ALASSO</b>	<b>1.1391</b>	<b>0.894</b>	<b>10.000</b>	<b>0.238</b>	<b>10.238</b>	<b>0.4207</b>
	AR1(0.5)	MCP	1.2361	0.820	9.992	1.226	11.218	0.5411
		LASSO	15.2163	0.000	3.204	3.610	6.814	5.4783
		ALASSO	9.7935	0.000	4.886	7.358	12.244	4.1601
		P-MCP	2.9193	0.000	9.992	38.278	48.270	1.4471
		<b>P-LASSO</b>	<b>1.1578</b>	<b>0.932</b>	<b>9.984</b>	<b>0.146</b>	<b>10.130</b>	<b>0.4815</b>
		<b>P-ALASSO</b>	<b>1.1542</b>	<b>0.878</b>	<b>9.992</b>	<b>0.390</b>	<b>10.382</b>	<b>0.4816</b>
$\beta/2$	Ind	MCP	2.1150	0.126	9.742	13.928	23.670	0.9829
		LASSO	4.3589	0.000	7.910	9.616	17.526	1.7079
		ALASSO	1.5986	0.056	9.440	3.466	12.906	0.7111
		P-MCP	2.9580	0.000	9.786	39.290	49.076	1.3689
		<b>P-LASSO</b>	<b>1.4913</b>	<b>0.314</b>	<b>9.514</b>	<b>2.236</b>	<b>11.750</b>	<b>0.6247</b>
		<b>P-ALASSO</b>	<b>1.4804</b>	<b>0.318</b>	<b>9.594</b>	<b>4.780</b>	<b>14.374</b>	<b>0.6161</b>
	AR1(0.5)	MCP	3.6634	0.000	7.940	25.170	33.110	1.7068
		LASSO	4.9064	0.000	2.182	2.144	4.326	2.8435
		ALASSO	3.9826	0.000	4.090	6.882	10.972	2.2960
		P-MCP	3.9857	0.000	8.074	46.080	54.154	1.8016
		<b>P-LASSO</b>	<b>3.4438</b>	<b>0.002</b>	<b>7.248</b>	<b>19.124</b>	<b>26.372</b>	<b>1.6696</b>
		<b>P-ALASSO</b>	<b>3.4339</b>	<b>0.006</b>	<b>7.496</b>	<b>34.142</b>	<b>41.638</b>	<b>1.6469</b>

have an improved performance. Although the performance of existing methods change from case to case, our methods beat the best of them in all the settings. MSE and  $l_2$  loss are usually 5% – 10% smaller, selection accuracy are better in terms of larger TM and smaller FP. There is a slightly drop in TP rate, which is the cost paid for other improvements. The advantage is not trivial since it is usually a few times larger than the standard errors. Besides, as the correlation  $\rho$  goes up, the advantage gets more and more

Table 4.4: *The standard error of statistics reported in Table 4.3*

Beta	$\Sigma$	Method	metrics					
			MSE	TM	TP	FP	Size	$\ \hat{\beta} - \beta^*\ $
$\beta$	Ind	MCP	0.0128	0.0120	0.0000	0.0365	0.0365	0.0047
		LASSO	0.0907	0.0000	0.0091	0.5378	0.5386	0.0156
		ALASSO	0.0143	0.0166	0.0000	0.0221	0.0221	0.0054
		P-MCP	0.0381	0.0000	0.0000	0.0858	0.0858	0.0066
		P-LASSO	0.0127	0.0056	0.0000	0.0066	0.0066	0.0044
		P-ALASSO	0.0127	0.0135	0.0000	0.0833	0.0833	0.0046
	AR1(0.5)	MCP	0.0252	0.0173	0.0080	0.1909	0.1891	0.0125
		LASSO	0.2186	0.0000	0.0692	0.6258	0.6636	0.0197
		ALASSO	0.1355	0.0000	0.0539	0.3228	0.3534	0.0225
		P-MCP	0.0388	0.0000	0.0080	0.1482	0.1467	0.0097
		P-LASSO	0.0267	0.0152	0.0085	0.1859	0.1841	0.0136
		P-ALASSO	0.0278	0.0172	0.0080	0.5008	0.4989	0.0139
$\beta/2$	Ind	MCP	0.0376	0.0141	0.0233	0.4133	0.4130	0.0157
		LASSO	0.1215	0.0000	0.1071	0.9814	1.0340	0.0218
		ALASSO	0.0282	0.0105	0.0318	0.1511	0.1554	0.0100
		P-MCP	0.0367	0.0000	0.0202	0.2323	0.2295	0.0075
		P-LASSO	0.0390	0.0124	0.0229	0.5817	0.5826	0.0159
		P-ALASSO	0.0411	0.0125	0.0218	1.1202	1.1198	0.0165
	AR1(0.5)	MCP	0.0811	0.0020	0.0799	0.2533	0.2302	0.0244
		LASSO	0.0616	0.0000	0.0522	0.5721	0.6110	0.0082
		ALASSO	0.0556	0.0000	0.0725	0.3204	0.3618	0.0147
		P-MCP	0.0820	0.0000	0.0774	0.1512	0.1160	0.0223
		P-LASSO	0.0810	0.0000	0.0786	1.0938	1.0844	0.0231
		P-ALASSO	0.0818	0.0020	0.0766	0.5350	0.5237	0.0229

obvious.

In addition, the performance of PPR is quite stable. LASSO usually performs the worst; MCP beats ALASSO when coefficients are large and correlation is low, ALASSO beats MCP in the other time. Therefore, their relative performance depends on the data. However, P-LASSO and P-ALASSO have a stable performance in all settings, and their performance is usually the best among the six methods. This is a big advantage in

Table 4.5: Comparison of methods when  $n = 100$ ,  $p = 1000$ ,  $p_0 = 10$ ,  $\beta$  is type 2. (based on 500 replications)

Beta	$\Sigma$	Method	metrics					
			MSE	TM	TP	FP	Size	$\ \hat{\beta} - \beta^*\ $
$\beta$	Ind	MCP	1.2228	0.784	9.984	0.328	10.312	0.4920
		LASSO	2.7894	0.000	9.990	77.778	87.768	1.2767
		ALASSO	1.5441	0.662	9.782	0.242	10.024	0.7070
		P-MCP	2.8636	0.000	9.992	38.030	48.022	1.3557
		<b>P-LASSO</b>	<b>1.1691</b>	<b>0.878</b>	<b>9.958</b>	<b>0.098</b>	<b>10.056</b>	<b>0.4370</b>
		<b>P-ALASSO</b>	<b>1.1655</b>	<b>0.832</b>	<b>9.972</b>	<b>0.268</b>	<b>10.240</b>	<b>0.4383</b>
	AR1(0.5)	MCP	1.2058	0.736	9.982	0.364	10.346	0.4798
		LASSO	2.8049	0.000	9.946	80.142	90.088	1.3808
		ALASSO	1.5584	0.476	9.626	0.512	10.138	0.7194
		P-MCP	2.9328	0.000	9.990	38.278	48.268	1.3941
		<b>P-LASSO</b>	<b>1.1742</b>	<b>0.850</b>	<b>9.936</b>	<b>0.120</b>	<b>10.056</b>	<b>0.4375</b>
		<b>P-ALASSO</b>	<b>1.1613</b>	<b>0.818</b>	<b>9.976</b>	<b>0.376</b>	<b>10.352</b>	<b>0.4361</b>
$\beta/2$	Ind	MCP	2.4884	0.000	8.634	17.278	25.912	1.1650
		LASSO	6.7262	0.000	3.456	0.724	4.180	2.4377
		ALASSO	1.8948	0.000	7.302	1.212	8.514	0.9065
		P-MCP	3.1728	0.000	8.772	38.440	47.212	1.4394
		<b>P-LASSO</b>	<b>1.9594</b>	<b>0.006</b>	<b>7.278</b>	<b>3.392</b>	<b>10.670</b>	<b>0.9163</b>
		<b>P-ALASSO</b>	<b>1.8489</b>	<b>0.008</b>	<b>7.778</b>	<b>7.628</b>	<b>15.406</b>	<b>0.8531</b>
	AR1(0.5)	MCP	2.1612	0.030	8.914	12.042	20.956	0.9888
		LASSO	4.7962	0.000	6.212	1.658	7.870	1.7488
		ALASSO	1.6649	0.010	8.184	1.170	9.354	0.7199
		P-MCP	3.1377	0.000	9.008	38.862	47.870	1.4429
		<b>P-LASSO</b>	<b>1.7294</b>	<b>0.044</b>	<b>8.146</b>	<b>2.136</b>	<b>10.282</b>	<b>0.7356</b>
		<b>P-ALASSO</b>	<b>1.6471</b>	<b>0.052</b>	<b>8.466</b>	<b>5.312</b>	<b>13.778</b>	<b>0.7041</b>

practice.

As for computation cost, our method involves two steps of penalization, and the dimensionality are almost the same in both steps. Thus the computation cost is roughly twice of the competitors'.

Comparing P-LASSO and P-ALASSO, they are quite similar. It seems P-ALASSO

Table 4.6: Comparison of methods when  $n = 200$ ,  $p = 3000$ ,  $p_0 = 20$ ,  $\beta$  is type 1. (based on 200 replications)

Beta	$\Sigma$	Method	metrics					
			MSE	TM	TP	FP	Size	$\ \hat{\beta} - \beta^*\ $
$\beta$	Ind	MCP	1.1075	0.985	20.000	0.015	20.015	0.4027
		LASSO	5.5447	0.000	19.945	24.390	44.335	2.0206
		ALASSO	1.3463	0.995	20.000	0.005	20.005	0.6066
		P-MCP	2.8484	0.000	20.000	69.005	89.005	1.3643
		<b>P-LASSO</b>	<b>1.1059</b>	<b>1.000</b>	<b>20.000</b>	<b>0.000</b>	<b>20.000</b>	<b>0.4017</b>
		<b>P-ALASSO</b>	<b>1.1095</b>	<b>0.950</b>	<b>20.000</b>	<b>0.055</b>	<b>20.055</b>	<b>0.4036</b>
	AR1(0.5)	MCP	1.1241	0.970	20.000	0.030	20.030	0.4519
		LASSO	29.1235	0.000	5.180	0.695	5.875	7.8606
		ALASSO	20.4145	0.000	8.085	4.600	12.685	6.2794
		P-MCP	2.8463	0.000	20.000	69.115	89.115	1.4101
		<b>P-LASSO</b>	<b>1.1230</b>	<b>1.000</b>	<b>20.000</b>	<b>0.000</b>	<b>20.000</b>	<b>0.4500</b>
		<b>P-ALASSO</b>	<b>1.1243</b>	<b>0.955</b>	<b>20.000</b>	<b>0.070</b>	<b>20.070</b>	<b>0.4528</b>
$\beta/2$	Ind	MCP	1.2761	0.435	19.925	0.915	20.840	0.5232
		LASSO	5.2045	0.000	18.215	12.380	30.595	1.8831
		ALASSO	1.3943	0.165	19.680	1.720	21.400	0.6080
		P-MCP	2.8547	0.000	19.980	68.675	88.655	1.3428
		<b>P-LASSO</b>	<b>1.2598</b>	<b>0.500</b>	<b>19.560</b>	<b>0.415</b>	<b>19.975</b>	<b>0.4810</b>
		<b>P-ALASSO</b>	<b>1.1936</b>	<b>0.585</b>	<b>19.880</b>	<b>0.475</b>	<b>20.355</b>	<b>0.4334</b>
	AR1(0.5)	MCP	1.6120	0.065	18.990	5.670	24.660	0.8726
		LASSO	5.6465	0.000	7.955	6.135	14.090	3.2256
		ALASSO	2.9239	0.000	19.655	67.565	87.220	1.4319
		P-MCP	2.8570	0.000	19.590	63.955	83.545	1.4041
		<b>P-LASSO</b>	<b>1.8660</b>	<b>0.080</b>	<b>18.015</b>	<b>3.155</b>	<b>21.170</b>	<b>0.9744</b>
		<b>P-ALASSO</b>	<b>1.7376</b>	<b>0.110</b>	<b>18.625</b>	<b>3.200</b>	<b>21.825</b>	<b>0.8861</b>

works slightly better when  $p$  or  $\rho$  is large. Besides, the simulation shows that P-MCP is much worse than Partial ALASSO/LASSO. This justify our intuition for partial penalization: when all the coefficients are small, the Magnitude Assumption is violated, MCP should have a bad performance. Remember that LASSO is worse than SCAD/MCP because it penalizes heavily on the large coefficients, while in the second stage, there are no large coefficients, thus lasso has no disadvantage here. That's why P-LASSO/P-ALASSO

Table 4.7: Comparison of methods when  $n = 200$ ,  $p = 3000$ ,  $p_0 = 20$ ,  $\beta$  is type 2. (based on 200 replications)

Beta	$\Sigma$	Method	metrics					
			MSE	TM	TP	FP	Size	$\ \hat{\beta} - \beta^*\ $
$\beta$	Ind	MCP	1.4053	0.085	18.795	0.720	19.515	0.6371
		LASSO	3.3666	0.000	18.720	76.765	95.485	1.4861
		ALASSO	2.0951	0.000	17.035	0.065	17.100	1.0223
		P-MCP	2.8922	0.000	19.305	68.610	87.915	1.3746
		<b>P-LASSO</b>	<b>1.3239</b>	<b>0.085</b>	<b>18.520</b>	<b>0.405</b>	<b>18.925</b>	<b>0.5587</b>
		<b>P-ALASSO</b>	<b>1.2972</b>	<b>0.085</b>	<b>18.610</b>	<b>0.425</b>	<b>19.035</b>	<b>0.5488</b>
	AR1(0.5)	MCP	1.3883	0.100	18.860	0.645	19.505	0.6089
		LASSO	3.3886	0.000	18.885	76.015	94.900	1.5486
		ALASSO	2.2361	0.000	17.155	0.080	17.235	1.0512
		P-MCP	2.9190	0.000	19.380	68.815	88.195	1.4015
		<b>P-LASSO</b>	<b>1.2831</b>	<b>0.115</b>	<b>18.655</b>	<b>0.315</b>	<b>18.970</b>	<b>0.5356</b>
		<b>P-ALASSO</b>	<b>1.2736</b>	<b>0.120</b>	<b>18.735</b>	<b>0.355</b>	<b>19.090</b>	<b>0.5292</b>
$\beta/2$	Ind	MCP	1.8445	0.000	12.835	0.475	13.310	0.8966
		LASSO	3.4445	0.000	13.320	5.820	19.140	1.4377
		ALASSO	1.6576	0.000	14.055	1.355	15.410	0.7825
		P-MCP	3.2507	0.000	15.375	65.940	81.315	1.4820
		<b>P-LASSO</b>	<b>1.8335</b>	<b>0.000</b>	<b>11.640</b>	<b>0.085</b>	<b>11.725</b>	<b>0.9011</b>
		<b>P-ALASSO</b>	<b>1.7610</b>	<b>0.000</b>	<b>12.245</b>	<b>0.140</b>	<b>12.385</b>	<b>0.8609</b>
	AR1(0.5)	MCP	1.9164	0.000	12.830	1.095	13.925	0.9244
		LASSO	4.8082	0.000	9.865	2.135	12.000	1.9765
		ALASSO	1.9648	0.000	11.635	0.870	12.505	0.9738
		P-MCP	3.2600	0.000	15.605	63.090	78.695	1.5025
		<b>P-LASSO</b>	<b>1.9856</b>	<b>0.000</b>	<b>10.835</b>	<b>0.185</b>	<b>11.020</b>	<b>0.9670</b>
		<b>P-ALASSO</b>	<b>1.8322</b>	<b>0.000</b>	<b>12.045</b>	<b>0.310</b>	<b>12.355</b>	<b>0.8857</b>

is better than P-MCP here.

Besides, As the correlation ( $\rho$ ) increases, the advantage of our methods increases. When  $\rho$  is large, we can imagine that some zero coefficients might be estimated as nonzero if they are highly correlated with a large nonzero coefficient, but their estimates should be quite close to 0. However, by LIC in the first stage, those coefficients are not selected, thus FP rate is much smaller than the traditional way, which also leads to the improve-

ment on other criterion.

In the simulation above, we set  $a = 3$  for MCP. Since  $a$  is another potential tuning parameter which may affect the comparison, we also tried  $a = 1.5$ . The results for all methods under  $a = 1.5$  are worse than the corresponding ones under  $a = 3$ . However, the advantage of our method is much more obvious when  $a = 1.5$ . This, in some sense, indicates that our method is more robust to the choice of  $a$ .

## 4.6.2 Real Data

We also applied our methods to the Africa Soil Property Prediction Data (available on Kaggle.com). The goal is to predict 5 target soil functional properties from diffuse reflectance infrared spectroscopy measurements. The soil functional properties are measured using conventional reference test, which is slow and expensive, and use chemicals. Diffuse reflectance infrared spectroscopy is potentially a substitute for the test. The measurement can be typically performed in about 30 seconds, the amount of light absorbed by a soil sample is measured, with minimal sample preparation, at hundreds of specific wavebands across a range of wavelengths to provide an infrared spectrum.

The data includes 1157 samples and 3594 measurements. The measurements are 3578 mid-infrared absorbance measurements and some potential spatial predictors from remote sensing data sources. And the five target properties include: Calcium, Phosphorus, pH, Sand Content, and Soil organic carbon(SOC). Therefore we are using the same design matrix to predict 5 different target variables, comparisons are made for each prediction.

For each target variable, we randomly split 1157 observations into training set (70%) and testing set (30%). we do modeling fitting and tuning parameter selection to the training set, then evaluate each methods on the testing set. The process is repeated for

50 times, and the average of MSE and Size are presented in Table 4.8. For each of the five methods in comparison, we actually tried four different criteria for parameter tuning, and only the best performance is recorded. (For MCP, the best is HBIC; while for the others, the best is AIC.)

Table 4.8: Comparison of methods - predicting 5 soil functional properties with Africa Soil Property Data

Target Variable	Measurement	MCP	LASSO	P-MCP	<b>P-ALASSO</b>	<b>P-LASSO</b>
Calcium	SIZE	9.81	66.95	42.72	29.48	134.30
	MSE	0.1836	<b>0.1476</b>	0.1649	0.1535	0.1598
Phosphorus	SIZE	2.86	24.31	29.40	28.74	30.38
	MSE	0.9303	<b>0.9107</b>	0.9309	0.9309	0.9300
pH	SIZE	10.14	84.47	51.22	36.20	170.94
	MSE	0.2284	0.2008	0.1528	0.1894	<b>0.1336</b>
SOC	SIZE	8.28	58.37	42.52	22.76	144.44
	MSE	0.2443	0.1909	0.1517	0.2031	<b>0.1277</b>
Sand	SIZE	8.66	55.82	41.46	31.42	128.30
	MSE	0.2436	0.1878	0.1287	0.1517	<b>0.1179</b>

As we can see from table 4.8, MCP always selects a much smaller model than the others, but its MSE is always the worst. Comparing the traditional methods, LASSO is averagely 20% smaller in MSE than MCP. P-LASSO usually selects a much larger set, but its MSE is the best: 35% smaller than LASSO for three target variables, and similar

to LASSO for the other two. Partial MCP and Partial ALASSO are also better than LASSO in general, but worse than P-LASSO.

The cost we pay for the smaller MSE is that the size is much larger. However, this does fit into our assumptions. It is not wise to focus on model selection if lots coefficients are not large enough, while it is wise to pick out large nonzero coefficients and focus on minimizing the errors. And that is what PPR has achieved.

To help us better understand the procedure, we also record the average number of variables selected in the first step in table 4.9. Recall that we use MCP and LIC to select large coefficients in the first step, thus the size of  $L$  should be always smaller than the size of MCP. Table 4.9 justifies our intuition, and it indicates that lots of small coefficients are included into the final model, and they do have a huge contribution for the decrease in MSE.

Table 4.9: Number of variables selected in the first step of PPR.

	Calcium	Phosphorus	pH	SOC	Sand
Size of L	9.40	0.04	8.14	5.76	7.36

## 4.7 Proofs

*Proof of Theorem 4.1.* When the selected model is the same as the true, the global minimizer  $\hat{\beta}$  is of form  $(\hat{\beta}_1, \mathbf{0})^T$ , hence we take the derivative wrt  $\beta_1$

$$\frac{\partial Q_n(\beta)}{\partial \beta_1} \Big|_{(\hat{\beta}_1^T, \mathbf{0})^T} = 0.$$

Since

$$\begin{aligned} \frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_j} \Big|_{(\hat{\boldsymbol{\beta}}_1, \mathbf{0})^T} &= -\mathbf{X}_{\cdot j}^T(\mathbf{Y}_1 - \mathbf{X}_1 \boldsymbol{\beta}_1^*) + \mathbf{X}_{\cdot j}^T \mathbf{X}_1 (\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) + n \left\{ P'_{\lambda_n}(|\beta_j^*|) \text{sgn}(\beta_j^*) \right. \\ &\quad \left. + \left[ P''_{\lambda_n}(|\beta_j^*|) + o_p(1) \right] (\hat{\beta}_j - \beta_j^*) \right\}, \quad j = 1, \dots, p_0 \end{aligned} \quad (4.34)$$

write the equation above into a matrix format, we have

$$\frac{1}{\sqrt{n}} \mathbf{X}_1^T (\mathbf{Y}_1 - \mathbf{X}_1 \boldsymbol{\beta}_1^*) + o_p(1) = \sqrt{n} \left[ (\mathbf{C}_{11,n} + \Gamma_n) (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) + \mathbf{b}_n \right].$$

It follows that

$$\sqrt{n} \mathbf{A}_n \left[ (\mathbf{C}_{11,n} + \Gamma_n) (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) + \mathbf{b}_n \right] = \frac{1}{\sqrt{n}} \mathbf{A}_n \mathbf{X}_1^T \boldsymbol{\varepsilon} + o_p(1)$$

Define  $Y_{ni} = \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_i) \varepsilon_i$ , where  $\mathbf{e}_i$  denotes the basis vector with the  $i$ -th element being 1 and all the others being 0. As we can see,  $\{Y_{ni} : 1 \leq i \leq n\}$  is a collection of independent  $q$ -dimensional random vectors with  $E(Y_{ni}) = \mathbf{0}$  and

$$\begin{aligned} \lim_{n \rightarrow \infty} \sum_{i=1}^n E(Y_{ni} Y_{ni}^T) &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{1}{n} \sigma^2 \mathbf{A}_n \mathbf{X}_1^T (\mathbf{e}_i \mathbf{e}_i^T) \mathbf{X}_1 \mathbf{A}_n^T = \lim_{n \rightarrow \infty} \frac{1}{n} \sigma^2 \mathbf{A}_n \mathbf{X}_1^T \left( \sum_{i=1}^n \mathbf{e}_i \mathbf{e}_i^T \right) \mathbf{X}_1 \mathbf{A}_n^T \\ &= \lim_{n \rightarrow \infty} \sigma^2 \mathbf{A}_n \mathbf{C}_{11,n} \mathbf{A}_n^T = \sigma^2 \mathbf{Q} \end{aligned}$$

by the multivariate Lindeberg central limit theorem and Slutsky's theorem, the multivariate normality is true if we can show for all  $\epsilon > 0$ ,

$$\sum_{i=1}^n E \|\mathbf{Y}_{ni}\|^2 \mathbf{I}(\|\mathbf{Y}_{ni}\| > \epsilon) \rightarrow 0,$$

In fact,

$$\begin{aligned}
& \sum_{i=1}^n E \|\mathbf{Y}_{ni}\|^2 \mathbf{I}(\|\mathbf{Y}_{ni}\| > \epsilon) \\
& \leq \sum_{i=1}^n \left\| \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_i) \right\|^2 E \epsilon_i^2 \mathbf{I} \left( \left\| \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_i) \right\| \cdot |\epsilon_i| > \epsilon \right) \\
& \leq \left( \sum_{i=1}^n \left\| \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_i) \right\|^2 \right) \cdot \max_{1 \leq j \leq n} E \epsilon_j^2 \mathbf{I} \left( \left\| \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_j) \right\| \cdot |\epsilon_j| > \epsilon \right)
\end{aligned}$$

Since

$$\begin{aligned}
\sum_{i=1}^n \left\| \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_i) \right\|^2 &= \text{trace} \left( \frac{1}{n} \mathbf{X}_1 \mathbf{A}_n^T \mathbf{A}_n \mathbf{X}_1^T \right) = \text{trace} (\mathbf{A}_n \mathbf{C}_{11,n} \mathbf{A}_n^T) = O(1) \\
\max_{1 \leq j \leq n} E \epsilon_j^2 \mathbf{I} \left( \left\| \frac{1}{\sqrt{n}} (\mathbf{A}_n \mathbf{X}_1^T \mathbf{e}_j) \right\| \cdot |\epsilon_j| > \epsilon \right) &\rightarrow 0, \quad \text{by Dominated Convergence Theorem}
\end{aligned}$$

we have  $\sum_{i=1}^n E \|\mathbf{Y}_{ni}\|^2 \mathbf{I}(\|\mathbf{Y}_{ni}\| > \epsilon) \rightarrow 0$ .

In conclusion,

$$\sqrt{n} \mathbf{A}_n \left[ (\mathbf{C}_{11,n} + \Gamma_n) (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) + \mathbf{b}_n \right] \rightarrow_d N(0, \sigma^2 \mathbf{Q}) \quad (4.35)$$

□

*Proof of Theorem 4.2.* First, we prove  $\tilde{\boldsymbol{\beta}}^{(0)}$  is the local minimizer:

Under (C.2), for any matrix  $\mathbf{Z}$  whose columns are a subset of the columns of  $\mathbf{X}_1$ , we have  $\rho_{\min}(\mathbf{C}_{11,n}) \leq \rho\left(\frac{\mathbf{Z}^T \mathbf{Z}}{n}\right) \leq \rho_{\max}(\mathbf{C}_{11,n})$ , where  $\rho(A)$  denotes the eigenvalue of matrix  $A$ . Therefore,  $\mathbf{C}_{11,n} \in \mathcal{U}(c_1, c_2)$  implies  $\frac{\mathbf{Z}^T \mathbf{Z}}{n} \in \mathcal{U}(c_1, c_2)$ . (We shall use this result later)

We take the partial derivative wrt  $\beta_j$ :

$$\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_j} = n P'_\lambda(|\beta_j|) \text{sgn}(\beta_j) - \mathbf{X}_j^T (\mathbf{Y} - \mathbf{X} \boldsymbol{\beta})$$

where  $\mathbf{X}_{\cdot j}$  is the  $j$ -th column of  $\mathbf{X}$ .

Define  $l_j = \frac{1}{n} \mathbf{X}_{\cdot j}^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ , based on the second-order sufficient KKT condition, any  $\tilde{\boldsymbol{\beta}}$  that satisfies

$$\begin{cases} l_j = P'_{\lambda_n}(|\tilde{\beta}_j|) \text{sgn}(\tilde{\beta}_j), & \text{if } \tilde{\beta}_j \neq 0, \\ l_j \leq P'_{\lambda_n}(0+), & \text{if } \tilde{\beta}_j = 0 \end{cases}$$

is a local minimizer of  $Q_n(\boldsymbol{\beta})$ .

For any penalty  $P_{\lambda_n}(\cdot)$  which belongs to the family, it must satisfy:  $P'_{\lambda_n}(0+) = \lambda_n$  and  $P'_{\lambda_n}(t) = 0$  when  $t \geq a\lambda_n$ . Thus, it suffices to show that

- (1)  $l_j = 0$  and  $|\tilde{\beta}_j^{(0)}| \geq a\lambda_n$  for  $j = 1, \dots, q_0$ ,
- (2)  $l_j \leq \lambda_n$  and  $|\tilde{\beta}_j^{(0)}| = 0$  for  $j = q_0 + 1, \dots, p$

We start with proving (1):

By the definition of  $\tilde{\boldsymbol{\beta}}^{(0)}$ , we have

$$l_j = \frac{1}{n} \mathbf{X}_{\cdot j}^T (\mathbf{Y} - \mathbf{Z}_1 (\mathbf{Z}_1^T \mathbf{Z}_1)^{-1} \mathbf{Z}_1^T \mathbf{Y}) = \frac{1}{n} \mathbf{X}_{\cdot j}^T (\mathbf{I} - P_{\mathbf{Z}_1}) \mathbf{Y}$$

where  $P_{\mathbf{Z}_1}$  is the projection matrix corresponding to  $\mathbf{Z}_1$ . Since  $\mathbf{X}_{\cdot j}$  is a column of  $\mathbf{Z}_1$  for  $j = 1, \dots, q_0$ ,  $l_j = 0$  for  $j = 1, \dots, q_0$ . Therefore, to prove (1), it suffices to show that, as  $n$  goes to infinity,

$$P(|\tilde{\beta}_j^{(0)}| \geq a\lambda_n \text{ for all } j, 1 \leq j \leq q_0) \rightarrow 1 \quad (4.36)$$

Let  $\tilde{\boldsymbol{\beta}}_{\mathbf{Z}_1}^{(0)}$  be the subvector of  $\tilde{\boldsymbol{\beta}}^{(0)}$  which includes the first  $q_0$  elements,  $\tilde{\boldsymbol{\beta}}_{\mathbf{Z}_2}^{(0)}$  is defined

similarly. We have

$$\begin{aligned}
\tilde{\beta}_{\mathbf{Z}_1}^{(0)} &= \frac{1}{n} \left( \frac{\mathbf{Z}_1^T \mathbf{Z}_1}{n} \right)^{-1} \mathbf{Z}_1^T (\mathbf{Z}_1 \beta_{\mathbf{Z}_1}^* + \mathbf{Z}_2 \beta_{\mathbf{Z}_2}^* + \boldsymbol{\varepsilon}) \\
&= \beta_{\mathbf{Z}_1}^* + \frac{1}{n} \left( \frac{\mathbf{Z}_1^T \mathbf{Z}_1}{n} \right)^{-1} \mathbf{Z}_1^T \mathbf{Z}_2 \beta_{\mathbf{Z}_2}^* + \frac{1}{n} \left( \frac{\mathbf{Z}_1^T \mathbf{Z}_1}{n} \right)^{-1} \mathbf{Z}_1^T \boldsymbol{\varepsilon} \\
&= V_1 + V_2 + V_3
\end{aligned} \tag{4.37}$$

By the triangle inequality, we have

$$|\tilde{\beta}_j^{(0)}| \geq |V_{1j}| - |V_{2j}| - |V_{3j}|, \quad j = 1, \dots, q_0$$

Thus we have

$$\min_{1 \leq j \leq q_0} |\tilde{\beta}_j^{(0)}| \geq \min_{1 \leq j \leq q_0} |V_{1j}| - \max_{1 \leq j \leq q_0} |V_{2j}| - \max_{1 \leq j \leq q_0} |V_{3j}|$$

By condition (C.5),

$$\min_{1 \leq j \leq q_0} |V_{1j}| > M \cdot n^{-1/2+d} \tag{4.38}$$

Next, we show that  $\max_{1 \leq j \leq q_0} |V_{3j}| = o_p(n^{-1/2+\delta})$ . For  $V_3$  we have

$$\sqrt{n}V_3 = \left( \frac{\mathbf{Z}_1^T \mathbf{Z}_1}{n} \right)^{-1} \frac{\mathbf{Z}_1^T}{\sqrt{n}} \boldsymbol{\varepsilon} = H \boldsymbol{\varepsilon},$$

and  $\sqrt{n}V_{3j} = \mathbf{e}_j^T H \boldsymbol{\varepsilon}$  where  $\mathbf{e}_j$  is the basis vector with the  $j$ -th element being 1 and all others 0's. And we have

$$\mathbf{e}_j^T H H^T \mathbf{e}_j = \mathbf{e}_j^T \left( \frac{\mathbf{Z}_1^T \mathbf{Z}_1}{n} \right)^{-1} \mathbf{e}_j \leq \rho_{\min}^{-1} \left( \frac{\mathbf{Z}_1^T \mathbf{Z}_1}{n} \right) \leq c_1^{-1} < \infty, \tag{4.39}$$

Note the fact that if  $E(\varepsilon_i)^{2k} < T_1$  and  $\|\boldsymbol{\alpha}\| \leq T_2$  where  $T_1$  and  $T_2$  are constants, then

$E(\boldsymbol{\alpha}^T \boldsymbol{\varepsilon})^{2k} = O_p(1)$ . Thus by the Chebyshev's inequality, we have

$$\begin{aligned}
P\left(\max_{1 \leq j \leq q_0} |\sqrt{n}V_{3j}| > n^d\right) &\leq \sum_{j=1}^{q_0} P(|\sqrt{n}V_{3j}| > n^d) \\
&\leq q_0 \cdot \frac{E(\sqrt{n}V_{3j})^{2k}}{n^{d \cdot 2k}} \\
&= q_0 \cdot O(n^{-2dk}) \\
&= O(n^{d-2dk}) \rightarrow 0 \text{ as } n \rightarrow \infty.
\end{aligned} \tag{4.40}$$

This implies  $\max_{1 \leq j \leq q_0} |V_{3j}| = o_p(n^{-1/2+d})$ .

At last, we prove that  $\max_{1 \leq j \leq q_0} |V_{2j}| = o(n^{-1/2+d})$  by (4.5):

$$\begin{aligned}
\max_{1 \leq j \leq q_0} |V_{2j}| &= \max_{1 \leq j \leq q_0} \left| e_j^T H \frac{\mathbf{Z}_2}{\sqrt{n}} \boldsymbol{\beta}_{\mathbf{Z}_2}^* \right| \\
&\leq \|\boldsymbol{\beta}_{\mathbf{Z}_2}^*\| \cdot \max_{1 \leq j \leq q_0} \left[ \|e_j^T H\| \rho_{\max}^{1/2} \left( \frac{\mathbf{Z}_2^T \mathbf{Z}_2}{n} \right) \right] \\
&\leq \sqrt{p_0 - q_0} \max_{q_0+1 \leq j \leq p_0} |\beta_j^*| \cdot \rho_{\max}^{1/2} \left( \frac{\mathbf{Z}_2^T \mathbf{Z}_2}{n} \right) \cdot \rho_{\min}^{-1/2} \left( \frac{\mathbf{Z}_1^T \mathbf{Z}_1}{n} \right) \\
&= o(n^{-1/2+d})
\end{aligned} \tag{4.41}$$

Combining (4.38)(4.40)(4.41), we can show that with probability tending to 1,

$$\min_{1 \leq j \leq q_0} |\tilde{\beta}_j^{(0)}| > M \cdot n^{(-1/2+d)} - o(n^{-1/2+d}) - o(n^{-1/2+d}) > a\lambda_n. \tag{4.42}$$

This finishes the proof of part (1). It remains to show that part (2) is true.

It is trivial that  $|\tilde{\beta}_j^{(0)}| = 0 \leq \lambda_n$  for  $j = q_0 + 1, \dots, p$ . Thus we only need to prove

$$P(|l_j| > \lambda_n \text{ for some } j > q_0) \rightarrow 0. \tag{4.43}$$

Expand  $l_j$ , we have

$$\begin{aligned}
l_j &= \frac{1}{n} \mathbf{X}_{\cdot j}^T (\mathbf{I} - P_{\mathbf{Z}_1}) \mathbf{Z}_2 \boldsymbol{\beta}_{\mathbf{Z}_2}^* + \frac{1}{n} \mathbf{X}_{\cdot j}^T (\mathbf{I} - P_{\mathbf{Z}_1}) \boldsymbol{\varepsilon} \\
&= S_{1j} + S_{2j}, \quad \text{for } j = q_0 + 1, \dots, p
\end{aligned} \tag{4.44}$$

We are going to bound  $\max_{j>q_0} |S_{1j}|$  and  $\max_{j>q_0} |S_{2j}|$ , respectively.

For  $S_{1j}$ , we have

$$\begin{aligned}
\max_{j>q_0} |S_{1j}| &\leq \left\| \frac{\mathbf{X}_{\cdot j}}{\sqrt{n}} \right\| \cdot \max_{j>q_0} \left[ (\mathbf{I} - P_{\mathbf{Z}_1}) \frac{\mathbf{Z}_2}{\sqrt{n}} \boldsymbol{\beta}_{\mathbf{Z}_2}^* \right] \\
&\leq 1 \cdot \|\boldsymbol{\beta}_{\mathbf{Z}_2}^*\| \cdot \rho_{\max}(\mathbf{I} - P_{\mathbf{Z}_1}) \cdot \rho_{\max}^{1/2} \left( \frac{\mathbf{Z}_2^T \mathbf{Z}_2}{n} \right) \\
&\leq \|\boldsymbol{\beta}_{\mathbf{Z}_2}^*\| \cdot \rho_{\max}^{1/2} \left( \frac{\mathbf{Z}_2^T \mathbf{Z}_2}{n} \right) = o(n^{-1/2+d})
\end{aligned} \tag{4.45}$$

While for  $S_{2j}$ , since  $\sqrt{n}S_{2j} = \frac{\mathbf{X}_{\cdot j}}{\sqrt{n}} (\mathbf{I} - P_{\mathbf{Z}_1}) \boldsymbol{\varepsilon}$ , similar to  $V_{3j}$  in part (1), we can show that  $E(\sqrt{n}S_{2j})^{2k} < \infty$ , therefore,

$$\begin{aligned}
P(\max_{j>q_0} |S_{2j}| > \lambda_n/2) &\leq \sum_{j=q_0+1}^p P(|\sqrt{n}S_{2j}| > \sqrt{n}\lambda_n/2) \\
&\leq p \cdot 2^{2k} (\sqrt{n}\lambda_n)^{-2k} E(\sqrt{n}S_{2j})^{2k} \\
&= O(p/(n\lambda_n^2)^k) \\
&\rightarrow 0
\end{aligned} \tag{4.46}$$

Therefore, we have

$$\begin{aligned}
P(|l_j| > \lambda_n, \text{ for some } j > q_0) &\leq P(|S_{1j}| + |S_{2j}| > \lambda_n \text{ for some } j > q_0) \\
&\leq P(\max_{j>q_0} |S_{1j}| + \max_{j>q_0} |S_{2j}| > \lambda_n) \\
&\leq P(\max_{j>q_0} |S_{1j}| > \lambda_n/2) + P(\max_{j>q_0} |S_{2j}| > \lambda_n/2) \quad (4.47)
\end{aligned}$$

It follows from (4.45) and (4.46) that both terms goes to 0 as  $n \rightarrow \infty$ . That is

$$P(|l_j| > \lambda_n, \text{ for some } j > q_0) \rightarrow 0 \text{ as } n \rightarrow \infty$$

This finishes the proof of part (2). □

*Proof of Lemma 4.1.* We take the partial derivative wrt  $\beta_j$ :

$$\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_j} = n \left\{ P'_\lambda(|\beta_j|) \text{sgn}(\beta_j) - \frac{\mathbf{X}_j^T \boldsymbol{\varepsilon}}{n} + \frac{\mathbf{X}_j^T \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)}{n} \right\}, \quad j = 1, \dots, p$$

where  $\mathbf{X}_j$  is the  $j$ -th column of  $\mathbf{X}$ .

we know that  $\tilde{\boldsymbol{\beta}}$  is the solution, thus it is a local minimizer, therefore it satisfies KKT condition, i.e. there exists a subgradient  $\mathbf{G} \in \partial P_{\lambda_n}(\tilde{\boldsymbol{\beta}})$  such that

$$\mathbf{G} - \frac{\mathbf{X}^T \boldsymbol{\varepsilon}}{n} + \frac{\mathbf{X}^T \mathbf{X}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{n} = \mathbf{0}$$

where the subdifferential of  $P_{\lambda_n}(\tilde{\boldsymbol{\beta}})$  is given by

$$\begin{aligned}
\partial P_{\lambda_n}(\tilde{\boldsymbol{\beta}}) &= \{\mathbf{G} = (G_1, \dots, G_p)^T \in \mathcal{R}^p : \\
&G_j = P'_\lambda(|\tilde{\beta}_j|) \text{sgn}(\tilde{\beta}_j) \text{ for } \tilde{\beta}_j \neq 0 \text{ and } G_j \in [-\lambda_n, \lambda_n] \text{ for } \tilde{\beta}_j = 0\} \quad (4.48)
\end{aligned}$$

If  $\tilde{\beta}_j = \beta_j^* = 0, \forall j > p_0$ , we have

$$\mathbf{G} - \frac{\mathbf{X}^T \boldsymbol{\varepsilon}}{n} + \frac{\mathbf{X}^T \mathbf{X}_1 (\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)}{n} = \mathbf{0} \quad (4.49)$$

The first  $p_0$  elements of (4.49) gives us that

$$n^{-1}(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) = (\mathbf{X}_1^T \mathbf{X}_1)^{-1} (n^{-1} \mathbf{X}_1^T \boldsymbol{\varepsilon} - \mathbf{G}_1) \quad (4.50)$$

While last  $p - p_0$  elements of (4.49) gives us that,  $\forall j > p_0$ ,

$$\left| \frac{\mathbf{X}_{\cdot j}^T \boldsymbol{\varepsilon}}{n} - \frac{\mathbf{X}_{\cdot j}^T \mathbf{X} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{n} \right| \leq P'_\lambda(0+) = \lambda_n \quad (4.51)$$

Therefore, from (4.50) and (4.51) we have

$$\begin{aligned} & P(\tilde{\beta}_j = 0, \forall j \geq p_0) \\ & \leq P \left( \left| \frac{\mathbf{X}_{\cdot j}^T \boldsymbol{\varepsilon}}{n} - \frac{\mathbf{X}_{\cdot j}^T \mathbf{X} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{n} \right| \leq \lambda_n, \quad \forall j > p_0 \right) \\ & \leq P \left( \left| \frac{\mathbf{X}_{\cdot j}^T \boldsymbol{\varepsilon}}{n} - \mathbf{X}_{\cdot j}^T \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} (n^{-1} \mathbf{X}_1^T \boldsymbol{\varepsilon} - \mathbf{G}_1) \right| \leq \lambda_n, \quad \forall j > p_0 \right) \\ & \leq P \left( \left| \frac{\mathbf{X}_{\cdot j}^T \boldsymbol{\varepsilon}}{n} - \mathbf{X}_{\cdot j}^T \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \frac{\mathbf{X}_1^T \boldsymbol{\varepsilon}}{n} \right| \leq \lambda_n + |\mathbf{X}_{\cdot j}^T \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{G}_1|, \quad \forall j > p_0 \right) \\ & \leq P \left( \left| \frac{\mathbf{X}_{\cdot j}^T}{n} (\mathbf{I} - \mathbf{P}_{X_1}) \boldsymbol{\varepsilon} \right| \leq \lambda_n + \|\mathbf{X}_{\cdot j}^T \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1/2}\| \cdot \|(\mathbf{X}_1^T \mathbf{X}_1)^{-1/2} \mathbf{G}_1\|, \quad \forall j > p_0 \right) \\ & \leq P \left( \left| \frac{\mathbf{X}_{\cdot j}^T}{n} (\mathbf{I} - \mathbf{P}_{X_1}) \boldsymbol{\varepsilon} \right| \leq \lambda_n + \|\mathbf{X}_{\cdot j}\| \rho_{\max}^{1/2}(\mathbf{P}_{X_1}) \cdot \frac{1}{\sqrt{n}} \rho_{\min}^{-1/2} \left( \frac{\mathbf{X}_1^T \mathbf{X}_1}{n} \right) \|\mathbf{G}_1\|, \quad \forall j > p_0 \right) \\ & \leq P \left( \left| \frac{\mathbf{X}_{\cdot j}^T}{n} (\mathbf{I} - \mathbf{P}_{X_1}) \boldsymbol{\varepsilon} \right| \leq \lambda_n + \lambda_n \cdot \rho_{\min}^{-1/2} \left( \frac{\mathbf{X}_1^T \mathbf{X}_1}{n} \right), \quad \forall j > p_0 \right) \end{aligned} \quad (4.52)$$

If  $\lambda_n = O(\frac{1}{\sqrt{n}})$ , then the RHS of (4.52) is on the order  $O(\frac{1}{\sqrt{n}})$ . While the LHS is a linear combination of *i.i.d* errors, and it is also on the order  $O(\frac{1}{\sqrt{n}})$ . Therefore, there exists  $0 < \alpha < 1$ , such that (4.52) is bounded above by  $C$ . i.e.

$$\limsup_{n \rightarrow \infty} P(\tilde{\beta}_j = 0, \forall j \geq p_0 + 1) \leq C.$$

□

*Proof of theorem 4.3.* Let the index set of true model be  $A_0 = \{j : 1 \leq j \leq p_0\}$ , and the index of large coefficients be  $A_1 = \{j : 1 \leq j \leq q_0\}$ , and the index set of the model selected by  $\lambda$  be  $A_\lambda = \{j : 1 \leq j \leq p, \hat{\beta}_j(\lambda) \neq 0\}$ . We define two sets of models as following:

$$\begin{aligned} \Lambda_{1-} &= \{\lambda : \lambda \geq 0, A_1 \not\subset A_\lambda\}, \\ \Lambda_{1+} &= \{\lambda : \lambda \geq 0, A_1 \subset A_\lambda, A_\lambda \neq A_1\}. \end{aligned}$$

If we consider the target set to be  $A_1$  (instead of  $A_0$ ), then these two sets of models are paralleled to the underfitted and the overfitted sets.

For a given model  $M$  (the index set of selected variables), define

$$\hat{\sigma}_M^2 = n^{-1} \|\mathbf{Y} - \mathbf{X}_M \hat{\boldsymbol{\beta}}_{M,ols}\|^2 \quad \text{where} \quad \hat{\boldsymbol{\beta}}_{M,ols} = (\mathbf{X}_M^T \mathbf{X}_M)^{-1} \mathbf{X}_M^T \mathbf{Y}$$

where  $\mathbf{X}_M$  denotes the submatrix of  $\mathbf{X}$  that corresponds to  $M$ . Recall that  $\hat{\sigma}_\lambda^2 = n^{-1} \|\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}(\lambda)\|^2$ , by the definition, we have  $\hat{\sigma}_{A_\lambda}^2 \leq \hat{\sigma}_\lambda^2$ .

Let  $\lambda_n$  be the tuning parameter that satisfies the conditions of theorem 4.2. We shall prove the following two cases:

Case I  $P(\inf_{\lambda \in \Lambda_{1-}} [\text{LIC}(\lambda) - \text{LIC}(\lambda_n)] > 0) \rightarrow 1,$

Case II  $P(\inf_{\lambda \in \Lambda_{1+}} [\text{LIC}(\lambda) - \text{LIC}(\lambda_n)] > 0) \rightarrow 1.$

Case I: By theorem 4.2, we know that  $P(\hat{\boldsymbol{\beta}}_{\lambda_n} = \tilde{\boldsymbol{\beta}}^{(0)}) \rightarrow 1,$  thus we have

$$\begin{aligned}
& P(\inf_{\lambda \in \Lambda_{1-}} [\text{LIC}(\lambda) - \text{LIC}(\lambda_n)] > 0) \\
&= P(\inf_{\lambda \in \Lambda_{1-}} [\text{LIC}(\lambda) - \text{HIC}(\lambda_n)] > 0, \hat{\boldsymbol{\beta}}(\lambda_n) = \tilde{\boldsymbol{\beta}}^{(0)}) \\
&\quad + P(\inf_{\lambda \in \Lambda_{1-}} [\text{LIC}(\lambda) - \text{LIC}(\lambda_n)] > 0, \hat{\boldsymbol{\beta}}(\lambda_n) \neq \tilde{\boldsymbol{\beta}}^{(0)}) \\
&= P\left(\inf_{\lambda \in \Lambda_{1-}} \left[ \log \frac{\hat{\sigma}_{A\lambda}^2}{\hat{\sigma}_{A_1}^2} + \frac{C_n}{n} \left( \sum_{j=1}^p L(\hat{\beta}_j(\lambda)) \right) - q_0 \right] > 0\right) + o(1) \\
&\geq P\left(\inf_{\lambda \in \Lambda_{1-}} \left[ \log \frac{\hat{\sigma}_{A\lambda}^2}{\hat{\sigma}_{A_1}^2} + \frac{C_n}{n} \left( \sum_{j=1}^p L(\hat{\beta}_j(\lambda)) \right) - q_0 \right] > 0\right) + o(1) \\
&\geq P\left(\inf_{\lambda \in \Lambda_{1-}} \left[ \log \frac{\hat{\sigma}_{A\lambda}^2}{\hat{\sigma}_{A_1}^2} - q_0 \frac{C_n}{n} \right] > 0\right) + o(1) \tag{4.53}
\end{aligned}$$

We are going to consider the two terms:  $\hat{\sigma}_{A\lambda}^2$  and  $\hat{\sigma}_{A_1}^2$ .

Since  $\mathbf{Y} = \mathbf{Z}_1 \boldsymbol{\beta}_{(1)} + \mathbf{Z}_2 \boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}$ , we have

$$\begin{aligned}
\hat{\sigma}_{A_1}^2 &= n^{-1} \mathbf{Y} (\mathbf{I} - \mathbf{P}_{Z_1}) \mathbf{Y} \\
&= n^{-1} (\mathbf{Z}_2 \boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon})^T (\mathbf{I} - \mathbf{P}_{Z_1}) (\mathbf{Z}_2 \boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}) \\
&= n^{-1} [\boldsymbol{\beta}_{(2)}^T \mathbf{Z}_2^T (\mathbf{I} - \mathbf{P}_{Z_1}) \mathbf{Z}_2 \boldsymbol{\beta}_{(2)} + 2 \boldsymbol{\beta}_{(2)}^T \mathbf{Z}_2^T (\mathbf{I} - \mathbf{P}_{Z_1}) \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T (\mathbf{I} - \mathbf{P}_{Z_1}) \boldsymbol{\varepsilon}] \tag{4.54}
\end{aligned}$$

Let  $\rho_{\max}(M)$  be the largest eigenvalue of matrix  $M$ , by condition (C.2), we have

$$\rho_{\max} \left( \frac{\mathbf{Z}_2^T \mathbf{Z}_2}{n} \right) \leq \rho_{\max}(\mathbf{C}_{11,n}) \leq c_2,$$

therefore, by (C.5),

$$\|\mathbf{Z}_2\boldsymbol{\beta}_{(2)}\| \leq \|\boldsymbol{\beta}_{(2)}\| \cdot \rho_{\max}^{1/2} \left( \frac{\mathbf{Z}_2^T \mathbf{Z}_2}{n} \right) \cdot n^{1/2} = o(n^d).$$

Note that the largest eigenvalue of  $(\mathbf{I} - \mathbf{P}_{Z_1})$  is 1, it follows that

$$\begin{aligned} \hat{\sigma}_{A_1}^2 &= n^{-1} [\boldsymbol{\beta}_{(2)}^T \mathbf{Z}_2^T (\mathbf{I} - \mathbf{P}_{Z_1}) \mathbf{Z}_2 \boldsymbol{\beta}_{(2)} + 2\boldsymbol{\beta}_{(2)}^T \mathbf{Z}_2^T (\mathbf{I} - \mathbf{P}_{Z_1}) \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T (\mathbf{I} - \mathbf{P}_{Z_1}) \boldsymbol{\varepsilon}] \\ &\leq n^{-1} \|\mathbf{Z}_2 \boldsymbol{\beta}_{(2)}\|^2 + 2n^{-1} \|\mathbf{Z}_2 \boldsymbol{\beta}_{(2)}\| \cdot \|\boldsymbol{\varepsilon}\| + n^{-1} \boldsymbol{\varepsilon}^T (\mathbf{I} - \mathbf{P}_{Z_1}) \boldsymbol{\varepsilon} \\ &= o(n^{-1+2d}) + o(n^{-1/2+d}) + n^{-1} \boldsymbol{\varepsilon}^T (\mathbf{I} - \mathbf{P}_{Z_1}) \boldsymbol{\varepsilon} \end{aligned} \quad (4.55)$$

The last term converges to  $\sigma^2$ , thus it is the dominant term. That is,  $\hat{\sigma}_{A_1}^2 \rightarrow \sigma^2$ .

Now for  $\hat{\sigma}_{A_\lambda}^2$ , let  $\boldsymbol{\mu} = \mathbf{Z}_1 \boldsymbol{\beta}_{(1)} + \mathbf{Z}_2 \boldsymbol{\beta}_{(2)}$ , we have

$$\begin{aligned} \hat{\sigma}_{A_\lambda}^2 &= n^{-1} \mathbf{Y} (\mathbf{I} - \mathbf{P}_{A_\lambda}) \mathbf{Y} \\ &= n^{-1} [\boldsymbol{\mu}^T (\mathbf{I} - \mathbf{P}_{A_\lambda}) \boldsymbol{\mu} + 2\boldsymbol{\mu}^T (\mathbf{I} - \mathbf{P}_{A_\lambda}) \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T (\mathbf{I} - \mathbf{P}_{A_\lambda}) \boldsymbol{\varepsilon}] \\ &\geq n^{-1} \boldsymbol{\mu}^T (\mathbf{I} - \mathbf{P}_{A_\lambda}) \boldsymbol{\mu} + 2n^{-1} \boldsymbol{\mu}^T (\mathbf{I} - \mathbf{P}_{A_\lambda}) \boldsymbol{\varepsilon} \\ &= n^{-1} (I_1 + I_2) \end{aligned} \quad (4.56)$$

By condition (C.6),  $\inf_{\lambda \in \Lambda_{1-}} |I_1| \geq \delta n$  for all sufficient large  $n$ . And we are going to show that  $\sup_{\lambda \in \Lambda_{1-}} |I_2| = o(I_1)$ . In fact, define  $\boldsymbol{\alpha}_\lambda^T = I_1^{-1/2} \boldsymbol{\mu}^T (\mathbf{I} - \mathbf{P}_{A_\lambda})$ , then  $\|\boldsymbol{\alpha}_\lambda\| = 1$ , and  $I_2 = 2\sqrt{I_1} \boldsymbol{\alpha}_\lambda^T \boldsymbol{\varepsilon}$ . By the sub-Gaussian tail inequality, we have

$$P \left( |\boldsymbol{\alpha}_\lambda^T \boldsymbol{\varepsilon}| > \sqrt{n/\log n} \right) \leq 2 \exp \left( -\frac{n}{2\sigma^2 \log n} \right).$$

The number of models by  $\Lambda_{1-}$  is at most  $\sum_{i=1}^{K_n} \binom{p}{i} \leq \sum_{i=1}^{K_n} p^i \leq 2p^{K_n}$ , thus we have

$$P \left( \sup_{\lambda \in \Lambda_{1-}} |\boldsymbol{\alpha}_\lambda^T \boldsymbol{\varepsilon}| > \sqrt{n/\log n} \right) \leq 4p^{K_n} \exp \left( -\frac{n}{2\sigma^2 \log n} \right) \rightarrow 0$$

as  $K_n \log p \log n = o(n)$ . Thus,

$$\sup_{\lambda \in \Lambda_{1-}} |I_2| = 2\sqrt{I_1} \sup_{\lambda \in \Lambda_{1-}} |\boldsymbol{\alpha}_\lambda^T \boldsymbol{\varepsilon}| = o(\sqrt{I_1} \sqrt{n/\log n}) = o(I_1)$$

That means,  $I_1$  is the dominant term in (4.56), thus,  $\inf_{\lambda \in \Lambda_{1-}} \hat{\sigma}_{A_\lambda}^2 \geq \delta$ .

Now let us go back to (4.53), with  $\inf_{\lambda \in \Lambda_{1-}} \hat{\sigma}_{A_\lambda}^2 \geq \delta$  and  $\hat{\sigma}_{A_1}^2 \rightarrow \sigma^2$ , since  $p_0 C_n = o(n)$ , we have

$$\begin{aligned} & P \left( \inf_{\lambda \in \Lambda_{1-}} [\text{LIC}(\lambda) - \text{LIC}(\lambda_n)] > 0 \right) \\ & \geq P \left( \inf_{\lambda \in \Lambda_{1-}} \left[ \log \frac{\hat{\sigma}_{A_\lambda}^2}{\hat{\sigma}_{A_1}^2} - q_0 \frac{C_n}{n} \right] \right) + o(1) \geq P \left( \log \frac{\delta}{\sigma^2} - \frac{q_0 C_n}{n} \right) \rightarrow 1 \end{aligned}$$

This finishes the proof of case I.

Case II Similar to (4.53), we have

$$\begin{aligned} & P \left( \inf_{\lambda \in \Lambda_{1+}} [\text{LIC}(\lambda) - \text{LIC}(\lambda_n)] > 0 \right) \\ & = P \left( \inf_{\lambda \in \Lambda_{1+}} [\text{LIC}(\lambda) - \text{HIC}(\lambda_n)] > 0, \hat{\boldsymbol{\beta}}(\lambda_n) = \tilde{\boldsymbol{\beta}}^{(0)} \right) \\ & \quad + P \left( \inf_{\lambda \in \Lambda_{1+}} [\text{LIC}(\lambda) - \text{LIC}(\lambda_n)] > 0, \hat{\boldsymbol{\beta}}(\lambda_n) \neq \tilde{\boldsymbol{\beta}}^{(0)} \right) \\ & \geq P \left( \inf_{\lambda \in \Lambda_{1+}} \left[ -\log \frac{\hat{\sigma}_{A_1}^2}{\hat{\sigma}_\lambda^2} + \frac{C_n}{n} \left( \sum_{j=1}^p L(\hat{\beta}_j(\lambda)) - q_0 \right) \right] > 0 \right) + o(1) \\ & \geq P \left( \inf_{\lambda \in \Lambda_{1+}} \left[ -\log \frac{\hat{\sigma}_{A_1}^2}{\hat{\sigma}_\lambda^2} + \frac{C_n}{n} (|A_\lambda| - q_0) \right] > 0 \right) + o(1) \end{aligned} \tag{4.57}$$

Now let us consider the term  $-\log \frac{\hat{\sigma}_{A_1}^2}{\hat{\sigma}_{A_\lambda}^2}$ . Applying  $\log(1+x) \leq x, \forall x \geq 0$ , we have

$$\begin{aligned}
\log \left( \frac{\hat{\sigma}_{A_1}^2}{\hat{\sigma}_{A_\lambda}^2} \right) &= \log \left( 1 + \frac{\mathbf{Y}^T(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})\mathbf{Y}}{\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{A_\lambda})\mathbf{Y}} \right) \\
&\leq \frac{\mathbf{Y}^T(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})\mathbf{Y}}{\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{A_\lambda})\mathbf{Y}} \\
&= \frac{\mathbf{Y}^T(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})\mathbf{Y}}{\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{A_1})\mathbf{Y} - \mathbf{Y}^T(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})\mathbf{Y}} \\
&= \frac{\mathbf{Y}^T(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})\mathbf{Y}}{n\hat{\sigma}_{A_1}^2 - \mathbf{Y}^T(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})\mathbf{Y}} \tag{4.58}
\end{aligned}$$

In case I, we have shown  $\hat{\sigma}_{A_1}^2 \rightarrow \sigma^2$ ; while for  $\mathbf{Y}^T(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})\mathbf{Y}$ , we substitute  $\mathbf{Y}$  by  $\mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \mathbf{Z}_2\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}$ ,

$$\begin{aligned}
&\mathbf{Y}^T(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})\mathbf{Y} \\
&= (\mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \mathbf{Z}_2\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon})(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})(\mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \mathbf{Z}_2\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}) \\
&= (\mathbf{Z}_2\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon})(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})(\mathbf{Z}_2\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}) \\
&= \boldsymbol{\beta}_{(2)}^T \mathbf{Z}_2^T (\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1}) \mathbf{Z}_2 \boldsymbol{\beta}_{(2)} + 2\boldsymbol{\beta}_{(2)}^T \mathbf{Z}_2^T (\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1}) \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T (\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1}) \boldsymbol{\varepsilon} \\
&\leq 2 [\boldsymbol{\beta}_{(2)}^T \mathbf{Z}_2^T (\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1}) \mathbf{Z}_2 \boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}^T (\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1}) \boldsymbol{\varepsilon}] \\
&= 2(J_1 + J_2) \tag{4.59}
\end{aligned}$$

The second equality is because  $A_1 \subset A_\lambda$  for all  $\lambda \in \Lambda_{1+}$ , and  $\mathbf{P}_{A_\lambda} \mathbf{Z}_1 = \mathbf{P}_{Z_1} \mathbf{Z}_1 = \mathbf{Z}_1$ .

For  $J_1$ , since  $(\mathbf{P}_{A_\lambda} - \mathbf{P}_{Z_1})$  is idempotent,

$$J_1 \leq \|\mathbf{Z}_2\boldsymbol{\beta}_{(2)}\|^2 \leq \|\boldsymbol{\beta}_{(2)}\|^2 \cdot \rho_{\max} \left( \frac{\mathbf{Z}_2^T \mathbf{Z}_2}{n} \right) \cdot n = O(\log p)$$

For  $J_2$ , we apply Theorem 1 in Daniel Hsu and Zhang (2012): for the subgaussian errors  $\boldsymbol{\varepsilon}$  with mean  $\mathbf{0}$  and variance  $\sigma^2 \mathbf{I}_n$ , the quadratic form satisfies ( $\|\cdot\|$  denotes the

spectral norm)

$$P \left\{ \boldsymbol{\varepsilon}^T \Sigma \boldsymbol{\varepsilon} > \sigma^2 \left( \text{tr}(\Sigma) + 2\sqrt{\text{tr}(\Sigma^2)t} + 2\|\Sigma\|t \right) \right\} \leq e^{-t} \quad (4.60)$$

If  $\Sigma$  is symmetric and idempotent, and  $\text{rank}(\Sigma) = K$ , then  $\text{tr}(\Sigma) = \text{tr}(\Sigma^2) = K$  and  $\|\Sigma\| \leq K^{1/2}$ . We apply the inequality (4.60), and let  $t > K$ , we shall have

$$\begin{aligned} & P \left\{ \boldsymbol{\varepsilon}^T \Sigma \boldsymbol{\varepsilon} > 3\sigma^2 K t \right\} \\ & \leq P \left\{ \boldsymbol{\varepsilon}^T \Sigma \boldsymbol{\varepsilon} > 3\sigma^2 \sqrt{K} t \right\} \\ & < P \left\{ \boldsymbol{\varepsilon}^T \Sigma \boldsymbol{\varepsilon} > \sigma^2 \left( \text{tr}(\Sigma) + 2\sqrt{\text{tr}(\Sigma^2)t} + 2\|\Sigma\|t \right) \right\} \\ & \leq e^{-t} \end{aligned} \quad (4.61)$$

let  $t = 2K_n \log p$ , since the projection matrices are always symmetric and idempotent, we apply (4.61),

$$\begin{aligned} & P \left( \sup_{\lambda \in \Lambda_{1+}} \boldsymbol{\varepsilon}^T (\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1}) \boldsymbol{\varepsilon} > 3\sigma^2 K_n^{1/2} t \right) \\ & \leq P \left( \sup_{\lambda \in \Lambda_{1+}} \boldsymbol{\varepsilon}^T \mathbf{P}_{A_\lambda} \boldsymbol{\varepsilon} > 3\sigma^2 K_n^{1/2} t \right) \\ & \leq P \left( \sup_{A \in \{M: |M|=K_n, M \subset \{1, \dots, p\}\}} \boldsymbol{\varepsilon}^T \mathbf{P}_A \boldsymbol{\varepsilon} > 3\sigma^2 K_n^{1/2} t \right) \\ & \leq \binom{p}{K_n} \cdot P(\boldsymbol{\varepsilon}^T \mathbf{P}_A \boldsymbol{\varepsilon} > 3\sigma^2 K_n^{1/2} t) \\ & \leq p^{K_n} e^{-2K_n \log p} \rightarrow 0 \end{aligned} \quad (4.62)$$

the second inequality is because:  $\forall A_\lambda, \lambda \in \Lambda_{1+}$ ,  $\exists$  an index set  $A$  such that  $|A| = K_n$  and  $A_\lambda \subset A$ , and we always have  $\boldsymbol{\varepsilon}^T \mathbf{P}_{A_\lambda} \boldsymbol{\varepsilon} \leq \boldsymbol{\varepsilon}^T \mathbf{P}_A \boldsymbol{\varepsilon}$ .

Therefore,  $J_2 = o(K_n^{3/2} \log p) = o(n)$ . Also remember that  $J_1 = O(\log p)$ , thus by

(4.59), we have

$$\mathbf{Y}^T(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})\mathbf{Y} = o(n) \quad (4.63)$$

Therefore, the denominator of (4.58) converges to  $n\sigma^2$ .

Besides, by the proof of theorem 3.5 in Wang et al. (2013), we can show that

$$P \left\{ \sup_{\lambda \in \Lambda_{1+}} [\boldsymbol{\varepsilon}^T(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})\boldsymbol{\varepsilon} / (|A_\lambda| - q_0)] > 16\sigma^2 \log p \right\} \rightarrow 0. \quad (4.64)$$

Also remember that  $J_1 = O(\log p)$ , thus by (4.59), we have

$$\sup_{\lambda \in \Lambda_{1+}} \mathbf{Y}^T(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})\mathbf{Y} / (|A_\lambda| - q_0) = O(\log p) \quad (4.65)$$

By (4.58) and (4.65), and following (4.57), we have

$$\begin{aligned} & P(\inf_{\lambda \in \Lambda_{1+}} [\text{LIC}(\lambda) - \text{LIC}(\lambda_n)] > 0) \\ & \geq P \left( \inf_{\lambda \in \Lambda_{1+}} \left[ -\log \frac{\hat{\sigma}_{A_1}^2}{\hat{\sigma}_\lambda^2} + \frac{C_n}{n} (|A_\lambda| - q_0) \right] > 0 \right) + o(1) \\ & \geq P \left( \frac{C_n}{n} - \sup_{\lambda \in \Lambda_{1+}} \left[ \log \frac{\hat{\sigma}_{A_1}^2}{\hat{\sigma}_\lambda^2} / (|A_\lambda| - q_0) \right] > 0 \right) + o(1) \\ & \geq P \left( \frac{C_n}{n} - \sup_{\lambda \in \Lambda_{1+}} \left[ \frac{\mathbf{Y}^T(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})\mathbf{Y} / (|A_\lambda| - q_0)}{n\hat{\sigma}_{A_1}^2 - \mathbf{Y}^T(\mathbf{P}_{A_\lambda} - \mathbf{P}_{A_1})\mathbf{Y}} \right] > 0 \right) + o(1) \\ & \geq P \left( \frac{C_n}{n} - \frac{O(\log p)}{n\sigma^2} > 0 \right) \rightarrow 1. \end{aligned} \quad (4.66)$$

This finishes the proof of case II. Together with case I, we have finished the proof of theorem 4.3.  $\square$

# Chapter 5

## Finite Variance Oracle Property

### 5.1 Introduction

Consider the regression model

$$y_i = \mathbf{x}_i' \beta + \epsilon_i, \quad i = 1, \dots, n, \quad (5.1)$$

where,  $y_i$  is the response,  $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})'$  is a nonrandom covariate vector,  $\beta = (\beta_1, \dots, \beta_p)'$  is the regression parameter, and  $\{\epsilon_i : i = 1, \dots, n\}$  are iid error variables with  $E\epsilon_1 = 0$  and  $E\epsilon_1^2 = \sigma^2 \in (0, \infty)$ . Here and in the following,  $A'$  denotes the transpose of a matrix  $A$ . We shall suppose that the dimension  $p = p_n$  of the regression model diverges with the sample size  $n$ , and that only a subset of the regression parameters are nonzero. Identification of the relevant covariates (corresponding to nonzero  $\beta_j$ s) and estimation of the nonzero regression parameters  $\beta_j$ s are of much importance in many applications, particularly when the dimension  $p$  of the full regression vector is very large. Several penalized regression methods have been proposed in the literature to address

these, such as the LASSO (Tibshirani, 1994), SCAD (Fan and Li, 2001), ALASSO (Zou, 2006), MCP (Zhang, 2010), among others. Consistent variable selection and asymptotic distribution of the resulting estimators have been proved under various sets of regularity conditions. See, for example, Bühlmann and van de Geer (2011), Bickel et al. (2009), Knight and Fu (2000), Meinshausen and Bühlmann (2006), Huang et al. (2008), Shen et al. (2012), Zhang (2010), among others. While consistent variable selection property of a method provides dimension reduction and provides a simpler model to work with, the asymptotic distributional results allow one to construct large sample statistical inference methods, often with nearly the same accuracy as that of the standard methods under the reduced true model. This is referred to as the *Oracle Property* by Fan and Li (2001) which plays an important role in high dimensional regression problems. However, the existing results require very strong conditions on the tails of the error distribution (e.g., exponential decay) when the dimension of the full regression parameter vector grows fast. Our goal here is to investigate validity of the Oracle Property in high dimensions under weak or near-minimal moment conditions.

To gain some insight into the effects of the tails of the error distribution, consider the following numerical example with 4 different choices of  $\epsilon_1$  in (5.1), given by

$$(a) \epsilon_1 \sim N(0, 1), (b) \epsilon_1 \sim \chi^2(1) - 1, (c) \epsilon_1 \sim \text{Pareto}(10) \text{ and } (d) \epsilon_1 \sim \text{Pareto}(3),$$

where for  $\epsilon_1 \sim \text{Pareto}(\alpha)$ , the tail probability  $P(|\epsilon_1| > x)$  decays at the rate  $x^{-\alpha}$  as  $x \rightarrow \infty$ . Thus, the distribution of  $\epsilon_1$  has Gaussian tails in (a), exponential tails (with a finite moment generating function) in (b), and only finite absolute moments of orders less than 10 and 3 in (c) and (d), respectively.

Figure 5.1 gives the probabilities of selecting the true model by the MCP method of Zhang (2010) for different combinations of  $(n, p)$  under the moment conditions (a)-(d).

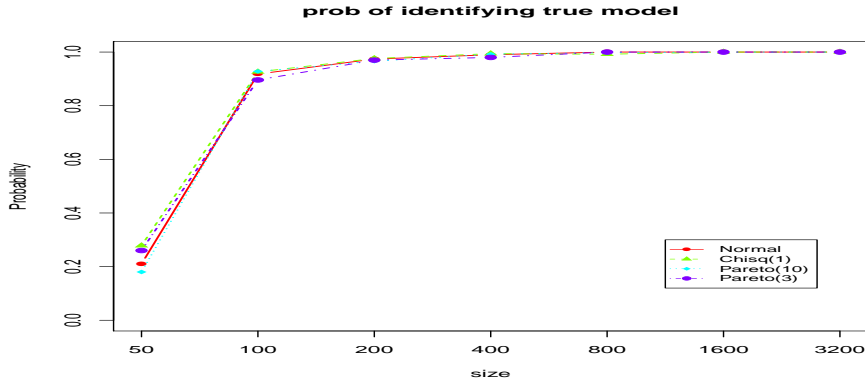


Figure 5.1: *Empirical probability of selecting the true model using MCP under error distributions (a)-(d)*

(See Section 4 for more details on the simulation set up.) The same values are reported in Table 1.

It appears from Figure 5.1 that the probability of selecting the true model is not very sensitive to the actual rate of decay of the tails of the error distribution in (a)-(d). We have observed similar patterns in a more extensive simulation study involving other combinations of  $(n, p)$  and other error distributions (results not reported here) that tends to suggest that existence of the exponential tails is not necessary for good variable selection property even in very high dimensional cases. A similar conclusion holds on the asymptotic normality of the MCP (and related) estimators of the nonzero components. Figure 5.2 gives normal Q-Q plots of the MCP estimates of the first nonzero component under two different error distributions (viz. cases (a) and (d)) for sample sizes  $n = 100, 1600$ . It is clear that the Q-Q plots in each column (respectively, for  $\epsilon_1 \sim N(0, 1)$  and  $\epsilon_1 \sim \text{Pareto}(3)$ ) show qualitatively similar patterns, with accuracy of normal approximation improving with the higher sample size. This suggests that the validity of the Oracle Property (i.e., both variable selection consistency and asymptotic

normality) should perhaps hold under much weaker moment conditions than the sufficient conditions given in the literature, for a wide range of growth rates of the dimension  $p$  as a function of  $n$ , and our theoretical results corroborate this.

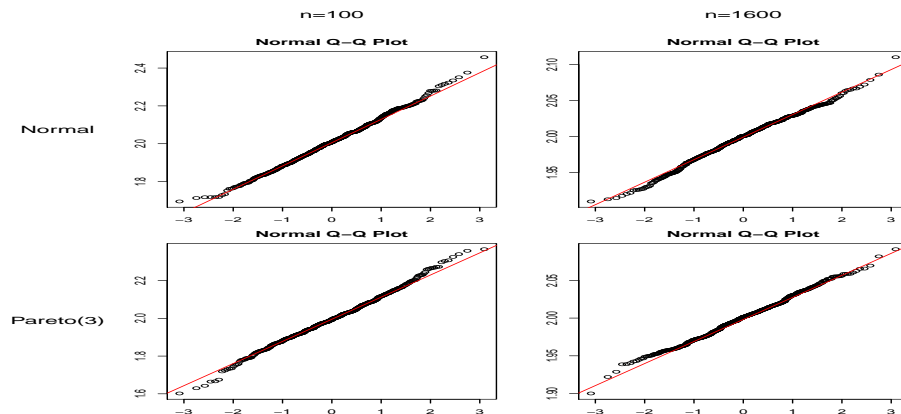


Figure 5.2: *Normal Q-Q plots of MCP estimates of a nonzero component under error distributions (a) and (d) for sample sizes  $n = 100, 1600$ , based on 500 simulation runs.*

To describe our main results and put it in the right context, let  $I_n \equiv \{j : 1 \leq j \leq p, \beta_j \neq 0\}$  denote the indices of the relevant covariates. For notation simplicity, without loss of generality (w.l.o.g.), suppose that  $I_n = \{1, \dots, p_0\}$  for some  $p_0 = p_{0n} \in (0, p]$ . We shall restrict attention to the case where  $p_0 \leq n$  and require that the ordinary least squares (OLS) estimator  $\hat{\beta}_n$  of the vector of nonzero regression parameters  $\beta^{(1)} \equiv (\beta_j : 1 \leq j \leq p_0)'$  under the reduced model  $\beta_{p_0+1} = \dots = \beta_p = 0$  be *uniquely* defined. For  $j = 1, \dots, p$ , let  $\hat{\beta}_{n,j}$  denote the  $j$ th component of a penalized regression estimator  $\hat{\beta}_n$  of the full regression parameter vector  $\beta$ . Define the set of selected variables  $\hat{I}_n = \{j : 1 \leq j \leq p, \hat{\beta}_{n,j} \neq 0\}$ . Then, variable selection consistency of the penalized regression method

is equivalent to requiring that

$$P(\hat{I}_n = I_n) \rightarrow 1 \quad \text{as } n \rightarrow \infty. \quad (5.2)$$

Also, a penalized regression method is said to have the Oracle Property if it satisfies (5.2) and if for any well-behaved (to be made precise later)  $q \times p$  matrix  $D_n$

$$D_n(\hat{\beta}_n - \beta) \rightarrow^d N_k(\mathbf{0}, \Sigma), \quad (5.3)$$

whenever  $D_n([\hat{\beta}_n - \beta^{(1)}]': \mathbf{0})' \rightarrow^d N_k(\mathbf{0}, \Sigma)$ . Here and elsewhere in the paper,  $\rightarrow^d$  denotes convergence in distribution,  $N_k(\boldsymbol{\mu}; \Sigma)$  denotes the normal distribution on  $\mathbb{R}^k$  with mean  $\boldsymbol{\mu}$  and variance  $\Sigma$ , and  $\mathbf{0}$  denotes a matrix (vector) of zeros of appropriate dimensions. Further,  $q$  does not depend on  $n$ . Note that condition (5.3) is equivalent to saying that any finite set of linear combinations of the penalized regression estimator  $\hat{\beta}_n$  has the same limit distribution as that of the Oracle OLS estimator  $(\hat{\beta}_n': \mathbf{0})'$  under the Oracle  $\beta_{p_0+1} = \dots = \beta_p = 0$ . Fan and Li (2001) first established the Oracle Property for the SCAD for a finite  $p$ , which was later extended to the increasing  $p_n$  case by Kim et al. (2008). Kim et al. (2008) allows  $p = O(n^\alpha)$ ,  $\alpha > 0$  if the noises satisfy  $E(\varepsilon_1^{2k}) < \infty$  where  $k$  is an integer; or  $p = O(\exp(cn))$ ,  $c > 0$  if noises are Gaussian distributed. Zou (2006) introduced the ALASSO method and also established its Oracle Property for a finite  $p$ . Available results on Oracle Property of the ALASSO method in the increasing  $p$  case have been obtained by Chattergee and Lahiri (2013), where they require  $E(|\varepsilon_1|^r) < \infty$ ,  $r > 3$  and  $p = O(n^\alpha)$  where  $\alpha$  depends  $r$ . Zhang (2010) introduced the MCP and established its variable selection consistency (and  $\ell_2$ -error bounds) for Gaussian errors, allowing  $\log p = o(n)$ . In this paper, we develop a way of proving the Oracle Property of penalized

regression methods under weak moment conditions. As a concrete example, we consider the ALASSO method and show that for a wide range of high dimensionality, allowing

$$\log p = O(n^\alpha) \text{ for some } \alpha \in (0, 1/2), \quad (5.4)$$

the Oracle Property holds for these methods just requiring finiteness of the second moment of  $\epsilon_1$ . In particular, for any high dimensional regression model where the full model dimension  $p$  grows at an *arbitrary* polynomial rate with the sample size, the Oracle Property holds under  $E\epsilon_1^2 < \infty$ . Since the polynomial growth rate of  $p$  is adequate in many applications, this result shows that the penalized regression methods can be employed in a wide range of high dimensional problems essentially under the same minimal moment conditions that is needed for the asymptotic normality of the OLS in a finite dimensional regression model.

The key tool used for proving the main results of the paper is a maximal inequality which could be potentially useful in proving similar theoretical properties of penalized regression methods. In addition to proving the finite variance case results, we also use the inequality to establish the validity of the Oracle Property beyond the growth rate (5.4). We show that under existence of suitable higher order absolute moments, the Oracle Property holds even when

$$\log p = O(n^\alpha) \text{ for some } \alpha \in (0, 1), \quad (5.5)$$

This may be compared with the existing results in the literature where exponential or Gaussian tails are often assumed.

The rest of the paper is organized as follows. In Sections 5.2, we prove the results for

ALASSO. In Section 5.3, we present the maximal inequality and some of its ramifications. Section 5.4 gives results from a moderately large simulation study that indicates that the Oracle Property holds quite generally under weak moment conditions for popular penalized regression methods including the ALASSO, SCAD and MCP. Proofs of the main results are given in Section 5.5.

## 5.2 Results on ALASSO

### 5.2.1 The ALASSO Method

In this section, we establish the Oracle Property of the ALASSO method of Zou (2006) under weak moment conditions. The ALASSO estimator of  $\beta$  in model (5.1) is defined as (cf. Zou (2006)):

$$\widehat{\beta}_n = \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^p} \sum_{i=1}^n (y_i - \mathbf{x}_i' \mathbf{u})^2 + \lambda_n \sum_{j=1}^p \frac{|u_j|}{\widetilde{\beta}_{n,j}^\gamma} \quad (5.6)$$

where  $\widetilde{\beta}_{n,j}$  is the  $j$ th component of a preliminary estimator  $\widetilde{\beta}_n$  of  $\beta$ ,  $\lambda_n \in (0, \infty)$  is a penalty parameter and  $\gamma \in (0, \infty)$  is a tuning parameter. When  $p \leq n$  and the design matrix is of full rank, a common choice of  $\widetilde{\beta}_n$  is given by the OLS estimator of  $\beta$ . For  $p > n$ , the OLS is not uniquely defined. In such cases, a popular choice of  $\widetilde{\beta}_n$  is given by the LASSO estimator of  $\beta$  (cf. Tibshirani (1994)):

$$\widetilde{\beta}_n = \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^p} \sum_{i=1}^n (y_i - \mathbf{x}_i' \mathbf{u})^2 + \lambda_n \sum_{j=1}^p |u_j|. \quad (5.7)$$

Note that the criteria functions of both ALASSO and the LASSO methods are convex and hence, the corresponding estimators can be found by convex optimization (cf. Efron

et al. (2004), Friedman et al. (2010)).

To prove the Oracle Property of the ALASSO estimator  $\widehat{\beta}_n$ , we need to impose a set of regularity conditions which are described in the next section.

## 5.2.2 Notation and Conditions

We now introduce some basic notation and notational conventions to be used in the rest of the paper. All asymptotic statements are driven by the sample size  $n$  going to infinity. However, we will often suppress the dependence on  $n$  to ease the notation. In particular, note that in model (5.1), *all* quantities, such as  $\beta$ ,  $\mathbf{x}_i$  etc. depend on  $n$ . Let  $X$  denote the  $n \times p$  design matrix with its  $i$ th row given by  $\mathbf{x}'_i$  of (5.1),  $i = 1, \dots, n$ . For a vector  $\mathbf{z} \in \mathbb{R}^p$ , write  $\mathbf{z} = (\mathbf{z}^{(1)'}, \mathbf{z}^{(2)'})'$  where  $\mathbf{z}^{(1)}$  is  $p_0 \times 1$ . Also, let  $D_n^{(1)}$  denote the  $q \times p_0$  submatrix of  $D_n$ . Let  $C_n = X'X/n$  and partition it as

$$C_n = \begin{bmatrix} C_{11,n} & C_{12,n} \\ C_{21,n} & C_{22,n} \end{bmatrix},$$

where  $C_{11,n}$  is of order  $p_0 \times p_0$ . Note that  $c_{ij} = n^{-1} \sum_{k=1}^n x_{ki}x_{kj}$  gives the  $(i, j)$ th element of  $C_n$ ,  $1 \leq i, j \leq p$ . Let  $\eta_{11,n}$  be the smallest eigenvalue of  $C_{11,n}$  and  $\kappa_n = \min\{|\beta_{j,n}| : j \in I_n\}$ .

We shall make use of the following conditions:

(C.1) (i)  $c_{jj} = 1$  for all  $j = 1, \dots, p$ .

(ii) There exists  $\delta \in (0, 1)$ , such that for all  $\mathbf{x} \in \mathbb{R}^{p_0}$ ,  $\mathbf{y} \in \mathbb{R}^{p-p_0}$  and  $n > \delta^{-1}$ ,

$$(\mathbf{x}'C_{12,n}\mathbf{y})^2 \leq \delta^2 (\mathbf{x}'C_{11,n}\mathbf{x}) \cdot (\mathbf{y}'C_{22,n}\mathbf{y}).$$

(iii)  $\max\{\|D_n^{(1)}C_{11,n}^{-1}\mathbf{x}_i^{(1)}\| : 1 \leq i \leq n\} = o(\sqrt{n})$  and  $D_n^{(1)}C_{11,n}^{-1}D_n^{(1)'} \rightarrow \Upsilon$  for some  $q \times q$  nonsingular matrix  $\Upsilon$ .

(C.2) (i)  $E(\epsilon_1) = 0$ ,  $E(\epsilon_1^2) = \sigma^2 \in (0, \infty)$ .

(ii) There exists a sequence  $a_n \rightarrow 0+$  such that  $n^{-1/2} \ll a_n \ll \kappa_n$  and

$$P\left(\max_{1 \leq j \leq p} |\tilde{\beta}_{j,n} - \beta_j| > a_n\right) = o(1).$$

(C.3) (i)  $\max\{|\beta_{j,n}| : j \in I_n\} = O(1)$ .

(ii) For some  $b_n \rightarrow \infty$  (to be specified in the statements of the theorems),

$$\frac{\lambda_n}{\sqrt{na_n^\gamma}} \gg \eta_{11}^{-1}p_0b_n + \left(\eta_{11}^{-1}p_0\lambda_n \sum_{j=1}^{p_0} |\beta_j|^{1-\gamma}\right)^{1/2}.$$

(iii)  $\frac{\lambda_n}{\sqrt{n}}\eta_{11}^{-1/2} \left(\sum_{j=1}^{p_0} |\beta_j|^{-2\gamma}\right)^{1/2} = o(1)$ .

We now comment on the conditions. Condition (C.1)(i) is a restatement of Zhang (2010)'s convention that we adopt here to streamline the exposition. Condition (C.1)(ii) requires the multiple correlation between the relevant variables (with  $\beta_{j,n} \neq 0$ ) and the spurious variables ( $\beta_{j,n} = 0$ ) to be strictly less than one, in absolute value. This condition is weaker than assuming orthogonality of the two sets of variables. Variants of this condition have been used in the literature, particularly in the context of the LASSO: see Meinshausen and Bühlmann (2006), Huang et al. (2008), Chatterjee and Lahiri (2011), and the references therein. Part (iii) of (C.1) specifies conditions on the co-efficients in the linear combinations of  $\hat{\beta}_n$  for the asymptotic normality. The first part is a minimal condition for the validity of the Lindeberg condition, while the second part guarantees

the existence of a limiting normal distribution. (C.1)(iii) is precisely the condition that we require for the matrix  $D_n$  be 'well-behaved'.

Condition (C.2) (i) gives the basic moment conditions on the error variables. As discussed in the Introduction, one of the major goals here is to prove the Oracle Property of the ALASSO method solely under finiteness of the variance of  $\epsilon_1$ , and (C.2) (i) will be shown to be adequate for  $p$  satisfying (5.4). To prove the Oracle Property for  $p$  growing at a faster rate, Additional moment conditions will be given in the statements of the respective results. Part (ii) of Condition (C.2) requires the initial estimator  $\tilde{\beta}_n$  to be consistent for the underlying regression parameter  $\beta$  at a suitable rate. In particular, it requires the initial estimator to be consistent in  $\|\cdot\|_\infty$ -norm at a (sub-root- $n$ ) rate such that its random variation around  $\beta$  in the  $\|\cdot\|_\infty$  norm does not suppress the smallest nonzero regression co-efficient, allowing separation of the noise and the signal with high probability. In the next section, we show that under mild conditions, the LASSO estimator with a suitable choice of the penalty parameter in the LASSO criterion serves as a viable choice for all  $p$  satisfying (5.5).

Finally, consider Condition (C.3). Part (i) of (C.3) says that all nonzero regression parameters lie in a bounded subinterval of  $\mathbb{R}$ , while parts (ii) and (iii) of (C.3) respectively specify a lower and an upper bounds on the penalty parameter  $\lambda_n$  that, in particular, depends on the tuning parameter  $\gamma$  of the ALASSO criterion and the convergence rate  $a_n$  of the initial estimator. Note that here we do allow the smallest nonzero regression coefficient and the smallest eigen value of  $C_{11}$  to go to zero at a suitable rate. The latter implies that  $C_{11}$  is nonsingular, and hence that  $p_0 \leq n$ .

### 5.2.3 Main Results

The first results assert validity of the Oracle Property of the ALASSO method for the polynomial growth of  $p$  as a function of  $n$ .

**Theorem 5.1.** *Suppose that Conditions (C.1)-(C.3) hold with  $b_n = \log n$  and that  $p = O(n^\nu)$  for some  $\nu \in (0, \infty)$ . Then, (5.2) and (5.3) hold.*

Recall that we say that  $\{D_n\}$  is ‘well-behaved’ when (C.1)(iii) holds. Thus, a method has the Oracle Property if (5.2) holds and (5.3) holds for all  $D_n$  satisfying (C.1)(iii).

Theorem 5.1 shows that in the case  $p$  grows at an arbitrary polynomial rate with the sample size, the Oracle Property of the ALASSO method holds whenever  $\epsilon_1$  has a finite variance.

We next consider the case where (5.4) holds, i.e.,

$$\log p = O(n^\alpha)$$

for some  $\alpha \in (0, 1/2)$ . In this case, the Oracle Property of the ALASSO method holds also under finiteness of the variance of  $\epsilon_1$  but the penalty parameter is required to satisfy a stronger condition, as specified below.

**Theorem 5.2.** *Suppose that  $\log p = O(n^\alpha)$  for some  $\alpha \in (0, 1/2)$  and that Conditions (C.1)-(C.3) hold with  $b_n = n^c$  for some  $c \in [\alpha, 1/2)$ . Then, the Oracle Property of the ALASSO method holds.*

Next we consider extending the validity of the Oracle Property of ALASSO allowing  $p$  to grow at a faster rate, namely, for

$$\log p = O(n^\alpha) \tag{5.8}$$

for some  $\alpha \in [1/2, 1)$ . For this, we need to assume that sufficiently higher order absolute moments of  $\epsilon_1$  are finite as specified in Theorem 5.3.

**Theorem 5.3.** *Suppose that  $\log p = O(n^\alpha)$  for some  $\alpha \in [1/2, 1)$  and that  $E|\epsilon_1|^r < \infty$  for some  $r > (1 - \alpha)^{-1}$ . Let Conditions (C.1)-(C.3) hold with  $b_n = n^c$  for some  $c \in [\alpha + \frac{1}{r} - \frac{1}{2}, 1/2)$ . Then, the ALASSO method has the Oracle Property.*

Note that when  $\alpha \geq 1/2$ ,  $r > (1 - \alpha)^{-1} \geq 2$ . Thus, we require higher than the second moment to establish validity of the Oracle Property. Further, in Theorems 5.2 and 5.3, Condition (C.3)(ii) must be satisfied with a larger  $\{b_n\}$  compared to that in Theorem 5.1. As a result, the Oracle Property of the ALASSO in the trans-polynomial dimensions  $p$  require stronger conditions on  $\kappa_n$  and  $\eta_{11}$ .

**Remark:** If the criterion function of the ALASSO method is changed to

$$\sum_{i=1}^n (y_i - \mathbf{x}'_i \beta)^2 + \lambda_n \sum_{j=1}^p \exp(|\tilde{\beta}_{n,j}|^{-1}) |\beta_j|,$$

then the Oracle Property of the modified ALASSO method can be extended to the case

$$\log p = o(n/\log n)$$

under finiteness of the absolute moments.

For Theorems 5.1-5.3 to be useful in practice, it is important that there are preliminary estimators  $\tilde{\beta}_n$  that satisfy (C.3)(ii) under the given moment conditions on  $\epsilon_1$ . The next result shows that the LASSO estimator defined in (5.7) with an appropriate choice of  $\lambda_n$  provides a viable choice for each of the three cases.

**Proposition 5.4.** *Suppose that Conditions (C.1)(i), (ii), (C.2) (i) and -(C.3)(i) hold. Suppose that  $\lambda_n$  in the LASSO criterion is chosen as  $\lambda_n = 6\sigma\sqrt{nb_n}$  where  $b_n$  is as in Theorems 5.1-5.3, under the respective moment conditions on  $\epsilon_1$  and growth conditions on  $p$ . Then, the LASSO estimator of  $\beta$  defined in (5.7) satisfies (C.3)(ii) with  $a_n = K(\sigma)n^{-1/2}\eta_{11}^{-1}p_0b_n$ .*

### 5.3 A Maximal Inequality

**Proposition 5.5.** *Let  $\epsilon_1, \dots, \epsilon_n$  be i.i.d random variables and let  $\{f_{ij}(\cdot) : 1 \leq j \leq p, 1 \leq i \leq n\}$  be Borel measurable functions from  $\mathbb{R} \rightarrow \mathbb{R}$  such that (i)  $|f_{ij}(x)| \leq d_{ij}(1 + |x|^s)$  for all  $x \in \mathbb{R}$ , and for some  $s, d_{ij} \in (0, \infty)$ , (ii)  $E f_{ij}(\epsilon_1) = 0$  and (iii)  $E|\epsilon_i|^{rs} < \infty$ . Suppose that the following condition holds,*

$$(C.4) \text{ For any } M_j \geq \max_{i,j} \{d_{ij}\}(1 + a^s) \text{ and } t_j^0 = M_j^{-1} \log \left( 1 + \frac{r x_j M_j^{r-1}}{(2+r)\tau_j} \right),$$

$$(i) \ r(x - \mu_j)M_j^{r-1} > (2+r)(e^r - 1)\tau_j.$$

$$(ii) \ 2(x - \mu_j)M_j > (2+r)e^r \sigma_j^2 \max\{e^2, 2 \log(re^r \sigma_j^2 M_j^{r-2} \tau_j^{-1})\},$$

$$\text{where } \xi_i = \epsilon_i I(|\epsilon_i| \leq a), \mu_j = \sum_{i=1}^n E f_{ij}(\xi_i), \sigma_j^2 = \sum_{i=1}^n E[f_{ij}^2(\xi_i)] \text{ and}$$

$$\tau_j = \sum_{i=1}^n E[|f_{ij}(\xi_i)|^r].$$

Then under condition (C.4), for any  $a, x \in (0, \infty)$ ,

$$\begin{aligned} & P \left( \max_{1 \leq j \leq p} \left| \sum_{i=1}^n f_{ij}(\epsilon_i) \right| > x \right) \\ & \leq \sum_{j=1}^p 2 \exp \left( -\frac{r}{2+r} x_j t_j^0 - M_j^{1-r} \tau_j t_j^0 \right) + n a^{-rs} E|\epsilon_1|^{rs} I(|\epsilon_1| > a). \end{aligned} \quad (5.9)$$

This maximum inequality is used for the proof of section 5.2, and it is powerful to bound the linear combinations of errors.

## 5.4 Simulation

In the simulation, we shall examine the performances of MCP, SCAD, and ALASSO in high dimensions. In order to justify our findings for different tail behaviors, we will consider the same design matrix and same coefficients but different error distributions ( $N(0, 1)$ ,  $\chi^2(1) - 1$ , Pareto(10), Pareto(3)).

For the simulation settings, the dimension of design matrix  $\mathbf{X}$  is  $n \times p$  with  $n = 200$  and  $p$  ranging from 100 to 3000. The predictors  $\{x_{ij}\}$ 's are generated from *i.i.d*  $N(0, 1)$  distribution. Let  $b_0 = (1, -0.6, 0.9, 0.75, -1, -0.55, -0.75, 0.8, -0.7, 0.9)'$ , we define  $\beta = (b_0, \mathbf{0}_{p-10})$ . The reason that we use  $\beta$  in such scale is to make the comparison more meaningful. We tried smaller  $\beta$ , which makes selection consistency fail; we also tried large  $\beta$ , which produces similar performances with all methods/errors. We generate the true target value by  $\hat{\mathbf{y}} = \mathbf{X}\beta$ , and create the observed target value by adding an error to  $\mathbf{y}$ . As for the errors, we generate *i.i.d* random variables  $\epsilon_i$  from four different distributions:

- (a)  $\epsilon_1 \sim N(0, 1)$ , (b)  $\epsilon_1 \sim \chi^2(1) - 1$ , (c)  $\epsilon_1 \sim \text{Pareto}(10)$  and (d)  $\epsilon_1 \sim \text{Pareto}(3)$ .

The distribution of  $\epsilon_1$  has Gaussian tails in (a), exponential and asymmetric tails (with a finite moment generating function) in (b). As for the Pareto distributions in (c) and (d), the numbers 10 and 3 are the shape parameters which control their tail behaviors. To make the signal to noise ratio comparable, we make the Pareto( $\alpha$ ) symmetric and adjust the scale parameters such that the variance is 1. Note that (c) and (d) have only finite absolute moments of orders less than 10 and 3, respectively.

For the tuning parameter  $\lambda$ , as suggested by Fan and Tang (2013), Wang et al. (2013), BIC criterion does not guarantee the selection consistency in high dimensions. We tried the information criterion  $IC = \log(\sigma_\lambda^2) + |\hat{I}_n| C_n/n$  with different  $C_n$  values:  $C_n = \log n$ ,  $C_n = \log \log n \log n$ , and  $C_n = \log \log n \log p$ . It turns out  $C_n = \log \log n \log p$

is uniformly better than the other two choices under our settings of  $(n, p)$  (in terms of selection accuracy), thus we select  $\lambda$  based on  $HIC = \log(\sigma_\lambda^2) + |\hat{I}_n| \log \log n \log p/n$ .

Since the oracle properties are for selection consistency and asymptotic distribution, thus we only report two performance measures for each situation:

TM: Proportion of times when the identified model is exactly the same as the true model.

This is presented in Table 5.1.

CC: Correlation coefficients for Normal Q-Q plots for the first element of  $\beta$ . To remove outliers, 1% tail of observed  $\hat{\beta}_1$  are removed from both sides. CC is presented in Table 5.2.

Here, we examine various situations with respect to  $p$ . Each simulation is based on 500 independent replications.

As for Table 5.2, we see that all the correlation coefficients for Normal Q-Q plots are above 0.99, which indicates that the estimated  $\beta_1$  always has a normal distribution, even with other error distributions. This table justifies our theory on asymptotic normality.

Previously, we keep  $n$  fixed at 200, and make  $p$  increase. We also run the simulation under another setup, where we keep  $n$  grow and always fix  $p = 2n$ . As we have shown in section 1, using MCP method, Figure 5.1 shows the empirical probability of selecting the true model, and Figure 5.2 shows the Normal Q-Q plots of  $\hat{\beta}_1$ . Here we also include similar plots in Figure 5.3 & 5.4 using ALASSO method, which shows a similar trend as MCP does.

It can be seen in Figure 5.4 that the empirical tail behavior of  $\hat{\beta}_1$  under Pareto errors is not as good as under Normal errors. However, after trimming 1% on both sides, the distribution is quite close to Normal, as shown in Table 5.2.

Table 5.1: *Probability of identifying the true model*

Method	Error Distribution	p						
		100	200	400	800	1500	2000	3000
MCP	N(0,1)	0.886	0.850	0.846	0.852	0.822	0.846	0.788
	$\chi^2(1) - 1$	0.874	0.850	0.830	0.886	0.776	0.844	0.812
	Pareto(10)	0.838	0.860	0.836	0.868	0.838	0.862	0.826
	Pareto(3)	0.880	0.862	0.834	0.848	0.820	0.880	0.812
SCAD	N(0,1)	0.638	0.522	0.450	0.450	0.230	0.352	0.212
	$\chi^2(1) - 1$	0.670	0.552	0.486	0.484	0.280	0.408	0.268
	Pareto(10)	0.606	0.508	0.438	0.442	0.242	0.306	0.232
	Pareto(3)	0.672	0.530	0.520	0.444	0.340	0.378	0.298
ALASSO	N(0,1)	0.666	0.668	0.722	0.728	0.682	0.756	0.766
	$\chi^2(1) - 1$	0.646	0.680	0.724	0.752	0.658	0.808	0.754
	Pareto(10)	0.636	0.650	0.682	0.752	0.642	0.774	0.724
	Pareto(3)	0.682	0.650	0.718	0.726	0.684	0.780	0.740

Table 5.2: *Correlation coefficient for Normal Q-Q plots for  $\hat{\beta}_1$*

Method	Error Distribution	p						
		100	200	400	800	1500	2000	3000
MCP	N(0,1)	0.9971	0.9982	0.9983	0.9974	0.9962	0.9974	0.9988
	$\chi^2(1) - 1$	0.9991	0.9956	0.9978	0.9987	0.9978	0.9974	0.9978
	Pareto(10)	0.9988	0.9979	0.9978	0.9976	0.9988	0.9976	0.9948
	Pareto(3)	0.9934	0.9971	0.9983	0.9983	0.9976	0.9976	0.9978
SCAD	N(0,1)	0.9970	0.9981	0.9987	0.9982	0.9965	0.9934	0.9989
	$\chi^2(1) - 1$	0.9990	0.9959	0.9985	0.9985	0.9950	0.9958	0.9956
	Pareto(10)	0.9989	0.9980	0.9974	0.9980	0.9985	0.9972	0.9955
	Pareto(3)	0.9929	0.9974	0.9973	0.9905	0.9538	0.9945	0.9769
ALASSO	N(0,1)	0.9973	0.9972	0.9984	0.9980	0.9963	0.9969	0.9989
	$\chi^2(1) - 1$	0.9987	0.9954	0.9984	0.9984	0.9980	0.9970	0.9967
	Pareto(10)	0.9986	0.9974	0.9975	0.9980	0.9979	0.9976	0.9950
	Pareto(3)	0.9931	0.9970	0.9988	0.9984	0.9969	0.9972	0.9983

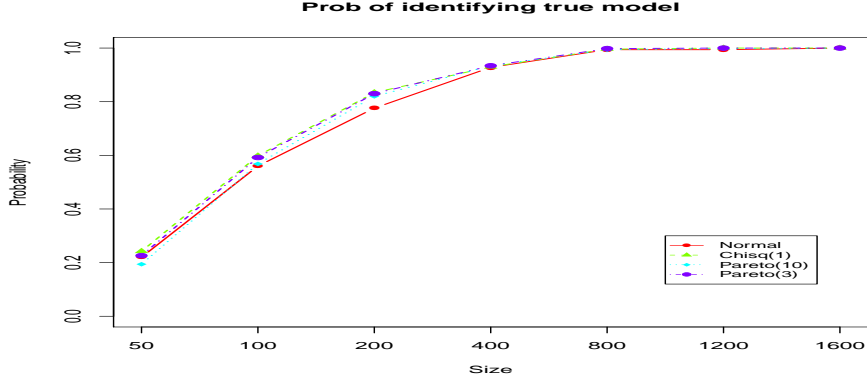


Figure 5.3: Empirical probability of selecting the true model using ALASSO under error distributions (a)-(d)

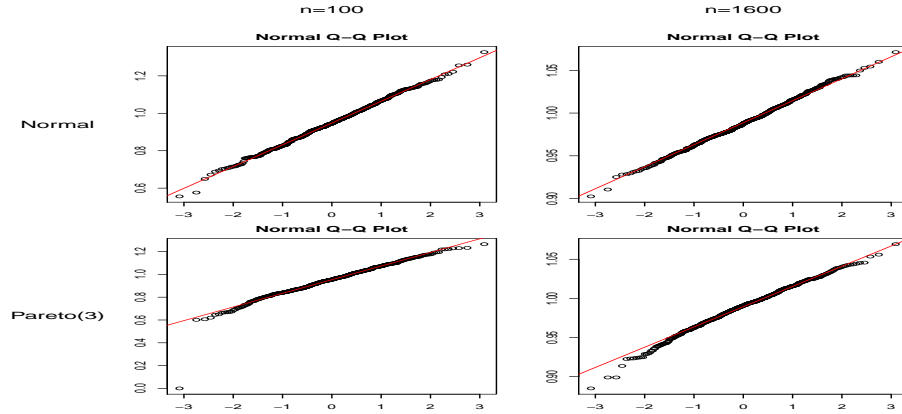


Figure 5.4: Normal Q-Q plots of ALASSO estimates of a nonzero component under error distributions (a) and (d) for sample sizes  $n = 100, 1600$ , based on 500 simulation runs.

## 5.5 Proofs

### 5.5.1 Lemmas

**Lemma 5.1.** Let  $\{\epsilon_i\}$  be i.i.d random variables with  $E\epsilon_1 = 0$ ,  $E\epsilon_1^2 = \sigma^2 \in (0, \infty)$ , and let  $\{x_{ij} : 1 \leq i \leq n, j \in J_n\} \subset \mathbb{R}$  be such that  $\sum_{i=1}^n x_{ij}^2 = n$  for all  $j \in J_n$  and  $\max\{|x_{ij}| : 1 \leq i \leq n, j \in J_n\} \cdot \max\{(E\epsilon_1^2 I(|\epsilon_1| > \sigma n^{1/2-\theta}))^{1/3}, n^{-\theta}\} = O(1)$  for some

$\theta \in (0, 1/2)$ . Then, there exists  $K \in (0, \infty)$  such that for  $n > K$ ,

$$\begin{aligned} & P \left( \left| \sum_{i=1}^n x_{ij} \epsilon_i \right| > 2\sigma\sqrt{n} \log n \text{ for some } j \in J_n \right) \\ & \leq 2|J_n| \cdot \exp(-K \log n \log \log n) + \sigma^{-2/3} (E\epsilon_1^2 I(|\epsilon_1| > \sigma n^{1/2-\theta}))^{1/3} \end{aligned} \quad (5.10)$$

**Remark:** Lemma 5.1 remains valid if the  $\{\epsilon_i\}_{i \geq 1}$  is replaced with a triangular array of row *i.i.d* random variables  $\{\epsilon_{11}, \dots, \epsilon_{nn}\}_{n \geq 1}$  with  $E\epsilon_{11} = 0$ ,  $E\epsilon_{1n}^2 = \sigma_n^2 \in (0, \infty) \forall n \geq 1$

**Proof of lemma 5.1:** Without loss of generality, suppose that  $\sigma^2 = E\epsilon_1^2 = 1$ , then

$$E\epsilon_1^2 I(|\epsilon_1| > a) \leq E\epsilon_1^2 = 1 \text{ for any } a \in (0, \infty).$$

Let  $\delta_{1n}$  and  $\delta_{2n}$  be sequences of positive numbers such that  $\delta_{in} \rightarrow 0+, i = 1, 2$ . The exact choice of  $\delta_{in}$  will be specified later. In Proposition 5.5, set  $a = \sqrt{n}\delta_{1n}$ , then

$$\begin{aligned} |\mu_j| &= \left| \sum_{i=1}^n x_{ij} E\epsilon_i I(|\epsilon_i| \leq a) \right| \\ &= \left| \sum_{i=1}^n x_{ij} \right| |E\epsilon_1 I(|\epsilon_1| > a)| \\ &\leq \left( \sum_{i=1}^n x_{ij}^2 \right)^{1/2} \sqrt{na}^{-1} E\epsilon_1^2 I(|\epsilon_1| > a) \\ &\leq \sqrt{n}\delta_{1n}^{-1} E\epsilon_1^2 I(|\epsilon_1| > \sqrt{n}\delta_{1n}) \\ &\leq \sqrt{n}\delta_{1n}^{-1} E\epsilon_1^2 I(|\epsilon_1| > n^{1/2-\theta}) \\ &\leq \sqrt{n} \quad \text{for all } j \in J_n \end{aligned} \quad (5.11)$$

If we set  $\delta_{1n} = \max\{(E\epsilon_1^2 I(|\epsilon_1| > n^{1/2-\theta}))^{1/3}, n^{-\theta}\} \leq 1$  and  $x = 2\sqrt{n} \log n$ , we have

$$\begin{cases} x_j = x - \mu_j > \sqrt{n} \log n, \\ \max\{|x_{ij}| : 1 \leq i \leq n\} (1 + \sqrt{n}\delta_{1n}) \leq k\sqrt{n} = M_j, \\ \tau_j = \sigma_j^2 = \sum_{i=1}^n x_{ij}^2 E\epsilon_1^2 I(|\epsilon_1| \leq \sqrt{n}\delta_{1n}) \leq nE\epsilon_1^2 \end{cases} \quad (5.12)$$

uniformly in  $j \in J_n$ , for  $n$  large.

Check that with this choice of  $x$ , condition (C.4) holds, uniformly in  $j \in J_n$  for  $n$  large.

Hence, by Proposition 5.5,

$$\begin{aligned} & P\left(\max_{j \in J_n} \left| \sum_{i=1}^n x_{ij} \epsilon_i \right| > x\right) \\ & \leq 2 \sum_{j \in J_n} \exp\left(-\frac{1}{2} \frac{\sqrt{n} \log n}{M_j} \log\left(1 + \frac{\sqrt{n} \log n M_j}{2n\sigma_j^2}\right)\right) + na^{-2} E|\epsilon_1|^2 I(|\epsilon_1| > a). \\ & \leq 2 \sum_{j \in J_n} \exp\left(-\frac{1}{2} \frac{\sqrt{n} \log n}{M_j} \log\left(1 + \frac{\log n M_j}{2E\epsilon_1^2 \sqrt{n}}\right)\right) + \delta_{1n}^{-2} E\epsilon_1^2 I(|\epsilon_1| > \sqrt{n}\delta_{1n}) \\ & \leq 2|J_n| \exp(-K \log n \log \log n) + \delta_{1n}^{-2} E\epsilon_1^2 I(|\epsilon_1| > n^{1/2-\theta}) \end{aligned} \quad (5.13)$$

**Lemma 5.2.** *Let  $\{\epsilon_i\}$  be i.i.d random variables with  $E|\epsilon_1|^r < \infty$  for some  $r \in [2, \infty)$ ,  $E\epsilon_1 = 0$ , and  $E\epsilon_1^2 = \sigma^2 \in (0, \infty)$ . Let  $\{x_{ij} : 1 \leq i \leq n, j \in J_n\} \subset \mathbb{R}$  be such that  $\sum_{i=1}^n x_{ij}^2 = n$  and  $\sum_{i=1}^n |x_{ij}|^r < Kn$  for all  $j \in J_n$ , and  $\max\{|x_{ij}| : 1 \leq i \leq n, j \in J_n\} \cdot \max\{(E|\epsilon_1|^r I(|\epsilon_1| > \sigma n^{1/r-\theta}))^{1/(r+1)}, n^{-\theta}\} = O(1)$  for some  $\theta \in (0, 1/r)$ . Then, for*

any  $\delta, c \in (0, 1/2)$ , there exists  $K = K(c, \theta, \delta) \in [1, \infty)$  such that for all  $n > K$ ,

$$\begin{aligned} & P \left( \max_{j \in J_n} \left| \sum_{i=1}^n x_{ij} \epsilon_i \right| > 2\sigma n^{1/2+c\delta} \right) \\ & \leq 2|J_n| \cdot \exp(-K \log n - Kn^{-1/r} n^{1/2+c-1/r} \log n) + [\sigma^{-r} (E|\epsilon_1|^r I(|\epsilon_1| > \sigma n^{1/r-\theta}))]^{1/(r+1)}. \end{aligned} \quad (5.14)$$

**Proof of lemma 5.2:** Without loss of generality, suppose that  $\sigma^2 = E\epsilon_1^2 = 1$ . As in the proof of lemma 5.1, we set  $a = n^{1/r} \delta_{1n}$ , where  $\delta_{1n} = \max\{(E|\epsilon_1|^r I(|\epsilon_1| > \sigma n^{1/r-\theta}))^{1/(r+1)}, n^{-\theta}\}$ . Then,

$$\begin{aligned} |\mu_j| & \leq n^{1/2} \left( \sum_{i=1}^n x_{ij}^2 \right)^{1/2} E|\epsilon_1| I(|\epsilon_1| > n^{1/r} \delta_{1n}) \\ & \leq n(n^{1/r} \delta_{1n})^{1-r} E|\epsilon_1|^r I(|\epsilon_1| > n^{1/r} \delta_{1n}) \\ & < \delta \sqrt{nn^c} \end{aligned} \quad (5.15)$$

uniformly in  $j \in J_n$ , for  $n$  large. Next apply the Proposition 5.5 with  $x = 2\delta \sqrt{nn^c}$ , we have

$$\begin{cases} x_j = x - \mu_j > \delta \sqrt{nn^c}, \\ \max \{|x_{ij}| : 1 \leq i \leq n, j \in J_n\} (1 + n^{1/r} \delta_{1n}) \leq kn^{1/r} = M_j, \\ \tau_j = \sum_{i=1}^n |x_{ij}|^r E|\epsilon_1|^r I(|\epsilon_1| \leq n^{1/r} \delta_{1n}) \leq Kn, \end{cases} \quad (5.16)$$

uniformly in  $j \in J_n$ , for  $n$  large.

Note that by Jensen's inequality, for  $n$  large,

$$n^{-1} \tau_j \geq (n^{-1} \sum_{i=1}^n x_{ij}^2)^{r/2} \cdot E|\epsilon_1|^r I(|\epsilon_1| \leq n^{1/r} \delta_{1n}) > 1/2, \quad \text{for all } j. \quad (5.17)$$

Now, using (5.16)(5.17), one can easily show that (C.4) holds uniformly in  $j \in J_n$  for  $n$

large. Also, by (5.16) and (5.17),

$$\begin{aligned}
M_j^{1-r} \tau_j t_j^0 &= (Kn^{1/2})^{-r} \log \left( 1 + \frac{rx_j M_j^{r-1}}{(2+r)\tau_j} \right) \\
&\geq K(r)n^{-1} n \log \left( 1 + K(r) \frac{\delta \sqrt{nn^c} n^{1-1/r}}{n} \right) \\
&\geq K(r, \delta, c) \log n,
\end{aligned} \tag{5.18}$$

and similarly,

$$x_j t_j^0 \geq K(r, \delta, c) n^{-1/r} \sqrt{nn^c} \log n. \tag{5.19}$$

Now the lemma follows from Proposition 5.5

### 5.5.2 Proof for ALASSO

For  $\mathbf{x} \in \mathbb{R}^p$ , write  $\mathbf{x} = (\mathbf{x}^{(1)'}, \mathbf{x}^{(2)'})'$  where  $\mathbf{x}^{(1)}$  is of dimension  $p_0 \times 1$ . Define the random vector  $W_n = n^{-1/2} \sum_{i=1}^n \mathbf{x}'_i \epsilon_i$  and write  $W'_n = (W_n^{(1)'}, W_n^{(2)'})$ . Let

$$V_n(\mathbf{u}) = \mathbf{u}' C_n \mathbf{u} - 2\mathbf{u}' W_n + \lambda_n \sum_{j=1}^p |\tilde{\beta}_{j,n}|^{-\gamma} \left( \left| \beta_{j,n} + \frac{u_j}{\sqrt{n}} \right| - |\beta_{j,n}| \right), \quad \mathbf{u} \in \mathbb{R}^p.$$

Note that

$$\hat{\beta}_n = \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^p} V_n(\mathbf{u}). \tag{5.20}$$

Let  $K, K(\cdot)$  denote generic constants with values in  $(0, \infty)$  that do not depend on  $n$ , but may depend on the arguments, if any. Also, unless otherwise specified, limits in order symbols are taken by letting  $n \rightarrow \infty$ .

**Proof of Theorem 5.1:** Fix  $\nu \in (0, \infty)$ . Define the sets

$$\begin{aligned} A_{1,n} &= \{ \|W_n^{(2)}\|_\infty \leq 2\sigma \log n \} \\ A_{2,n} &= \{ \|\tilde{\beta}_n - \beta_n\|_\infty \leq a_n \}. \end{aligned}$$

Note that by Condition (C.3) and Lemma 5.2,  $P(A_{kn}^c) = o(1)$  for  $k = 1, 2$  when  $p = O(n^\nu)$ . Hence, writing  $A_n = A_{1,n} \cap A_{2,n}$ , it follows that  $P(A_n) \rightarrow 1$ . Further, on  $A_n$ ,

$$\|W_n^{(1)}\| \leq 2\sigma\sqrt{p_0} \log n. \quad (5.21)$$

By condition (C.1) on the set  $A_n$ , for  $n$  large,

$$\begin{aligned} V_n(\mathbf{u}) &\geq (1 - \delta) \left[ \mathbf{u}^{(1)'} C_{11,n} \mathbf{u}^{(1)} + \mathbf{u}^{(2)'} C_{22,n} \mathbf{u}^{(2)} \right] - 2\mathbf{u}' W_n + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^p \frac{|u_j|}{|\tilde{\beta}_{j,n}|^\gamma} \\ &\quad - 2\lambda_n \sum_{j=1}^{p_0} \frac{|\beta_{j,n}|}{|\tilde{\beta}_{j,n}|^\gamma} \\ &\geq (1 - \delta) \|\mathbf{u}^{(1)}\|^2 \eta_{11,n} - 2\|\mathbf{u}^{(1)}\| \|W_n^{(1)}\| - 2\lambda_n \sum_{j=1}^{p_0} \frac{|\beta_{j,n}|}{|\tilde{\beta}_{j,n}|^\gamma} \\ &\quad + \sum_{j=p_0+1}^p |u_j| \left( \frac{\lambda_n}{\sqrt{n}} \cdot \frac{1}{|\tilde{\beta}_{j,n}|^\gamma} - 2|W_{j,n}| \right) \\ &\geq \|\mathbf{u}^{(1)}\| \{ (1 - \delta) \|\mathbf{u}^{(1)}\| \eta_{11,n} - K\sqrt{p_0} \log n \} \\ &\quad - 2\lambda_n \sum_{j=1}^{p_0} |\beta_{j,n}|^{1-\gamma} (1 + K(\gamma)a_n) \\ &\quad + K(\gamma) \cdot \sum_{j=p_0+1}^p |u_j| [\lambda_n \cdot n^{-1/2} a_n^{-\gamma} - K \log n]. \end{aligned} \quad (5.22)$$

Next note that for any  $a > 0, c > 0$ , the quadratic form  $Q(x) = ax^2 - bx - c$  is positive whenever  $x > (2a)^{-1}[b + \sqrt{b^2 + 4ac}]$ . Hence, it follows from (5.22) and Conditions (C.1)-

(C.3) that

$$V_n(\mathbf{u}) > 0 \quad \text{for all } \mathbf{u} \in B_n \equiv \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}^{(1)}\| > M_n\} \quad (5.23)$$

where

$$M_n \equiv K(\delta, \gamma) \left\{ \eta_{11,n}^{-1} \sqrt{p_0} \log n + \left( \eta_{11,n}^{-1} \lambda_n \sum_{j=1}^{p_0} |\beta_{j,n}|^{1-\gamma} \right)^{1/2} \right\}.$$

Next, note that for any  $\mathbf{u} = (\mathbf{u}^{(1)}, \mathbf{u}^{(2)})$  and  $p_0 + 1 \leq j \leq p$ , the absolute value of the  $j$ th component of  $\mathbf{u}^{(1)'} C_{12,n}$  is bounded above by

$$\sum_{l=1}^{p_0} |u_l c_{l,j,n}| \leq \left( \sum_{l=1}^{p_0} u_l^2 \right)^{1/2} \left( \sum_{l=1}^{p_0} c_{l,j,n}^2 \right)^{1/2} \leq \|\mathbf{u}^{(1)}\| [\text{tr}(C_{11,n})]^{1/2} c_{j,j,n}^{1/2}.$$

Since by (C.1),  $\max\{c_{j,j,n} : 1 \leq j \leq p\} = O(1)$ , by (C.2) it follows that on the set  $A_n$  and for  $n \geq n_0$  (for some  $n_0 \geq 1$ ), uniformly over  $\{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}^{(1)}\| \leq M_n, \mathbf{u}^{(2)} \neq \mathbf{0}\}$ ,

$$\begin{aligned} & V_n(\mathbf{u}) - V_n(\mathbf{u}^{(1)}, \mathbf{0}) \\ &= \mathbf{u}^{(1)'} C_{12,n} \mathbf{u}^{(2)} + \mathbf{u}^{(2)'} C_{22,n} \mathbf{u}^{(2)} - 2\mathbf{u}^{(2)'} W_n^{(2)} + \frac{\lambda_n}{\sqrt{n}} \sum_{j=p_0+1}^p \frac{|u_j|}{|\tilde{\beta}_{j,n}|^\gamma} \\ &\geq \sum_{j=p_0+1}^p |u_j| \left[ \frac{\lambda_n}{\sqrt{n} |\tilde{\beta}_{j,n}|^\gamma} - 2|W_{j,n}| - \left| \left( \mathbf{u}^{(1)'} C_{12,n} \right)_j \right| \right] \\ &\geq \sum_{j=p_0+1}^p |u_j| \left[ K(\gamma) \lambda_n n^{-1/2} a_n^{-\gamma} - 4\sigma \log n - M_n p_0^{1/2} \right] \\ &> 0. \end{aligned} \quad (5.24)$$

Since  $V_n(\mathbf{0}) = 0$ , from (5.23) and (5.24), it follows that the minimum of  $V_n(\mathbf{u})$  can not be attained at any  $\mathbf{u}$  satisfying either ' $\|\mathbf{u}^{(1)}\| > M_n$ ' or ' $\|\mathbf{u}^{(1)}\| \leq M_n, \mathbf{u}^{(2)} \neq \mathbf{0}$ '. Thus,

$$\text{argmin } V_n(\mathbf{u}) \in B_n \equiv \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}^{(1)}\| \leq M_n, \mathbf{u}^{(2)} = \mathbf{0}\},$$

whenever (5.23) and (5.24) hold. Hence, on the set  $A_n$  and for  $n \geq n_0$ ,

$$\begin{aligned} \sqrt{n} \left( \widehat{\beta}_n - \beta_n \right) &\equiv \operatorname{argmin} V_n(\mathbf{u}) = \operatorname{argmin}_{\|\mathbf{u}^{(1)}\| \leq M_n} V_n(\mathbf{u}^{(1)}, \mathbf{0}) \\ &= \left( \mathbf{U}_n^{(1)'}, \mathbf{0}' \right)', \end{aligned} \quad (5.25)$$

where  $\mathbf{U}_n^{(1)} = C_{11,n}^{-1} \left( W_n^{(1)} - n^{-1/2} \lambda_n \tilde{\mathbf{s}}_n^{(1)} / 2 \right)$  and

$\tilde{\mathbf{s}}_n^{(1)} = (\tilde{s}_{1,n}, \dots, \tilde{s}_{p_0,n})'$  with  $\tilde{s}_{j,n} \equiv \operatorname{sgn}(\beta_{j,n}) |\tilde{\beta}_{j,n}|^{-\gamma}$ ,  $1 \leq j \leq p_0$ . Also, let  $\mathbf{U}_n = \left( \mathbf{U}_n^{(1)'}, \mathbf{0}' \right)' \in$

$\mathbb{R}^p$ . Then, it follows that

$$P\left(\hat{I}_n = I_n\right) = P\left(\sqrt{n}(\widehat{\beta}_n - \beta)^{(2)} = \mathbf{0}\right) \quad (5.26)$$

$$\geq P(A_n) \rightarrow 1 \quad \text{as } n \rightarrow \infty. \quad (5.27)$$

Thus, the ALASSO is variable selection consistent under the conditions of Theorem 5.1.

Next note that by Conditions (C.1)-(C.3) and Taylor's expansion, on  $A_n$ ,

$$\left| \tilde{s}_{j,n} \right| \leq \left[ |\beta_{j,n}| (1 - a_n) \right]^{-\gamma} \leq K |\beta_{j,n}|^{-\gamma}. \quad (5.28)$$

uniformly in  $j \in \{1, \dots, p_0\}$ , for all  $n \geq n_0$ . Also, by (C.1) and (C.3),

$$\begin{aligned} &\frac{\lambda_n}{\sqrt{n}} \left\| D_n^{(1)'} C_{11}^{-1} \tilde{\mathbf{s}}_n^{(1)} \right\| \\ &\leq \frac{\lambda_n}{\sqrt{n}} \left\| D_n^{(1)'} C_{11}^{-1/2} \right\| \left\| C_{11}^{-1/2} \right\| \left\| \tilde{\mathbf{s}}_n^{(1)} \right\| \\ &\leq K \frac{\lambda_n}{\sqrt{n}} \eta_{11}^{-1/2} \left( \sum_{j=1}^{p_0} |\beta_j|^{-2\gamma} \right)^{1/2} \\ &= o(1). \end{aligned} \quad (5.29)$$

Now using Condition (C.2), (5.29) and Lindeberg Central Limit Theorem, it can be shown that

$$\sqrt{n}D_n(\hat{\beta}_n - \beta) \rightarrow^d N(\mathbf{0}, \sigma^2\Upsilon).$$

where  $\rightarrow^d$  denotes convergence in distribution on  $\mathbb{R}^k$  and where  $\Sigma$  is a  $k \times k$  matrix with  $(i, j)$ th element given by  $\tau(x_i, x_j)$ ,  $1 \leq i, j \leq k$ .

**Proofs of Theorems 5.2 and 5.3** Follows by retracing the steps in the proof of Theorem 5.1, where the set  $A_{1n}$  is redefined by replacing  $\log n$  with  $n^c$  for a given  $c$ , in each case. We omit the routine details.

**Proof of Proposition 5.4:** First consider the case  $p = O(n^\nu)$  for some  $\nu \in (0, \infty)$  and suppose that condition (C.1) (C.2)(i) and (C.3)(i) hold. Let

$$V_{1n}(\mathbf{u}) = \mathbf{u}'C_n\mathbf{u} - 2\mathbf{u}'W_n + \lambda_n \sum_{j=1}^p \left( \left| \beta_j + \frac{u_j}{\sqrt{n}} \right| - |\beta_j| \right), \quad \mathbf{u} \in \mathbb{R}^p.$$

Then the LASSO estimator  $\tilde{\beta}_n$  satisfies

$$\sqrt{n}(\tilde{\beta}_n - \beta) = \underset{\mathbf{u} \in \mathbb{R}^p}{\operatorname{argmin}} V_{1n}(\mathbf{u}). \quad (5.30)$$

Define the sets  $A_{1n} = \{\|W_n\|_\infty \leq 2\sigma b_n\}$  where  $b_n = \log n$ . Then by condition (C.1)(i)(ii)

and (C.3)(i), on  $A_{1n}$  we have

$$\begin{aligned}
V_{1n}(\mathbf{u}) &\geq (1 - \delta) \left[ \mathbf{u}^{(1)'} C_{11,n} \mathbf{u}^{(1)} + \mathbf{u}^{(2)'} C_{22,n} \mathbf{u}^{(2)} \right] - 2\mathbf{u}' W_n \\
&\quad + \lambda_n \sum_{j=1}^{p_0} \left\{ \left( |\beta_j| - \frac{|u_j|}{\sqrt{n}} \right) - |\beta_j| \right\} + \frac{\lambda_n}{\sqrt{n}} \sum_{j=p_0+1}^p |u_j| \\
&\geq (1 - \delta) \|\mathbf{u}^{(1)}\|^2 \eta_{11} - 2\|\mathbf{u}^{(1)}\| \|W_n^{(1)}\| - \frac{\lambda_n}{\sqrt{n}} \|\mathbf{u}^{(1)}\| \sqrt{p_0} + \sum_{j=p_0+1}^p |u_j| \left( \frac{\lambda_n}{\sqrt{n}} - 2\sigma b_n \right) \\
&> 0
\end{aligned} \tag{5.31}$$

for all  $\mathbf{u} \in B_{1n} = \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}^{(1)}\| > M_{1n}\}$  where  $M_{1n} = \eta_{11}^{-1} [4\sigma b_n \sqrt{p_0} + \frac{\lambda_n}{\sqrt{n}} \sqrt{p_0}]$ .

Further, on  $A_{1n}$ , for  $\|\mathbf{u}^{(1)}\| \leq M_{1n}$  and large  $n$ ,

$$\begin{aligned}
V_{1n}(\mathbf{u}) &= \mathbf{u}' C_n \mathbf{u} - 2\mathbf{u}' W_n + \lambda_n \sum_{j=1}^{p_0} \left( |\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j| \right) + \frac{\lambda_n}{\sqrt{n}} \sum_{j=p_0+1}^p |u_j| \\
&\geq (1 - \delta) \|\mathbf{u}^{(1)}\|^2 \eta_{11}^{-1} - \lambda_n \sum_{j=1}^{p_0} |u_j / \sqrt{n}| + \frac{\lambda_n}{\sqrt{n}} \sum_{j=p_0+1}^p |u_j| - 2\mathbf{u}' W_n \tag{5.32}
\end{aligned}$$

Note that for any  $a, b \in \mathbb{R}$ ,  $|b| + |a + b| = |-b| + |a + b| \geq |-b + (a + b)| = |a|$ , i.e.

$|a + b| \geq |a| - |b|$ . Hence,

$$\begin{aligned}
V_{1n}(\mathbf{u}) &\geq (1 - \delta) \|\mathbf{u}^{(1)}\|^2 \eta_{11}^{-1} - \frac{\lambda_n}{\sqrt{n}} \|\mathbf{u}^{(1)}\| \sqrt{p_0} - 2\|\mathbf{u}^{(1)}\| \|W_n^{(1)}\| + \sum_{j=p_0+1}^p |u_j| (\lambda_n / \sqrt{n} - 2|w_j|) \\
&\geq -K(\sigma) M_n \left( \frac{\lambda_n}{\sqrt{n}} \sqrt{p_0} + b_n \right) + \sum_{j=p_0+1}^p |u_j| (\lambda_n / \sqrt{n} - 4\sigma b_n) \\
&> 0
\end{aligned} \tag{5.33}$$

provided  $\lambda / \sqrt{n} \geq 6\sigma b_n$ ,  $\|\mathbf{u}^{(2)}\|_1 \geq \sqrt{n} a_n$  and  $\lambda_n / \sqrt{n} \cdot \sqrt{n} a_n \geq M_n (\lambda \sqrt{p_0} / \sqrt{n} + b_n)$ .

### 5.5.3 Proof of the Maximal Inequality

**Proof of Proposition 5.5:** Note that

$$\begin{aligned}
& P \left( \max_{1 \leq j \leq p} \left| \sum_{i=1}^n f_{ij}(\epsilon_i) \right| > x \right) \\
& \leq P \left( \max_{1 \leq j \leq p} \left| \sum_{i=1}^n f_{ij}(\epsilon_i) \right| > x, \max_{1 \leq i \leq n} |\epsilon_i| \leq a \right) + nP(|\epsilon_i| > a) \\
& \leq P \left( \max_{1 \leq j \leq p} \left| \sum_{i=1}^n f_{ij}(\xi_i) \right| > x \right) + nP(|\epsilon_i| > a) \quad \text{where } \xi_i = \epsilon_i I(|\epsilon_i| \leq a) \\
& \leq \sum_{j=1}^p P \left( \left| \sum_{i=1}^n f_{ij}(\xi_i) \right| > x \right) + na^{-rs} \zeta(a), \tag{5.34}
\end{aligned}$$

where  $\zeta(t) = E|\epsilon_1|^{rs} I(|\epsilon_1| > t)$ ,  $t > 0$ .

For a fixed  $j \in \{1, \dots, p\}$ , and consider  $P(|\sum_{i=1}^n f_{ij}(\xi_i)| > x)$ . Note that  $f_{ij}(\xi_i) \leq d_{ij}(1 + a^s)$  for all  $i, j$ . Let  $M_{ij} = d_{ij}(1 + a^s)$  and  $M_j = \max\{M_{ij} : 1 \leq i \leq n\}$ . Define

$$h(x) = \begin{cases} (e^x - 1 - x)/x^r, & \text{if } x \geq r \\ 0, & \text{if } x < r \end{cases} \tag{5.35}$$

For any  $t > 0$ ,

$$\begin{aligned}
& E \exp [tf_{ij}(\xi_i)] \\
& = E \exp [tf_{ij}(\xi_i)] \cdot I(f_{ij}(\xi_i) \leq r/t) + E \exp [tf_{ij}(\xi_i)] \cdot I(f_{ij}(\xi_i) > r/t) \\
& \leq E \left[ 1 + tf_{ij}(\xi_i) + \frac{1}{2}t^2 f_{ij}^2(\xi_i) e^r \right] \cdot I(tf_{ij}(\xi_i) \leq r) \\
& \quad + E [1 + tf_{ij}(\xi_i) + h(tf_{ij}(\xi_i)) (tf_{ij}(\xi_i))^r] \cdot I(tf_{ij}(\xi_i) > r) \\
& \leq 1 + tE f_{ij}(\xi_i) + \frac{1}{2}e^r t^2 E[f_{ij}^2(\xi_i)] + h(M_{i,j}t) \cdot t^r E[|f_{ij}(\xi_i)|^r]. \tag{5.36}
\end{aligned}$$

Hence, it follows that

$$\begin{aligned}
& P\left(\sum_{i=1}^n f_{ij}(\xi_i) > x\right) \\
&= P\left(\exp\left(t \sum_{i=1}^n f_{ij}(\xi_i)\right) > e^{tx}\right), \quad t > 0 \\
&\leq e^{-tx} E \exp\left[t \sum_{i=1}^n f_{ij}(\xi_i)\right] \\
&\leq e^{-tx} \prod_{i=1}^n \left\{1 + t E f_{ij}(\xi_i) + \frac{1}{2} e^r t^2 E[f_{ij}^2(\xi_i)] + h(M_{i,j}t) \cdot t^r E[|f_{ij}(\xi_i)|^r]\right\} \\
&\leq \exp\left(-tx + t\mu_j + \frac{1}{2} e^r t^2 \sigma_j^2 + h(M_j t) t^r \tau_j\right), \tag{5.37}
\end{aligned}$$

where  $\mu_j = \sum_{i=1}^n E f_{ij}(\xi_i)$ ,  $\sigma_j^2 = \sum_{i=1}^n E[f_{ij}^2(\xi_i)]$  and  $\tau_j = \sum_{i=1}^n E[|f_{ij}^2(\xi_i)|^r]$ . We shall now change  $t > 0$  suitably to obtain the desired bound.

Let  $x_j = x - \mu_j$ ,  $t_{1,j} = \frac{\delta x_j}{e^r \sigma_j^2}$  and  $t_{2,j} = t_j^o = M_j^{-1} \log(1 + \tau_j^{-1}(1 - \delta)(x - \mu_j) M_j^{r-1})$ , where  $\delta = 2/(2+r) \in (0, 1)$ .

Note that  $Q_1(t) = \frac{1}{2} e^r t^2 \sigma_j^2 - \delta x_j t$  attains its minimum value at  $t_{1,j} = \frac{\delta x_j}{e^r \sigma_j^2}$ ,  $Q_1'(t) < 0$  for all  $0 < t < t_{1,j}$  and  $Q_1'(t) > 0$  for all  $t > t_{1,j}$ . Also by condition (C.4),  $M_j t_{2,j} \geq r$ ,  $M_j t_{1,j} = 2x_j M_j / [(2+r)e^r \sigma_j^2] > e^2$  and  $M_j t_{1,j} > 2 \log(re^r \sigma_j^2 M_j^{r-2} / \tau_j)$ . Hence, with the fact that  $e^x > 1 + 2 + x^2/2 > 2x$  for all  $x \geq 2$ , one can show that

$$\begin{aligned}
M_j t_{2,j} &= \log(1 + \tau_j^{-1}(1 - \delta)x_j M_j^{r-1}) \\
&< \log(2\tau_j^{-1}(1 - \delta)x_j M_j^{r-1}) \\
&= \log(M_j t_{1,j}) + \log(2(1 - \delta)e^r \sigma_j^2 M_j^{r-1} / [\delta \tau_j]) \\
&< M_j t_{1,j} / 2 + M_j t_{1,j} / 2 = M_j t_{1,j}. \tag{5.38}
\end{aligned}$$

Hence, setting  $t = t_{2j}$  in (5.37), we have

$$\begin{aligned}
& P\left(\sum_{i=1}^n f_{ij}(\xi_i) > x\right) \\
& \leq \exp\left(-t_{2j}x_j + \frac{1}{2}e^r t_{2j}^2 \sigma_j^2 + h(M_j t_{2j}) t_{2j}^r \tau_j\right) \\
& = \exp\left(-t_{2j}x_j + \frac{1}{2}e^r t_{2j}^2 \sigma_j^2 + [e^{M_j t_{2j}} - 1 - M_j t_{2j}] M_j^{-r} \tau_j\right) \\
& = \exp\left(t_{2j}\left[\frac{1}{2}e^r t_{2j} \sigma_j^2 - x_j\right] + (1 - \delta)x_j M_j^{-1} - M_j^{1-r} \tau_j t_{2j}\right) \\
& \leq \exp\left(t_{2j}\left[\frac{1}{2}e^r t_{1j} \sigma_j^2 - x_j\right] + (1 - \delta)x_j M_j^{-1} - M_j^{1-r} \tau_j t_{2j}\right) \\
& = \exp\left(-t_{2j}[1 - \delta/2]x_j + (1 - \delta)x_j M_j^{-1} - M_j^{1-r} \tau_j t_{2j}\right) \\
& = \exp\left(-t_{2j}[1 - \delta]x_j - \frac{\delta}{2}x_j t_{2j} + (1 - \delta)x_j M_j^{-1} - M_j^{1-r} \tau_j t_{2j}\right) \\
& \leq \exp\left(-[1 - \delta]x_j t_{2j} - M_j^{1-r} \tau_j t_{2j}\right)
\end{aligned} \tag{5.39}$$

since  $\frac{\delta}{2}x_j t_{2j} \geq \frac{\delta}{2}x_j r M_j^{-1} = \frac{r}{2+r}x_j M_j^{-1} = (1 - \delta)x_j M_j^{-1}$ .

Replacing  $f_{ij}(\cdot)$  by  $-f_{ij}(\cdot)$ , one can similarly derive a bound on  $P(\sum_{i=1}^n f_{ij}(\xi_i) < -x)$ .

Hence, the result forms.

## REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akadémiai Kiado.
- Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.
- Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.
- Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.
- Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.
- Candes, E. and Tao, T. (2007). The dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . *The Annals of Statistics*, pages 2313–2351.
- Chatterjee, A. and Lahiri, S. N. (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics*, 41(3):1232–1259.
- Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625.
- Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.
- Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press.
- Claeskens, G., Hjort, N. L., Shen, X., Dougherty, D. P., Johnson, W. O., Ishwaran, H., Rao, J. S., Cook, R. D., Li, L., Tsai, C.-L., Raftery, A. E., Zheng, Y., Hjort, N. L., and Claeskens, G. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98(464):pp. 900–945.
- Daniel Hsu, S. M. K. and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32:407–499.

- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.
- Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory*, 57(8):5467–5484.
- Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society, Series B*, 75(3):531–552.
- Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–148.
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.
- Gilmour, S. G. (1996). The interpretation of mallows’s  $c_p$ -statistic. *The Statistician*, pages 49–56.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.
- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (2000). Correction to: “Bayesian model averaging: a tutorial” [Statist. Sci. **14** (1999), no. 4, 382–417; MR1765176 (2001a:62033)]. *Statistical Science*, 15(3):193–195.
- Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603–1618.
- Huang, J. and Xie, H. (2007). Asymptotic oracle properties of SCAD-penalized least squares estimators. 55:149–166.
- Kim, Y., Choi, H., and Oh, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673.
- Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.
- Mallows, C. L. (1973). Some comments on  $c_p$ . *Technometrics*, 15(4):661–675.
- Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.
- Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

- Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of American Statistical Association*, 107(497):223–232.
- Sun, W., Wang, J., and Fang, Y. (2013). Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research*, 14(1):3419–3440.
- Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.
- Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568.
- Wang, L., Kim, Y., and Li, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *The Annals of Statistics*, 0:1–32.
- Wang, T. and Zhu, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*, 102(7):1141–1151.
- Xie, H. and Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, 37(2):673–696.
- Yang, Y. (December 2005). Can the strengths of Aic and Bic be shared? a conflict between model identification and regression estimation. *Biometrika*, 92(4):937–950.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67.
- Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.
- Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.
- Zheng, Z., Fan, Y., and Lv, J. (2014). High dimensional thresholded regression and shrinkage effect. *Journal of the Royal Statistical Society, Series B*, 76(3):627–649.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320.
- Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *ANNALS OF STATISTICS*, 36(4):1509–1533.