이 학 박 사 학 위 논 문

Calibrating Nonconvex Penalized Regression for Logistic Model in Ultra-high Dimension

초고차원 자료에 대한 교정 비볼록 벌점화 로지스틱 회귀분석

2019년 8월

서울대학교 대학원

통계학과

최 세 민

Calibrating Nonconvex Penalized Regression for Logistic Model in Ultra-high Dimension

초고차원 자료에 대한 교정 비볼록 벌점화 로지스틱 회귀분석

지도교수 김 용 대

이 논문을 이학박사 학위논문으로 제출함
2019년 6월

서울대학교 대학원
통계학과
최 세 민

최세민의 이학박사 학위논문을 인준함
2019년 6월

위 원 장	박 병 욱	(인)
부위원장	김 용 대	(인)
위 원	오 희 석	(인)
위 원	임 채 영	(인)
위 원	권 성 훈	(인)

Calibrating Nonconvex Penalized Regression for Logistic Model in Ultra-high Dimension

By

Semin Choi

A Thesis
Submitted in fulfillment of the requirement
for the degree of
Doctor of Philosophy
in Statistics

Department of Statistics
College of Natural Sciences
Seoul National University
August, 2019

ABSTRACT

Calibrating Nonconvex Penalized Regression for Logistic Model in Ultra-high Dimension

Semin Choi

The Department of Statistics

The Graduate School

Seoul National University

In high dimensional linear regression, penalized regression methods are used for estimation and variable selection simultaneously. The LASSO is a penalized regression method which is easy to compute the solution, but the LASSO solution is hard to satisfy the variable selection consistency. Nonconvex penalized regression

methods such as the SCAD and the MCP have the oracle property which contains variable selection consistency. However, direct computation of the global solution to the nonconvex penalized regression is infeasible. The calibrated CCCP is developed which can obtain the oracle estimator as the unique local minimum.

We propose the calibrated CCCP for logistic model. We prove that the calibrated CCCP for logistic model produces a consistent solution path which contains the oracle estimator with probability tending to one. Since the loss function for logistic model is not quadratic, we apply the MLQA-CCCP algorithm for the penalized objective function. Furthermore, we extend the theoretical result to the case of Huber loss instead of the logistic loss. The numerical experiments support our theoretical results.

Keywords: High dimensional regression, penalized regression, logistic loss, Huber loss, variable selection, oracle estimator, MCP, SCAD.

Student Number: 2012 − 20238

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

High dimensional data is the dataset where the number of covariates p is much larger than the number of samples n. High dimensional data analysis arises in many applications including genomics, economics, and neuroscience. In statistics, linear regression is an common approach to modeling the relationship between a response variable and covariates. In linear regression model, the least square method is the most popular for estimating the regression coefficients of covariates. However, it does not work in the high dimensional data, since the inverse matrix of the sample

covariance does not exist. As an alternative to the least square method, there are two approaches: subset selection and penalized regression.

In high dimesional data analysis, subset selection methods such as best subset selection have computational burden and they are unstable. Hence, the penalized regression methods such as LASSO [Tibshirani, 1996], MCP [Zhang et al., 2010] and SCAD [Fan and Li, 2001] are used for high dimensional data analysis. The penalized regression methods simultaneously select the relevant covariates for modeling the response variable, and estimate the regression coefficients.

The LASSO, based on l_1 penalty function, is widely used for the high dimensional data analysis. The LASSO is computationally easy since it is a convex penalized regression method. For example, the LASSO solution can be easily computed by the LARS algorithm or coordinate descent methods. Also, the LASSO has some desired theoretical properties such as the minimax optimal rate. The minimax optimal rate of the LASSO solution can be derived under restricted eigenvalue condition.

However, it is well known that the LASSO need strong irrep-

resentable condition to satisfy variable selection consistency and that the strong irrepresentable condition is hard to be satisfied for high-dimensional data [Zhao and Yu, 2006; Zou, 2006]. On the other hand, the MCP and the SCAD, based on nonconvex penalty functions, do not need such conditions for variable selection consistency. Such nonconvex penalized regression methods have the oracle property under mild conditions for fixed p [Fan and Li, 2001]. For high dimensional data, the oracle estimator itself is a local minimum of SCAD (or MCP) penalized linear regression [Kim et al., 2008b]. However, due to multiple local minima, direct computation of the global solution is infeasible.

The calibrated CCCP [Wang et al., 2013] is a two stage method for the nonconvex penalized objective function. It can obtain the oracle estimator as the unique local minimum. We propose the calibrated CCCP for logistic model. We prove that the calibrated CCCP for logistic model produces a constistent solution path which contains the oracle estimator with probability tending to one. Since the loss function for logistic model is not quadratic, we apply the MLQA-CCCP algorithm [Lee et al., 2016] for the penalized objective function. Furthermore, we extend the theoretical

result to the case of Huber loss instead of the logistic loss. The numerical experiments support our theoretical results.

1.2 Outline of the thesis

In this thesis, the contents are organized as follows. In Chapter 2, we review the penalized regression methods in high-dimensional data analysis. The theoretical properties such as minimax optimal rate, variable selection consistency and oracle property are introduced. Furthermore, we review the calibrated CCCP for linear model and its oracle property. In Chapter 3, we propose the calibrated CCCP for logistic model, and derive the oracle property for the proposed method. Since the loss function is not quadratic, we apply the MLQA-CCCP algorithm for computation. The numerical experiments are also included in chapter 4. In Chapter 5, we summarize the contents of this thesis and provide some discussions for future works.

Chapter 2

Literature Review: Penalized Regression on High Dimensional Regression

2.1 Introduction

Consider the linear regression model,

$$\mathbf{y}_n = \mathbf{X}_n \beta_n^* + \epsilon_n,$$

where $\mathbf{y}_n$ is an $n \times 1$ vector of response, $\mathbf{X}_n = (\mathbf{x}_1^n, ..., \mathbf{x}_n^n)^T = (x_{ij})$ is an $n \times p_n$ design matrix of covariate where $\mathbf{x}_i^n \in \mathbb{R}^{p_n}$ for $i = 1, ..., n$, β_n^* is a $p_n \times 1$ vector of regression coefficients and ϵ_n is an $n \times 1$ vector of random error.

There are two approaches for high dimensional linear regression: subset selection and penalization. There are many subset selection methods such as forward selection, backward elimination and stepwise selection. They select a subset of covariates with some criterion, and calculate the regression coefficients with the selected covariates. Since the number of the selected covariates is less than the number of samples, the least square estimator can be calculated for the subset selection methods even in high dimensional data. However, these subset selection methods are computationally intensive and unstable.

As an alternative to subset selection methods, many penalization methods have been proposed which can select the relevant variables and estimate the coefficients of covariates simultaneously. The LASSO and the nonconvex penalized methods are the two main streams of penalized regression methods in high dimensional regression.

The LASSO, based on l_1 penalty function, is widely used for the high dimensional data analysis. The LASSO is computationally easy since it is a convex penalized regression method. For example, the LASSO solution can be easily computed by the LARS algorithm or coordinate descent methods. Also, the LASSO has some desired theoretical properties such as the minimax optimal rate. The minimax optimal rate of the LASSO solution can be derived under restricted eigenvalue condition.

However, it is well known that the LASSO need strong irrepresentable condition to satisfy variable selection consistency and that the strong irrepresentable condition is hard to be satisfied for high-dimensional data [Zhao and Yu, 2006; Zou, 2006]. On the other hand, the MCP and the SCAD, based on nonconvex penalty functions, do not need such conditions for variable selection consistency. Such nonconvex penalized regression methods have the oracle property under mild conditions for fixed p [Fan and Li, 2001]. For high dimensional data, the oracle estimator itself is a local minimum of SCAD (or MCP) penalized linear regression [Kim et al., 2008b]. However, due to multiple local minima, direct computation of the global solution is infeasible.

In the next two sections, we will review the properties of the LASSO and nonconvex penalized methods.

2.2 LASSO

The LASSO [Tibshirani, 1996] estimator $\hat{\beta}_n^L$ is defined as follows:

$$\hat{\beta}_n^L := \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathbf{y}_n - \mathbf{X}_n \beta\|_2^2 + \lambda_n \|\beta\|_1 \right\}$$

The LASSO can achieve sparsity which means that the estimator produces exatly zero regression coefficients. Hence the LASSO simultaneously select the relevant covariates for modeling the response variable, and estimate the regression coefficients. To show that why the LASSO can achieve sparsity, we represent the LASSO as follows:

$$\hat{\beta}_n^L := \underset{\beta}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{y}_n - \mathbf{X}_n \beta\|_2^2 \text{ subject to } \|\beta\|_1 \leq t_n.$$

It is known that the two form of the LASSO introduced above

are equivalent. In the latter form of the LASSO, as shown in Figure 2.1, the solution of the LASSO occurs at the point of contact between the ellipsoid and the diamond. In Figure 2.1, The ellipsoid means that the set of points which has the same values of $\frac{1}{2n}\|\mathbf{y}_n - \mathbf{X}_n\beta\|_2^2$. Hence, the exact zero element of the solution can be obtained.

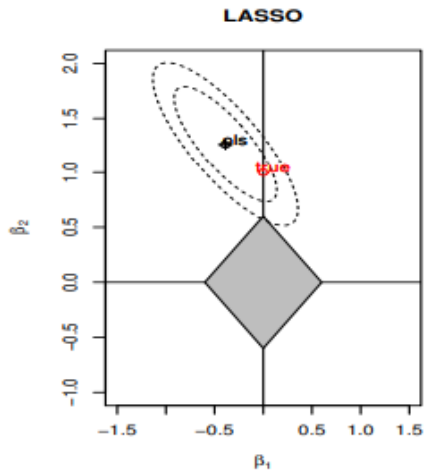

Figure 2.1: Sparsity of the LASSO

Since the objective function for the LASSO is convex, it has the unique global minimum and so is easy to be optimized. For a fixed tuning parameter λ_n, shooting algorithm [Fu, 1998] was introduced for solving the l_1 penalized least square problem. To

compute the entire solution path for tuning parameter λ_n, the algorithms such as the LARS algorithm [Efron et al., 2004] can be used.

The LASSO is not only computationally attractive, but it also has good theoretical properties. Before explaining the theoretical properties of the LASSO, we introduce the results for the minimax lower bound of the l_2-estimation loss and the l_2-prediction loss [Raskutti et al., 2011; Ye and Zhang, 2010]. Under regularity conditions, with probability tending to 1,

$$\min_{\hat{\beta}_n} \max_{\beta_n^* \in \mathbb{B}_0(q_n)} \|\hat{\beta}_n - \beta_n^*\|_2^2 \geq O\Big(\frac{q_n \log p_n}{n}\Big)$$

$$\min_{\hat{\beta}_n} \max_{\beta_n^* \in \mathbb{B}_0(q_n)} \frac{1}{n}\|\mathbf{X}_n(\hat{\beta}_n - \beta_n^*)\|_2^2 \geq O\Big(\frac{q_n \log p_n}{n}\Big)$$

where $\mathbb{B}_0(q_n) := \{\beta \in \mathbb{R}^{p_n} : \sum_{j=1}^{p_n} I(\beta_j \neq 0) \leq q_n\}$.

Consider the LASSO estimator $\hat{\beta}_n^L(\lambda_n)$ with the tuning parameter $\lambda_n = O\Big(\sqrt{\frac{\log p_n}{n}}\Big)$. Under "Restricted Eigenvalue Condition" [Bickel et al., 2009], with probability tending to 1,

$$\|\hat{\beta}_n^L(\lambda_n) - \beta_n^*\|_2^2 \leq C_1 \frac{q_n \log p_n}{n}$$

$$\frac{1}{n}\|\mathbf{X}_n(\hat{\beta}_n^L(\lambda_n) - \beta_n^*)\|_2^2 \leq C_2 \frac{q_n \log p_n}{n}$$

where $C_1, C_2 > 0$ are constants and $q_n := \sum_{j=1}^{p_n} I(\beta_{nj}^* \neq 0)$. Hence, the l_2-estimation loss and the l_2-prediction loss for the LASSO estimator achieves the minimax lower bound.

Even though the prediction loss for the LASSO estimator achieves the minimax lower bound, from the perspective of variable selection the LASSO has not attractive theoretical property. We first introduce the definition of variable selection consistency. Let $A_n(\lambda_n) := \{j : \hat{\beta}_{nj}^L(\lambda_n) \neq 0\}$ and $A_{0n} := \{j : \beta_{nj}^* \neq 0\}$, where $\hat{\beta}_n^L(\lambda_n) = (\hat{\beta}_{n1}^L(\lambda_n), ..., \hat{\beta}_{np_n}^L(\lambda_n))^T$. Then, we say that the lasso variable selection is consistent if $\lim_n P(A_n(\lambda_n) = A_{0n}) = 1$. The optimal estimation rate is available only when $\lambda_n = O(1/\sqrt{n})$ for fixed $p_n = p$, but it leads to inconsistent variable selection, i.e., $\lim_n P(A_n(\lambda_n) = A_{0n}) < 1$ [Zou, 2006].

Zhao and Yu [2006] introduced the so-called "weak irrepresentable condition" which is a necessary condition of the consistency of the lasso variable selection. Without loss of generality, assume that $A_{on} = \{1, 2, ..., q_n\}$ and let $\frac{1}{n}\mathbf{X}_n^T\mathbf{X}_n := \mathbf{C}^n$. Let $\mathbf{C}^n = \begin{bmatrix} \mathbf{C}_{11}^n & \mathbf{C}_{12}^n \\ \mathbf{C}_{21}^n & \mathbf{C}_{22}^n \end{bmatrix}$ where $\mathbf{C}_{11}^n$ is a $q_n \times q_n$ matrix. Then, the weak irrepresentable condition is as follows:

$$|\mathbf{C}_{21}^n(\mathbf{C}_{11}^n)^{-1}\text{sign}(\beta_{n(1)})| < 1,$$

where $\beta^*_{n(1)} := (\beta^*_{n1}, ..., \beta^*_{nq_n})^T$.

Zhao and Yu [2006] also introduced the so-called "strong irrepresentable condition" which is a sufficient condition of the consistency of the lasso variable selection. The strong irrepresentable condition is that there exists a positive constant η satisfying

$$|\mathbf{C}^n_{21}(\mathbf{C}^n_{11})^{-1}\text{sign}(\beta_{n(1)})| \leq 1 - \eta.$$

The irrepresentable condition mean that the regression coefficients of the inactive variables on q_n active variables should be uniformly bounded by a constant less than equal to one. Zhao and Yu [2006] also empirically showed that irrepresentable conditions rarely holds for large p_n and q_n, by sampling $\mathbf{C}^n$ from white Wishart distribution. Hence, to achieve the variable selection consistency, the LASSO requires a quite strong condition.

Furthermore, the LASSO solution is biased. Since the same amount of shrinkage is enforced on all nonzero coefficients, they cannot achieve unbiasedness.

2.3 Nonconvex penalized regression

Due to the variable selection inconsistency and the biasedness of the LASSO, nonconvex penalized methods can be good alterna-

tives to the LASSO since they have variable selection consistency and unbiasedness.

The penalized least square estimator is defined as the minimizer of $Q_{\lambda_n}(\beta)$ where

$$Q_{\lambda_n}(\beta) = \frac{1}{2n}\|\mathbf{y}_n - \mathbf{X}_n\beta\|_2^2 + \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_j|), \tag{2.1}$$

where $p_{\lambda_n}(|\beta_j|)$ is a penalty function. For example, if $p_{\lambda_n}(|\beta_j|) = \lambda_n|\beta_j|$, the above estimator is the LASSO estimator.

There are many nonconvex penalties, including the bridge penalty, the SCAD, and the MCP.

For the bridge penalty, the penalty function is defined as follows :

$$p_{\lambda_n}(|\beta_j|) = \lambda_n|\beta_j|^v,$$

where $0 < v < 1$ is a constant.

For the MCP, the penalty function is defined as follows :

$$p_{\lambda_n}(|\beta_j|) = \left(\lambda_n|\beta_j| - \frac{\beta_j^2}{2a}\right) I(0 \le |\beta_j| < a\lambda_n) + \frac{a\lambda_n^2}{2} I(|\beta_j| \ge a\lambda_n),$$

where $a > 0$ is a constant.

For the SCAD, the penalty function is defined as follows :

$$\begin{aligned} p_{\lambda_n}(|\beta_j|) = \; & \lambda_n|\beta_j| I(|\beta_j| < \lambda_n) \\ & + \left(\lambda_n|\beta_j| - \frac{\beta_j^2 - 2\lambda_n|\beta_j| + \lambda_n^2}{2(a-1)} \right) I(\lambda_n \leq |\beta_j| \leq a\lambda_n) \\ & + \frac{(a+1)\lambda_n^2}{2} I(|\beta_j| > a\lambda_n) \end{aligned}$$

As shown in Figure 2.2, the penalty function for the LASSO is convex and the penalty functions for the bridge penalty, the MCP and the SCAD are nonconvex. Hence, for the MCP and the SCAD, $Q_{\lambda_n}(\beta)$ has multiple local minima and direct computation of the global solution to the nonconvex penalized regression is infeasible. There are many optimization algorithms for finding a local minimum. In here, we introduce the CCCP algorithm [Kim et al., 2008b] for finding a local minimum of the $Q_{\lambda_n}(\beta)$.

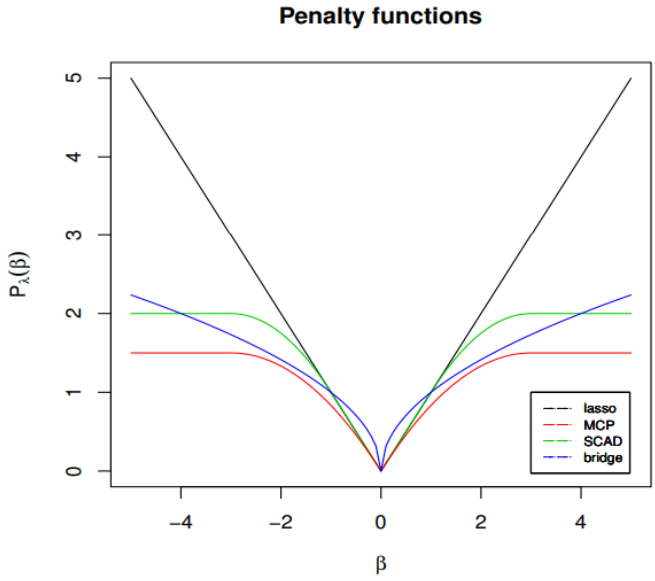

Figure 2.2: The Graph of Some Penalty Functions

For the penalized regression in the equation 2.1, we consider a penalty function $p_{\lambda_n}(|\beta_j|)$ which has the decomposition

$$p_{\lambda_n}(|\beta_j|) = J_{\lambda_n}(|\beta_j|) + \lambda_n|\beta_j|,$$

where $J_{\lambda_n}(|\beta_j|)$ is a differentiable concave function.

For the bridge penalty, the penalty function cannot be decomposed as above.

For the SCAD,

$$J_\lambda(|\beta_j|) = -\frac{\beta_j^2 - 2\lambda|\beta_j| + \lambda^2}{2(a-1)} I(\lambda \le |\beta_j| \le a\lambda) + \left[\frac{(a+1)\lambda^2|}{2} - \lambda|\beta_j|\right] I(|\beta_j| > a\lambda).$$

For the MCP,

$$J_\lambda(|\beta_j|) = -\frac{\beta_j^2}{2a} I(0 \le |\beta_j| < a\lambda) + \left[\frac{a\lambda^2}{2} - \lambda|\beta_j|\right] I(|\beta_j| \ge a\lambda).$$

The penalized objective function $Q_{\lambda_n}(\beta)$ in the equation 2.1 can be rewritten as

$$Q_{\lambda_n}(\beta) = \frac{1}{2n}\|\mathbf{y}_n - \mathbf{X}_n\beta\|_2^2 + \sum_{j=1}^{p_n} J_{\lambda_n}(|\beta_j|) + \lambda_n \sum_{j=1}^{p_n} |\beta_j|.$$

Given a current solution $\beta^{(k)}$ with the initial value $\beta^{(0)} = 0$, the tight convex upper bound is

$$Q_{\lambda_n}(\beta|\beta^{(k)}) = \frac{1}{2n}\|\mathbf{y}_n - \mathbf{X}_n\beta\|_2^2 + \sum_{j=1}^{p_n} \nabla J_{\lambda_n}(|\beta_j^{(k)}|)\beta_j + \lambda_n \sum_{j=1}^{p_n} |\beta_j|,$$

where $\nabla J_{\lambda_n}(t) := \partial \nabla J_{\lambda_n}(t)/\partial t$. We then update the current solution by $\beta^{(k+1)} = \text{argmin}_\beta\, Q_{\lambda_n}(\beta|\beta^{(k)})$ until it converges.

Since minimizing $Q_{\lambda_n}(\beta|\beta^{(k)})$ can be viewed as l_1 penalized least square problem, we can directly apply the LARS or the shooting algorithm.

We summarize the CCCP algorithm in Algorithm 1.

Algorithm 1 The CCCP algorithm

1: **for** $k = 0,...,$ until converges **do**

2:

$$\beta^{(k+1)} := \underset{\beta}{\operatorname{argmin}}\, Q_{\lambda_n}(\beta|\beta^{(k)})$$

where $Q_{\lambda_n}(\beta|\beta^{(k)}) = \frac{1}{2n}\|\mathbf{y}_n - \mathbf{X}_n\beta\|_2^2 + \sum_{j=1}^{p_n} \nabla J_{\lambda_n}(|\beta_j^{(k)}|)\beta_j + \lambda_n \sum_{j=1}^{p_n} |\beta_j|$.

3: **end for**

Let us introduce the definition of the oracle estimator. The oracle estimator $\hat{\beta}_n^{(o)}$ is defined as follows:

$$\hat{\beta}_n^{(o)} = \underset{\beta:\beta_{A_{0n}^c}=0}{\operatorname{argmin}} \frac{1}{2n}\|\mathbf{y}_n - \mathbf{X}_n\beta\|_2^2$$

It means that the oracle estimator is the ideal estimator obtained only with signal variables without penalization. Fan and Li [2001] showed the oracle property for fixed $p_n = p$. On high dimensions, Kim et al. [2008b] showed that the oracle estimator itself is a local minimum of the SCAD or the MCP penalized linear regression. However, it is not guaranteed that our local minimum is the oracle estimator. Kwon and Kim [2012] showed that if the empirical risk is strictly convex, the oracle estimator asymptotically becomes the global minimizer of the SCAD penalized regression. Kim and Kwon [2012] showed that under Sparse Riesz Condition, the ora-

cle estimator asymptotically becomes the unique local minimizer of the SCAD penalized linear regression. Despite of the above theoretical properties, with finite samples, the solution path is still not unique and is not guaranteed to contain the oracle estimator.

2.4 The calibrated CCCP [Wang et al., 2013]

Recall that the objective function $Q_{\lambda_n}(\beta|\beta^{(k)})$ for the nonconvex penalized linear regression in the CCCP algorithm as follows:

$$Q_{\lambda_n}(\beta|\beta^{(k)}) = \frac{1}{2n}\|\mathbf{y}_n - \mathbf{X}_n\beta\|_2^2 + \sum_{j=1}^{p_n} \nabla J_{\lambda_n}(|\beta_j^{(k)}|)\beta_j + \lambda_n \sum_{j=1}^{p_n} |\beta_j|,$$

where $\nabla J_{\lambda_n}(t) := \partial \nabla J_{\lambda_n}(t)/\partial t$.

The calibrated CCCP algorithm consists of the following two steps.

Algorithm 2 The calibrated CCCP algorithm

Let $\hat{\beta}_n^{(1)}(\lambda_n) = \text{argmin}_\beta \, Q_{\tau_n\lambda_n}(\beta|0_{p_n})$, where the choice $\tau_n > 0$ will be discussed later.

2: Let $\hat{\beta}_n(\lambda_n) = \text{argmin}_\beta \, Q_{\lambda_n}(\beta|\hat{\beta}_n^{(1)}(\lambda_n))$.

Note that the first step is viewed as l_1 penalized linear regression with the tuning parameter $\tau_n\lambda_n$. The second step is also viewed as l_1 penalized regression with the tuning parameter λ_n.

The original CCCP algorithm iteratively finds a solution until it converges, but the calibrated CCCP algorithm finds a solution with only two steps.

Let $\hat{\beta}_n^{(o)}$ be the oracle estimator. If $n\tau_n^2\lambda_n^2 \to \infty$, $\log p_n = O(n\tau_n^2\lambda^2)$ and $\tau_n q_n = o(1)$, where $q_n := |A_{0n}| = |\{j : \beta_{nj}^* \neq 0\}$, then under some mild conditions and "restricted eigenvalue condition", Wang et al. [2013] showed that

$$P(\hat{\beta}_n(\lambda_n) = \hat{\beta}_n^{(o)}) \to 1,$$

as $n \to \infty$.

Hence, under the restricted eigenvalue condition, the calibrated CCCP algorithm finds the oracle estimator with probability tending to 1.

2.5 Review of compatibility condition

The key assumption of our theoretical result is "compatibility condition" **(A2)**[Van de Geer et al., 2008] in section 3.3. In this section, we review some conditions related to compatibility condition. Assume that the diagonal elements of the Gram matrix $\Phi_n = \mathbf{X}_n^T\mathbf{X}_n/n$ are all equal to 1.

For a real number $1 \leq u \leq p$, we introduce the following quantities that we will call *restricted eigenvalues*:

$$\phi_{\min}(u) = \min_{\delta \in \mathbb{R}^p : 1 \leq \mathcal{M}(\delta) \leq u} \frac{\delta^T \Phi_n \delta}{|\delta|_2^2}$$
$$\phi_{\max}(u) = \max_{\delta \in \mathbb{R}^p : 1 \leq \mathcal{M}(\delta) \leq u} \frac{\delta^T \Phi_n \delta}{|\delta|_2^2}$$

where $\mathcal{M}(\delta) = |\{j : \delta_j \neq 0\}|$.

And we introduce the following quantities called *restricted correlations*:

$$\theta_{m_1,m_2} = \max \left\{ \frac{c_1^T \mathbf{X}_{n,I_1}^T \mathbf{X}_{n,I_2} c_2}{n|c_1|_2|c_2|2} : I_1 \cap I_2 = \emptyset, |I_i| \leq m_i, c_i \in \mathbb{R}^{I_i} \backslash \{0\}, i = 1, 2. \right\}.$$

Then "Restricted Isometry Condition" [Candes and Tao, 2005] is defined as follows:

$$\delta_s + \theta_{s,s} + \theta_{s,2s} < 1$$

with $s = \|\beta\|_0$ and $\delta_s = \max\{\phi_{\max}(s) - 1, 1 - \phi_{\min}(s)\}$.

Under "Restricted Isometry Condition", Candes and Tao [2005] showed that the sparse signal is exactly recovered by l_1 minimization for the noiseless error. A sufficient condition for restricted isometry condition is that $\delta_s + \delta_{2s} + \delta_{3s} < 1$, since $\theta_{s,s'} \leq \delta_{s+s'}$.

"Uniform Uncertainty Principle" [Candes et al., 2007] is defined as follows:

$$\delta_{2s} + \theta_{s,2s} < 1.$$

Candes et al. [2007] derived the optimal order for the l_2 estimation loss of the Dantzig selector with Gaussian noise.

There are some stable recovery results with slightly different conditions. Candes et al. [2005] derived stable recovery for bounded e rror with the condition $\delta_{3s} + 3\delta_{4s} < 2$. Candes [2008] derived the similar result with the condition $\delta_{2s} < \sqrt{2} - 1$. Cai et al. [2009b] extend the stable recovery result for Gaussian error with the condition $\delta_{1.5s} + \theta_{s,1.5s} < 1$. Cai et al. [2009a] proved the stable recovery result with the condition $\delta_{1.25s} + \theta_{1.25s,s} < 1$ or $\delta_{1.75s} < \sqrt{2} - 1$. These conditions in Cai et al. [2009a] are strictly weaker than the conditions in Candes [2008] and Cai et al. [2009b].

On the other hand, there are some stable recovery results with other types of assumptions, so called *mutual incoherence property*. Let $m = \max_{1 \le i \ne j \le p} |(\Phi_n)_{i,j}|$. Donoho et al. [2005] derived stable recovery for bounded error with the condition $(4s - 1)m < 1$. Cai et al. [2010] extend the stable recovery result for Gaussian error with the condition $(2s - 1)m < 1$. Furthermore, they showed that

this condition is a sharp condition in the sense that if $(2s-1)m = 1$, there is an counter example that can not recover the coefficients stably.

The RIP framework does not contain the MIP framework, or vice versa. An advantage of MIP is that it can be used to deterministically verify whether a given matrix satisfy the condition.

We introduce the restricted eigenvalue condition [Bickel et al., 2009]. FOr some integer s such that $1 \leq s \leq p$ and a positive number c_0, the restricted eigenvalue condition $\mathrm{RE}(s, c_0)$ is that the following condition holds.

$$\kappa(s, c_0) := \min_{\substack{J_0 \subseteq \{1,\ldots,p\}, \\ |J_0| \leq s}} \min_{\substack{\delta \neq 0 \\ |\delta_{J_0^c}|_1 \leq c_0 |\delta_{J_0}|_1}} \frac{|\mathbf{X}_n \delta|_2}{\sqrt{n}|\delta_{J_0}|_2} > 0 \qquad (2.2)$$

For some integer s, m such that $1 \leq s \leq p/2$ and $m \geq s$, $s + m \leq p$, denote by J_1 the subset of $\{1, ..., p\}$ corresponding to the m largest absolute value coordinates of δ outside of J_0, and define $J_{01} := J_0 \cup J_1$. Then, the slightly stronger version of the above restricted eigenvalue condition, $\mathrm{RE}(s, m, c_0)$, is that the following condition holds.

$$\kappa(s, m, c_0) := \min_{\substack{J_0 \subseteq \{1,\ldots,p\}, \\ |J_0| \leq s}} \min_{\substack{\delta \neq 0 \\ |\delta_{J_0^c}|_1 \leq c_0 |\delta_{J_0}|_1}} \frac{|\mathbf{X}_n \delta|_2}{\sqrt{n}|\delta_{J_{01}}|_2} > 0.$$

In Bickel et al. [2009], RE(s, c_0) is used to derived the upper bound of l_2 prediction loss, and RE(s, c_0) is used to derived the upper bound of l_2 estimation loss. *Uniform Uncertainty Principle* mentioned above is a sufficient condition of the restricted eigenvalue condition RE$(s, 1)$. Generally, $\delta_{2s} + c_0\theta_{s,2s} < 1$ implies RE(s, c_0). On the other hand, it is obvious that the restricted eigenvalue condition is stronger than compatibility condition. Hence, compatibility condition in our assumptions is quite a weak assumption for deriving estimation error of the LASSO estimator.

Other related conditions of compatibility condition are also introduced in Van De Geer et al. [2009], which is summarized in figure 2.3 [Van De Geer et al., 2009] below.

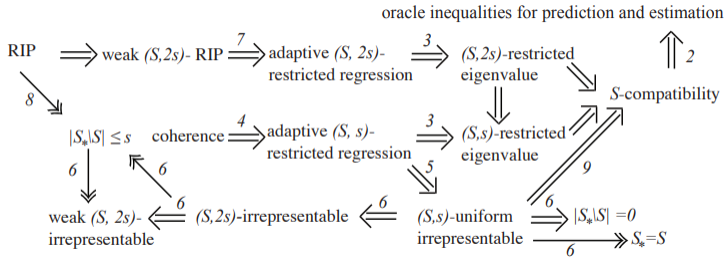

Figure 2.3: Compatibility condition and its related conditions

2.6 Algorithms for l_1 penalized regression

Each step of the proposed calibrated CCCP algorithm for logistic model in section 3.2 need an algorithm for l_1 penalized regression. In this section, we review some algorithms for l_1 penalized regression with general convex loss. Rosset et al. [2004] suggested a path-following algorithm for convex, twice differentiable loss function with an l_1 penalty. If the change in the regularization parameter at every iteration is ϵ, then the solution path we generate is guaranteed to be within $O(\epsilon^2)$ from the true path of penalized optimal solutions. Zhao and Yu [2004] developed a path-following algorithm for any convex loss function with an l_1 penalty. This algorithm finds the exact solutions at uniformly spaced values of $\|\beta\|_1$. Park and Hastie [2007] developed a path-following algorithm for generalized linear model with an l_1 penalty. This algorithm ensures that the solutions are exact at the locations where the active set changes. In Kim et al. [2008a], a gradient decent algorithm for general convex loss functions with an l_1 penalty was introduced. Friedman et al. [2010] developed the local quadratic approximation algorithm. The l_1-penalized quadratic functionis easy to be minimized by using the LARS or coordinate descent algorithms.

There are some modified version of the LQA algorithm. Among them, Yuan et al. [2012] adapts a line search method at every iteration of the inner loop in the coordinate descent algorithm. Hence the number of line searches is proportional to the dimension of parameters. On the other hand, Lee et al. [2016] adapts one line search per one iteration of the outer loop and has descent property. The proposed calibrated CCCP for logistic model apply MLQA algorithm in Lee et al. [2016]. Details will be introduced in section 3.2.

Chapter 3

The calibrated CCCP for logistic model

3.1 Introduction

Let P_n^* be the distribution of (X^n, Y), where $X^n = (X_1, ..., X_{p_n})^T$ is a random vector and Y is a binary random variable. Let $\{(\mathbf{x}_i^n, y_i)\}_{i=1}^n$ be i.i.d copies of (X^n, Y). We consider the logistic regression model as follows:

$$P(Y = 1|X^n = \mathbf{x}) = \frac{\exp(\mathbf{x}^T\beta_n^*)}{1 + \exp(\mathbf{x}^T\beta_n^*)},$$

where $\beta_n^* = (\beta_{n1}^*, \beta_{n2}^*, ..., \beta_{n,p_n}^*)^T$ is the true parameter.

Then, the negative log-likelihood function is

$$P_n\gamma_\beta = \frac{1}{n}\sum_{i=1}^{n}\gamma_\beta(\mathbf{x}_i^n, y_i),$$

where $\gamma_\beta(\mathbf{x}, y) := -y\mathbf{x}^T\beta + \log(1 + \exp(\mathbf{x}^T\beta))$.

Let $\mathbf{X}_n = (\mathbf{x}_1^n, ..., \mathbf{x}_n^n)^T = (x_{ij})$ be an $n \times p_n$ design matrix. For simplicity, we assume that the covariates are normalized, i.e., $\sigma_j^2 := \mathbb{E}[X_j^2] = 1$ for $j = 1, ..., p_n$.

We consider the problem of minimizing the following objective function

$$Q_{\lambda_n}(\beta) = P_n\gamma_\beta + \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_j|), \tag{3.1}$$

where $p_{\lambda_n}(\cdot)$ is a penalty function which depends on a tuning parameter $\lambda_n > 0$.

3.2 The proposed algorithm

First, we introduce the original CCCP algorithm for logistic model. For the penalized regression in the equation 3.1, we consider a penalty function $p_{\lambda_n}(|\beta_j|)$ which has the decomposition

$$p_{\lambda_n}(|\beta_j|) = J_{\lambda_n}(|\beta_j|) + \lambda_n|\beta_j|,$$

where $J_{\lambda_n}(|\beta_j|)$ is a differentiable concave function.

For the SCAD,

$$J_\lambda(|\beta_j|) = -\frac{\beta_j^2 - 2\lambda|\beta_j| + \lambda^2}{2(a-1)} I(\lambda \le |\beta_j| \le a\lambda) + \left[\frac{(a+1)\lambda^2|}{2} - \lambda|\beta_j|\right] I(|\beta_j| > a\lambda).$$

For the MCP,

$$J_\lambda(|\beta_j|) = -\frac{\beta_j^2}{2a} I(0 \le |\beta_j| < a\lambda) + \left[\frac{a\lambda^2}{2} - \lambda|\beta_j|\right] I(|\beta_j| \ge a\lambda).$$

The penalized objective function $Q_{\lambda_n}(\beta)$ in the equation 3.1 can be rewritten as

$$Q_{\lambda_n}(\beta) = \frac{1}{2n}\|\mathbf{y}_n - \mathbf{X}_n\beta\|_2^2 + \sum_{j=1}^{p_n} J_{\lambda_n}(|\beta_j|) + \lambda_n \sum_{j=1}^{p_n} |\beta_j|.$$

Given a current solution $\beta^{(k)}$ with the initial value $\beta^{(0)} = 0$, the tight convex upper bound is

$$Q_{\lambda_n}(\beta|\beta^{(k)}) = \frac{1}{2n}\|\mathbf{y}_n - \mathbf{X}_n\beta\|_2^2 + \sum_{j=1}^{p_n} \nabla J_{\lambda_n}(|\beta_j^{(k)}|)\beta_j + \lambda_n \sum_{j=1}^{p_n} |\beta_j|,$$

where $\nabla J_{\lambda_n}(t) := \partial \nabla J_{\lambda_n}(t)/\partial t$. We then update the current solution by $\beta^{(k+1)} = \operatorname{argmin}_\beta Q_{\lambda_n}(\beta|\beta^{(k)})$ until it converges.

For the logistic model, $P_n\gamma_\beta + \sum_{j=1}^{p_n} \nabla J_{\lambda_n}(|\beta_j^{(k)}|)\beta_j$ is not quadratic with respect to β, where $\gamma_\beta(\mathbf{x}, y) := -y\mathbf{x}^T\beta + \log(1 + \exp(\mathbf{x}^T\beta))$. Hence, to update the current solution by $\beta^{(k+1)} = \operatorname{argmin}_\beta Q_{\lambda_n}(\beta|\beta^{(k)})$

for each k, we introduce the modified local quadratic approximation(MLQA) algorithm [Lee et al., 2016].

$$Q_{\lambda_n}(\beta|\beta^{(k)}) = P_n\gamma_\beta + \sum_{j=1}^{p_n} \nabla J_{\lambda_n}(|\beta_j^{(k)}|)\beta_j + \lambda_n \sum_{j=1}^{p_n} |\beta_j|.$$

Let $L(\beta|\beta^{(k)}) = P_n\gamma_\beta + \sum_{j=1}^{p_n} \nabla J_{\lambda_n}(|\beta_j^{(k)}|)\beta_j$. Given a current estimator $\tilde{\beta}$, a local quadratic approximation $\tilde{L}$ of L around $\tilde{\beta}$ is as follows:

$$\begin{aligned}\tilde{L}(\beta|\tilde{\beta}, \beta^{(k)}) := L(\tilde{\beta}|\beta^{(k)}) \quad &+\nabla L(\tilde{\beta}|\beta^{(k)})^T(\beta - \tilde{\beta}) \\ &+(\beta - \tilde{\beta})^T\nabla^2 L(\tilde{\beta}|\beta^{(k)})(\beta - \tilde{\beta})/2,\end{aligned}$$

where $\nabla L(\beta|\beta^{(k)}) = \partial L(\beta|\beta^{(k)})/\partial\beta$ and $\nabla^2 L(\beta|\beta^{(k)}) = \partial^2 L(\beta|\beta^{(k)})/\partial\beta^2$.

With the local quadratic approximation $\tilde{L}(\beta|\tilde{\beta}, \beta^{(k)})$ defined as above, let

$$\tilde{Q}_{\lambda_n}(\beta|\tilde{\beta}, \beta^{(k)}) = \tilde{L}(\beta|\tilde{\beta}, \beta^{(k)}) + \lambda_n \sum_{j=1}^{p_n} |\beta_j|.$$

Then the MLQA algorithm for minimizing $Q_{\lambda_n}(\beta|\beta^{(k)})$ is as follows:

1. Set the initial estimator $\tilde{\beta} := \beta^{(k)}$.

2. Repeat the following steps until converges:

   - Find $\hat{\beta}^a = \text{argmin}_\beta \tilde{Q}_{\lambda_n}(\beta|\tilde{\beta}, \beta^{(k)})$.

- Find $\hat{h} = \text{argmin}_{h>0} Q_{\lambda_n}(h\hat{\beta}^a + (1-h)\tilde{\beta}|\beta^{(k)})$.
- Update $\tilde{\beta}$ by $\beta^{\hat{h}} = \hat{h}\hat{\beta}^a + (1-\hat{h})\tilde{\beta}$.

The MLQA algorithm has descent property which is that if $\tilde{\beta}$ is not the minimizer of $Q_\lambda(\beta|\beta^{(k)})$, then there exists a $h > 0$ such that $Q_\lambda(\beta^h|\beta^{(k)}) < Q_\lambda(\tilde{\beta}|\beta^{(k)})$.

Recall that

$$Q_{\lambda_n}(\beta|\beta^{(k)}) = P_n\gamma_\beta + \sum_{j=1}^{p_n} \nabla J_{\lambda_n}(|\beta_j^{(k)}|)\beta_j + \lambda_n \sum_{j=1}^{p_n} |\beta_j|,$$

where $\gamma_\beta(\mathbf{x}, y) := -y\mathbf{x}^T\beta + \log(1 + \exp(\mathbf{x}^T\beta))$ for the logistic model.

The proposed calibrated CCCP algorithm for logistic model consists of the following two steps.

1. Let $\hat{\beta}_n^{(1)}(\lambda_n) = \text{argmin}_\beta Q_{\tau_n\lambda_n}(\beta|0_{p_n})$, where the choice $\tau_n > 0$ will be discussed later.
2. Let $\hat{\beta}_n(\lambda_n) = \text{argmin}_\beta Q_{\lambda_n}(\beta|\hat{\beta}_n^{(1)}(\lambda_n))$.

We use the modified local quadratic approximation algorithm [Lee et al., 2016] for each step of the calibrated nonconvex penalized regression.

3.3 Assumptions

We introduce the assumptions for the main theorem.

(A1) It holds that

$$K_{p_n} := \max_{1 \le k \le p_n} \|X_k\|_\infty < \infty,$$

where $\|X_k\|_\infty$ denotes the sup-norm.

(A2) *Compatibility Condition* [Van de Geer et al., 2008]

$$\kappa_n = \kappa(q_n) := \min_{\substack{J_0 \subseteq \{1,\dots,p_n\}, \\ |J_0| \le q_n}} \min_{\substack{\delta \ne 0, \\ |\delta_{J_0^c}|_1 \le 3|\delta_{J_0}|_1}} \frac{\sqrt{q_n}\|(X^n)^T\delta\|}{|\delta_{J_0}|_1} > 0$$

where $\|\cdot\|$ denotes the $L_2(Q^n)$-norm and Q^n is the distribution of X^n.

(A3)

1. The penalty function $p_\lambda(t)$ is assumed to be increasing and concave for $t \in [0, +\infty)$ with a continuous derivative $\dot{p}_\lambda(t)$ on $(0, +\infty)$.
2. $\nabla J_\lambda(|t|) = \lambda \text{sign}(t)$ for $|t| > a\lambda$, where $a > 1$ is a constant.
3. $|\nabla J_\lambda(|t|)| \le |t|$ for $|t| \le b\lambda$, where $b \le a$ is a positive constant.

(A4) Assume that $\lambda_n = o(d_n^*)$ and $\tau_n = o(1)$,

where $d_n^* := \min\{|\beta_{nj}^*| : \beta_{nj}^* \neq 0\}$.

(A5) There exists a positive constant ρ such that

$\lambda_{\min}(n^{-1}(\mathbf{X}_n)_{A_{0n}}^T(\mathbf{X}_n)_{A_{0n}}) \geq \rho$, where $A_{0n} = \{j : \beta_{nj}^* \neq 0\}$

and $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue.

First, we discuss for the compatibility condition **(A2)**. It is a necessary condition of the restricted eigenvalue condition which is used to prove the same theoretical property for the linear model. It is also used to derive the upper bound of l_2 prediction loss for the LASSO. Hence, it is viewed as a mild and reasonable assumption.

Next, we discuss for the assumption **(A3)**. Recall that $p_\lambda(t) = J_\lambda(t) + \lambda(t)$. For the SCAD and the MCP, we can easily check that **(A3)** holds. On the other hand, the brid penalty $p_\lambda(t) = \lambda t^v$ for $0 \leq v \leq 1$ does not satisfy **(A3)**, since the second condition of **(A3)** implies that $p'_\lambda(t) = 0$ for $|t| > a\lambda$ and the third condition of **(A3)** implies that $\lim_{t\to 0+} p'_\lambda(t) = \lambda$. The figure 3.1 and 3.2 are the graph of SCAD and bridge penalty, respectively.

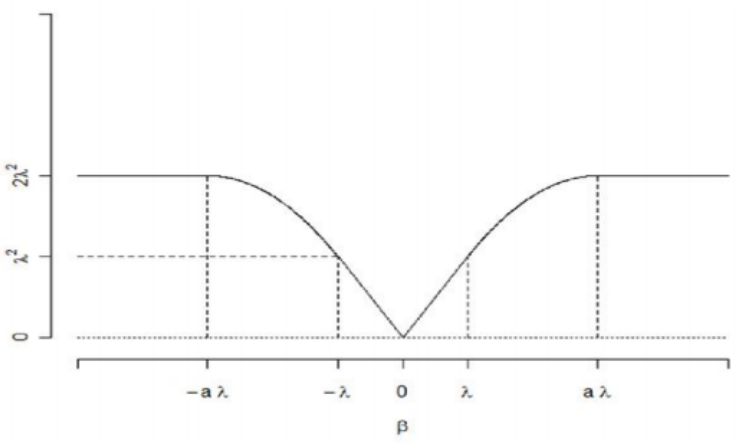

Figure 3.1: The Graph of SCAD penalty

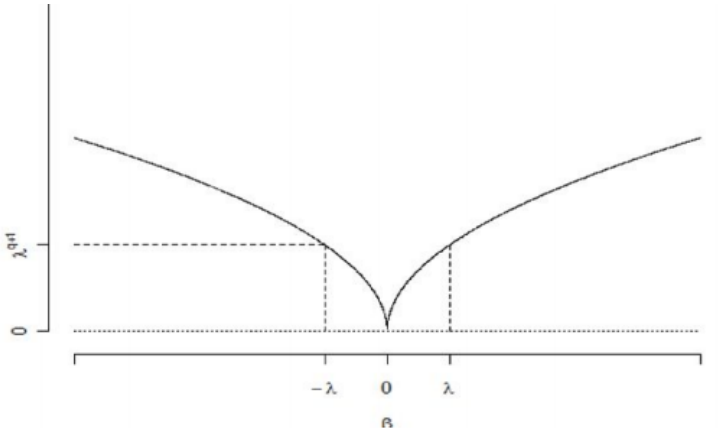

Figure 3.2: The Graph of bridge penalty with $v = 0.5$

3.4 Theoretical properties

We introduce the main theorem of this thesis.

Theorem 1. *Assume that conditions (A1)-(A5) hold. Let* $A_{0n} := \{j : \beta^*_{nj} \neq 0\}$ *and let* $\hat{\beta}^{(o)}_n$ *be the oracle estimator.*
If $\tau_n \kappa_n^{-2} q_n = o(1)$, $n\lambda_n \to \infty$, $\frac{\log p_n}{n} = o(1)$ *and* $\tau_n \lambda_n \asymp \sqrt{\frac{\log p_n}{n}}$, *then*

$$P(\hat{\beta}_n(\lambda_n) = \hat{\beta}^{(o)}_n) \to 1,$$

as $n \to \infty$.

Proof. Recall that

$$Q_{\lambda_n}(\beta|\hat{\beta}^{(1)}_n) = P_n\gamma_\beta + \sum_{j=1}^{p_n} \nabla J_{\lambda_n}(|\hat{\beta}^{(1)}_{nj}|)\beta_j + \lambda_n \sum_{j=1}^{p_n} |\beta_j|,$$

Then, we want to show that

$P\Big(\hat{\beta}^{(o)}_n$ is the minimizer of $Q_{\lambda_n}(\beta|\hat{\beta}^{(1)}_n)\Big) \to 1$ as $n \to \infty$.

Since $Q_{\lambda_n}(\beta|\hat{\beta}^{(1)}_n)$ is a convex function of β, the KKT condition is necessary and sufficient for characterizing the minimum.

To verify that $\hat{\beta}^{(o)}_n$ is the minimizer of $Q_{\lambda_n}(\beta|\hat{\beta}^{(1)}_n)$, it is sufficient to show that

$$-\frac{1}{n}\sum_{i=1}^{n} x_{ij}[y_i - b'(\mathbf{x}_i^T\hat{\beta}^{(o)}_n)] + \nabla J_{\lambda_n}(|\hat{\beta}^{(1)}_{nj}|) + \lambda_n \text{sign}(\hat{\beta}^{(o)}_{nj}) = 0, \quad j \in A_{0n} \tag{3.2}$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}[y_i - b'(\mathbf{x}_i^T\hat{\beta}_n^{(o)})] + \nabla J_{\lambda_n}(|\hat{\beta}_{nj}^{(1)}|)\right| \leq \lambda_n, \quad j \notin A_{0n}, \ (3.3)$$

where $b(\cdot) := \log(1 + \exp(\cdot))$.

Hence, it is sufficient to prove that $P\Big((3.2) \text{ and } (3.3) \text{ hold}\Big) \to 1$ as $n \to \infty$.

Recall that

$$\hat{\beta}_n^{(1)} = \underset{\beta}{\operatorname{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n}[y_i\mathbf{x}_i^T\beta - b(\mathbf{x}_i^T\beta)] + \tau_n\lambda_n\|\beta\|_1 \right\},$$

For (3.2), first note that $-\frac{1}{n}\sum_{i=1}^{n} x_{ij}[y_i - b'(\mathbf{x}_i^T\hat{\beta}_n^{(o)})] = 0$ for $j \in A_{0n}$. Hence, it is sufficient to show that

$P\Big(\nabla J_{\lambda_n}(|\hat{\beta}_{nj}^{(1)}|) = -\lambda_n \text{sign}(\hat{\beta}_{nj}^{(o)}) \text{ for all } j \in A_{0n}\Big) \to 1 \text{ as } n \to \infty.$

Let $F_{n1} := \{\|\hat{\beta}_n^{(1)} - \beta_n^*\|_1 \leq C\tau_n\kappa_n^{-2}q_n\lambda_n\}$ for some constant $C > 0$. Since $\tau_n\kappa_n^{-2}q_n = o(1)$, on the event F_{n1}, $\|\hat{\beta}_n^{(1)} - \beta_n^*\|_\infty \leq \|\hat{\beta}_n^{(1)} - \beta_n^*\|_1 \leq \lambda_n/2$ for all n sufficiently large.

Let $F_{n2} := \{\|\hat{\beta}_n^{(o)} - \beta_n^*\|_\infty \leq \lambda_n/2\}$. Since $\lambda_n = o(d_n^*)$, on the event $F_{n1} \cap F_{n2}$, we have $\text{sign}(\hat{\beta}_{nj}^{(1)}) = \text{sign}(\hat{\beta}_{nj}^{(o)})$, for $j \in A_{0n}$ and $\min_{j\in A_{0n}} |\hat{\beta}_{nj}^{(1)}| \geq a\lambda_n$. Hence, by the condition **(A3)**, on the event $F_{n1} \cap F_{n2}$, $\nabla J_{\lambda_n}(|\hat{\beta}_{nj}^{(1)}|) = -\lambda_n \text{sign}(\hat{\beta}_{nj}^{(1)}) = -\lambda_n \text{sign}(\hat{\beta}_{nj}^{(o)})$. By Lemma 1, for $\tau_n\lambda_n \asymp \sqrt{\log(2p_n)/n}$,

$$P(F_{n1}) \geq 1 - \exp[-16na_n],$$

where $a_n = \left(\sqrt{\frac{2\log(2p_n)}{n}} + \frac{\log(2p_n)}{n} K_p \right)$.

By Lemma 2, we have $P(F_{n2}) \geq 1 - 2q_n \exp(-Anq_n^2\lambda_n^2)$, where $A > 0$ is a constant.

For (3.3), it is sufficient to show that

$$P\left(\left| \frac{1}{n} \sum_{i=1}^{n} x_{ij}[y_i - b'(\mathbf{x}_i^T \hat{\beta}_n^{(o)})] \right| \leq \lambda_n/2 \text{ for all } j \notin A_{0n} \right) \to 1$$

and

$$P\left(\left| \nabla J_{\lambda_n}(|\hat{\beta}_{nj}^{(1)}|) \right| \leq \lambda_n/2 \text{ for all } j \notin A_{0n} \right) \to 1$$

as $n \to \infty$.

On the event F_{n1}, we have $\max_{j \notin A_{0n}} |\hat{\beta}_{nj}^{(1)}| \leq \lambda_n/2$ for all n sufficiently large. Hence, (3.3) holds on the event $F_{n1} \cap F_{n3}$, where $F_{n3} := \left\{ \max_{j \in A_{0n}^c} \left| \frac{1}{n} \sum_{i=1}^{n} x_{ij}[y_i - b'(\mathbf{x}_i^T \hat{\beta}_n^{(o)})] \right| \leq \lambda_n/2 \right\}$.

By Lemma 3, $P(F_{n3}) \geq 1 - 2(p_n - q_n)\exp(-Bn\lambda_n^2)$, where $B > 0$ is a constant. Hence, (3.2) and (3.3) hold with probability at least $1 - \exp(-16na_n) - 2q_n \exp(-Anq_n^2\lambda_n^2) - 2(p_n - q_n)\exp(-Bn\lambda_n^2)$, for all n sufficiently large.

With the assumptions of the theorem 1, $1 - \exp(-16na_n) - 2q_n \exp(-Anq_n^2\lambda_n^2) - 2(p_n - q_n)\exp(-Bn\lambda_n^2) \to 1$ as $n \to \infty$. □

Define

$$\mathbf{Z}_M := \sup_{\|\beta - \beta^*\|_1 \leq M} |(P_n - P_n^*)(\gamma_\beta - \gamma_\beta^*)|$$

where $P_n^* \gamma_\beta := \mathbb{E}_{(\mathbf{X},Y)}[\gamma_\beta(\mathbf{x}, y)]$.

Sublemma 1 (Corollary A.1 in Van de Geer et al. [2008])**.** *Under the conditions **(A1)** and **(A4)**, for all $M > 0$ and all $t > 0$,*

$$P\left(\mathbf{Z}_M \geq 4a_n M\left(1+t\sqrt{2(1+8a_nK_p)}+\frac{8t^2 a_n K_p}{3}\right)\right) \leq \exp[-16na_n^2t^2],$$

where $a_n := \left(\sqrt{\frac{2\log(2p_n)}{n}} + \frac{\log(2p_n)}{n}K_p\right)$.

Lemma 1. *Assume that conditions **(A1)** and **(A2)** hold,*

$f_\beta \in \mathbf{F}_\eta := \{f_\beta : \|f_\beta - f_{\beta_n^*}\|_\infty \leq \eta\}$ *for all* $\|\beta - \beta_n^*\|_1 \leq M$,

$\tau_n\lambda_n \asymp \sqrt{\log(2p_n)/n}$, *and* $\sqrt{\log(2p_n)/n} \to 0$ *as* $n \to \infty$. *Then,*

$$P(\|\hat{\beta}_n^{(1)} - \beta_n^*\|_1 \leq C\tau_n\kappa_n^{-2}q_n\lambda_n) \geq 1 - \exp[-16na_n],$$

where $a_n := \left(\sqrt{\frac{2\log(2p_n)}{n}} + \frac{\log(2p_n)}{n}K_p\right)$.

Proof. Let $\lambda_0 := 4a_n\left(1 + a_n^{-1/2}\sqrt{2(1+8a_nK_p)} + \frac{8K_p}{3}\right)$.

For a fixed $0 < \delta < 1$, let $\epsilon^* := \frac{8C_1q_n\lambda_n^2}{\delta\kappa_n^2}$ and $M := \epsilon^*/\lambda_0$,

where $q_n = |\{j : \beta_{nj}^* \neq 0\}|$, $\lambda_n \geq \lambda_0/8$ and $C_1 > 0$ is a constant.

Since $P(\mathbf{Z}_M \leq \epsilon^*) \geq 1 - \exp[-16na_n]$ by sublemma 1, it is sufficient to show that on the event $\{\mathbf{Z}_M \leq \epsilon^*\}$,

$$\|\hat{\beta}_n^{(1)} - \beta_n^*\|_1 \leq M.$$

Let

$$s := \frac{M}{M + \|\hat{\beta}_n^{(1)} - \beta_n^*\|_1},$$

and $\tilde{\beta} := s\hat{\beta}_n^{(1)} + (1-s)\beta_n^*$. Write $\mathcal{E}(\beta) := P_n^*\gamma_\beta - P_n^*\gamma_{\beta_n^*}$ and $\tilde{\mathcal{E}} := \mathcal{E}(\tilde{\beta})$.

We first note that $\|\tilde{\beta} - \beta_n^*\|_1 \leq M$. By the convexity of $\beta \mapsto \gamma_\beta$, and of $\|\cdot\|_1$,

$$\begin{aligned} P_n\gamma_{\tilde{\beta}} + \lambda_n\|\tilde{\beta}\|_1 &\leq s[P_n\gamma_{\hat{\beta}_n^{(1)}} + \lambda_n\|\hat{\beta}_n^{(1)}\|_1] + (1-s)[P_n\gamma_{\beta_n^*} + \lambda_n\|\beta_n^*\|_1] \\ &\leq P_n\gamma_{\beta_n^*} + \lambda_n\|\beta_n^*\|_1. \end{aligned}$$

Thus,

$$\begin{aligned} \tilde{\mathcal{E}} + \lambda_n\|\tilde{\beta}\|_1 &= -(P_n - P_n^*)(\gamma_{\tilde{\beta}} - \gamma_{\beta_n^*}) + P_n(\gamma_{\tilde{\beta}} - \gamma_{\beta_n^*}) + \lambda_n\|\tilde{\beta}\|_1 \\ &\leq -(P_n - P_n^*)(\gamma_{\tilde{\beta}} - \gamma_{\beta_n^*}) + \lambda_n\|\beta_n^*\|_1 \\ &\leq \mathbf{Z}_M + \lambda_n\|\beta_n^*\|_1. \end{aligned}$$

So, on $\{\mathbf{Z}_M \leq \epsilon^*\}$,

$$\tilde{\mathcal{E}} + \lambda_n\|\tilde{\beta}\|_1 \leq \epsilon^* + \lambda_n\|\beta_n^*\|_1. \tag{3.4}$$

Therefore, we have

$$\tilde{\mathcal{E}} + \lambda_n\|\tilde{\beta}_{\text{out}}\|_1 \leq \epsilon^* + \lambda_n\|\tilde{\beta}_{\text{in}} - \beta_n^*\|_1, \tag{3.5}$$

wheere $A_{0n} := \{j : \beta_{nj}^* \neq 0\}$, $\beta_{nj,\text{in}} := \beta_{nj}I(j \in A_{0n})$ and $\beta_{nj,\text{out}} := \beta_{nj}I(j \notin A_{0n})$.

Case 1. $\lambda_n\|\tilde{\beta}_{\text{in}} - \beta_n^*\|_1 \leq \frac{\epsilon^*}{2}$. Then, (3.5) implies

$$\tilde{\mathcal{E}} + \lambda_n\|\tilde{\beta}_{\text{out}}\|_1 \leq \frac{3}{2}\epsilon^*$$

and hence

$$\tilde{\mathcal{E}} + \lambda_n \|\tilde{\beta} - \beta_n^*\|_1 \leq 2\epsilon^*$$

But then,

$$\|\tilde{\beta} - \beta_n^*\|_1 \leq \frac{2\epsilon^*}{\lambda_n} = \frac{2\lambda_0 M}{\lambda_n} \leq \frac{M}{4}$$

since $\lambda_n \geq 8\lambda_0$. This implies $\|\hat{\beta}_n^{(1)} - \beta_n^*\|_1 \leq M$.

Case 2. $\lambda_n \|\tilde{\beta}_{\text{in}} - \beta_n^*\|_1 \geq \frac{\epsilon^*}{2}$. Then, (3.5) implies

$$\lambda_n \|\tilde{\beta}_{\text{out}}\|_1 \leq \epsilon^* + \lambda_n \|\tilde{\beta}_{\text{in}} - \beta_n^*\|_1 \leq 3\lambda_n \|\tilde{\beta}_{\text{in}} - \beta_n^*\|_1.$$

This means that we can apply the compatibility condition (Assumption **(A2)**). Recall that inequality (3.4) implies

$$\tilde{\mathcal{E}} + \lambda_n \|\tilde{\beta}_{\text{out}}\|_1 \leq \epsilon^* + \lambda_n \|\beta_n^*\|_1 - \lambda_n \|\tilde{\beta}_{\text{in}}\|_1 \leq \epsilon^* + \lambda_n \|\tilde{\beta}_{\text{in}} - \beta_n^*\|_1.$$

Since $\|\tilde{\beta}_{\text{out}}\|_1 = \|\tilde{\beta} - \beta_n^*\|_1 - \|\tilde{\beta}_{\text{in}} - \beta_n^*\|_1$,

$$\tilde{\mathcal{E}} + \lambda_n \|\tilde{\beta} - \beta_n^*\|_1 \leq \epsilon^* + 2\lambda_n \|\tilde{\beta}_{\text{in}} - \beta_n^*\|_1.$$

With the compatibility condition, we find

$$\|\tilde{\beta}_{\text{in}} - \beta_n^*\|_1 \leq \frac{q_n}{\kappa_n} \|f_{\tilde{\beta}} - f_{\beta_n^*}\|.$$

Thus,

$$\tilde{\mathcal{E}} + \lambda_n \|\tilde{\beta} - \beta_n^*\|_1 \leq \epsilon^* + \frac{2\lambda_n \sqrt{q_n}}{\kappa_n} \|f_{\tilde{\beta}} - f_{\beta_n^*}\|.$$

Now, since $f_{\tilde{\beta}} \in \mathbf{F}_\eta$, we can use the margin condition (Assumption **(A3)**):

$$\frac{2\lambda_n\sqrt{q_n}}{\kappa_n}\|f_{\tilde{\beta}} - f_{\beta_n^*}\| \leq C_1\frac{8q_n\lambda_n^2}{\delta\kappa_n^2} + \delta\tilde{\mathcal{E}}.$$

It follows that

$$\tilde{\mathcal{E}} + \lambda_n\|\tilde{\beta} - \beta_n^*\|_1 \leq \epsilon^* + C_1\frac{8q_n\lambda_n^2}{\delta\kappa_n^2} + \delta\tilde{\mathcal{E}} = 2\epsilon^* + \delta\tilde{\mathcal{E}}.$$

Hence,

$$(1-\delta)\tilde{\mathcal{E}} + \lambda_n\|\tilde{\beta} - \beta_n^*\|_1 \leq 2\epsilon^*.$$

This yields

$$\|\tilde{\beta} - \beta_n^*\|_1 \leq \frac{2\lambda_0}{\lambda_n}M \leq \frac{M}{4}.$$

But $\|\tilde{\beta} - \beta_n^*\|_1 \leq M/4$ implies $\|\hat{\beta}_n^{(1)} - \beta_n^*\|_1 \leq M/3 \leq M$. So in both Case 1 and Case 2, we arrive at $\|\hat{\beta}_n^{(1)} - \beta_n^*\|_1 \leq M$. □

Lemma 2. *Assume that conditions (A3)-(A5) hold. Then, for some constant $A > 0$ that depends only on ρ,*

$$P(\|\hat{\beta}_n^{(o)} - \beta_n^*\|_\infty \leq \lambda_n/2) \geq 1 - 2q_n\exp(-Anq_n^2\lambda_n^2).$$

Proof. WLOG, assume that β_n^* is the q_n-dimensional parameter with $\beta_{nj}^* \neq 0$ for all $j = 1, ..., q_n$, and let $\hat{\beta}_n$ be the least square estimator of β_n^*.

Let

$$P_n^{(1)}\gamma_{\beta_n^*} := \left.\frac{\partial P_n\gamma_\beta}{\partial\beta}\right|_{\beta=\beta_n^*} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i[-Y_i + b'(\mathbf{x}_i^T\beta_n^*)].$$

By Taylor expansion,

$$0 = P_n^{(1)}\gamma_{\hat{\beta}_n} = P_n^{(1)}\gamma_{\beta_n^*} + H_n(\hat{\beta}_n - \beta^*)$$

where $H_n = \left[\frac{\partial P_n\gamma_\beta}{\partial\beta\partial\beta^T}\right]_{\beta=\tilde{\beta}}$
for some $\tilde{\beta}_j \in (\min\{\hat{\beta}_{nj}, \beta_{nj}^*\}, \max\{\hat{\beta}_{nj}, \beta_{nj}^*\})$.

Hence,

$$\begin{aligned}
\frac{\lambda_{\min}(H_n)}{q_n}\|\hat{\beta}_n - \beta_n^*\|_\infty &\le \frac{1}{q_n}\|H_n(\hat{\beta}_n - \beta_n^*)\|_2 \\
&\le \|H_n(\hat{\beta}_n - \beta_n^*)\|_\infty \\
&= \|P_n^{(1)}\gamma_{\beta_n^*}\|_\infty \\
&= \max_j\left|\frac{1}{n}\sum_{i=1}^{n}x_{ij}[-Y_i + b'(\mathbf{x}_i^T\beta_n^*)]\right|.
\end{aligned}$$

Then,

$$
\begin{aligned}
&P(\|\hat{\beta}_n - \beta_n^*\|_\infty > \lambda_n/2) \\
&\le P\bigg(\max_j \bigg| \frac{1}{n}\sum_{i=1}^n x_i j[-Y_i + b'(\mathbf{x}_i^T\beta_n^*] \bigg| > \frac{q_n\lambda_n}{2\lambda_{\min}(H_n)} \bigg) \\
&\le \sum_{i=1}^{q_n} P\bigg(\bigg| \frac{1}{n}\sum_{i=1}^n [b''(\mathbf{x}_i^T\beta_n^*)^{1/2}x_{ij}] \frac{-Y_i + b'(\mathbf{x}_i^T\beta_n^*)}{b''(\mathbf{x}_i^T\beta_n^*)^{1/2}} \bigg| > \frac{q_n\lambda_n}{2\lambda_{\min}(H_n)} \bigg) \\
&\le \sum_{j=1}^{q_n} 2\exp\bigg(-\frac{nq_n^2\lambda_n^2}{8\lambda_{\min}(H_n)^2\|v_j\|_2^2} \bigg) \\
&\le 2q_n\exp(-Anq_n^2\lambda_n^2).
\end{aligned}
$$

□

Lemma 3. *Assume that conditions (A3)-(A5) hold. Then, for some constnat $B > 0$,*

$$
\begin{aligned}
P\bigg(\max_{j\in A_{0n}^c} \Big| -\frac{1}{n}\sum_{i=1}^n x_{ij}[Y_i - b'(\mathbf{x}_i^T\hat{\beta}_n^{(o)})] \Big| \le \lambda_n/2 \bigg) \\
\ge 1 - 2(p_n - q_n)\exp(-Bn\lambda_n^2)
\end{aligned}
$$

Proof.

$$
\begin{aligned}
P(F_{n3}^c) &= P\bigg(\max_{j\in A_{0n}^c} \Big| -\frac{1}{n}\sum_{i=1}^n x_{ij}[Y_i - b'(\mathbf{x}_i^T\hat{\beta}_n^{(o)})] \Big| \le \lambda_n/2 \bigg) \\
&\le \sum_{j\in A_{0n}^c} P\bigg(\Big| -\frac{1}{n}\sum_{i=1}^n x_{ij}[Y_i - b'(\mathbf{x}_i^T\hat{\beta}_n^{(o)})] \Big| \le \lambda_n/2 \bigg) \\
&\le 2(p_n - q_n)\exp(-Bn\lambda_n^2),
\end{aligned}
$$

for some $B > 0$. □

In the case of the Huber loss, all assumptions used for logistic loss are satisfied. Hence, Theorem 1 can be directly extended to the Huber loss.

Chapter 4

Experiments

4.1 Simulation studies

We investigate the signal recovery and estimation properties of the proposed method via numerical studies. We compare the results of oracle MLE, LASSO, SCAD, MCP, and our proposed method. The oracle MLE assumes the availability of the knowledge of the true underlying model. The SCAD estimator is obtained by the original CCCP algorithm without calibration. The MCP estimator is obtained with $a = 3$.

The logistic model is defined as follows.

$$P(y_i = 1|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \beta_n^*)}{1 + \exp(\mathbf{x}_i^T \beta_n^*)}.$$

Let $\beta^* = (3, 1.5, 0, 0, 2, \mathbf{0}_{p-5})^T$, $(n, p) = (300, 2000)$ and $\mathbf{x} \sim N(\mathbf{0}_p, \Sigma)$ where $\Sigma_{ij} = 0.5^{|i-j|}, 1 \leq i, j \leq p$.

We sampled $(\mathbf{x}, y)$ 100 times, so the all reported results are average of 100 simulation results.

The tuning parameter λ_n is chosen by cross-validation(CV) and generalized information criterion(GIC) in Wang et al. [2013]. For cross-validation, 5-fold cross validation is used to select the tuning parameter.

Methods	TP	FP	TM	Misclassification rate
Oracle	3.00	0.00	1.00	0.122
LASSO(CV)	3.00	47.52	0.00	0.139
SCAD(CV)	2.14	4.16	0.05	0.162
MCP(CV)	2.17	4.61	0.04	0.158
New-SCAD(CV)	2.96	1.02	0.55	0.130
LASSO(GIC)	2.99	15.23	0.00	0.136
SCAD(GIC)	2.82	0.82	0.62	0.152
MCP(GIC)	2.94	1.03	0.42	0.129
New-SCAD(GIC)	2.99	0.15	0.95	0.126

Table 4.1: Simulation study : $q_n = 3, d_n^* = 3$

- TP(True Positive) : the average number of nonzero coefficients correctly estimated to be nonzero.
- FP(False Positive) : the average number of zero coefficients incorrectly estimated to be nonzero.
- TM : the proportion of the true model being exactly identified

In table 4.1, our proposed method (New-SCAD(GIC)) is better than any other methods in the sense of misclassification rate. Also, our proposed method is the best in the sense of the proportion of the true model being exactly identified. Hence, our proposed method can nicely identify the true signal and estimate regression coefficients of covariates.

In the next simulation study, we increase the size of q_n.

For $q_n = 5$, let $\beta^* = (3, 1.5, 0, 0, 2, 0, 2.5, 1, \mathbf{0}_{p-8}^T)^T$.

All other settings are same with the above simulation study.

Methods	TP	FP	TM	Misclassification rate
Oracle	5.00	0.00	1.00	0.147
LASSO(CV)	5.00	42.17	0.00	0.171
SCAD(CV)	4.29	7.31	0.04	0.179
MCP(CV)	4.21	5.83	0.04	0.182
New-SCAD(CV)	4.82	1.23	0.42	0.173
LASSO(GIC)	4.97	13.26	0.00	0.177
SCAD(GIC)	4.85	1.12	0.43	0.170
MCP(GIC)	4.65	0.91	0.37	0.156
New-SCAD(GIC)	4.98	0.38	0.84	0.165

Table 4.2: Simulation study : $q_n = 5, d_n^* = 3$

In table 4.2, our proposed method is the best result except the MCP in the sense of misclassification rate. It means that MCP was better than our method in the sense of misclassification rate. However, in the sense of the proportion of the true model being exactly identified, our proposed method was the definitely best.

In the last simulation study, we decrease the size of d_n^*.

For $d_n^* = 2$, let $\beta^* = (2, 1, 0, 0, 1.5, \mathbf{0}_{p-5}^T)^T$.

All other settings are same with the above simulation study.

Methods	TP	FP	TM	Misclassification rate
Oracle	3.00	0.00	1.00	0.139
LASSO(CV)	3.00	44.14	0.00	0.147
SCAD(CV)	2.32	4.16	0.05	0.165
MCP(CV)	2.45	4.61	0.04	0.168
New-SCAD(CV)	2.81	1.57	0.48	0.151
LASSO(GIC)	3.00	23.17	0.00	0.156
SCAD(GIC)	2.88	0.97	0.59	0.162
MCP(GIC)	2.92	1.53	0.39	0.154
New-SCAD(GIC)	2.99	0.21	0.94	0.146

Table 4.3: Simulation study : $q_n = 3, d_n^* = 2$

In table 4.3, our proposed method (New-SCAD(GIC)) is better than any other methods in the sense of misclassification rate. Also, our proposed method is the best in the sense of the proportion of the true model being exactly identified. Hence, our proposed method can nicely identify the true signal and estimate regression

coefficients of covariates.

In the all cases, our proposed method is the definitely best in the sense of the proportion of the true model being exactly identified. It is consistent with our theoretical results.

4.2 Real data analysis

To demonstrate the application, we analyze the lung cancer dataset of Huang et al. [2016]. In the lung cancer dataset, the number of covariates $p = 22401$ and the number of samples $n = 164$.

As same with the simulation studies, we compare the results of LASSO, SCAD, MCP, and our proposed method. Since we do not know the knowledge of the true underlying model, the oracle MLE cannot be calculated.

The SCAD estimator is obtained by the original CCCP algorithm without calibration. The MCP estimator is obtained with $a = 3$.

The tuning parameter λ_n is chosen by cross-validation(CV) and generalized information criterion(GIC) in Wang et al. [2013]. For cross-validation, 5-fold cross validation is used to select the tuning parameter.

The results are based on 100 random partitions of the original dataset. Since we do not know the knowledge of the true underlying model, TP, FP, TM in the simulation study cannot be calculated. Instead, we report the average model size for 100 random partitions.

Method	Average model size	Misclassification rate
LASSO(CV)	14.8	0.031
SCAD(CV)	10.7	0.037
MCP(CV)	12.1	0.028
New-SCAD(CV)	13.9	0.038
LASSO(GIC)	13.7	0.035
SCAD(GIC)	10.9	0.047
MCP(GIC)	12.2	0.029
New-SCAD(GIC)	12.9	0.041

Table 4.4: The results for the lung cancer dataset

In table 4.4, the result of our proposed method is not good in the sense of misclassification rate. The LASSO and the MCP are better than our proposed method. Instead, our proposed method is based on the SCAD penalty, and the result of our proposed method is better than the original SCAD with CCCP algorithm. Hence, we can conclude that the estimator with the calibrated CCCP is more accurate than the estimator with the original CCCP algorithm.

Average model size is related to sparsity. As well known, the LASSO selects variables more than any other methods. Our pro-

posed method selects variables more than the MCP and the SCAD.

Chapter 5

Conclusion

In this thesis, we propose the calibrated CCCP algorithm for logistic model. Since the loss function for logistic model is not quadratic, we apply the MLQA-CCCP algorithm [Lee et al., 2016] for the penalized objective function.Also, we show that the calibrated CCCP algorithm for logistic model finds the oracle estimator as the unique local minimum with probability tending to 1. Furthermore, we extend the theoretical result to the case of Huber loss instead of the logistic loss.

In the future work, we will extend this result to general convex loss. Also, our theoretical result is based on the unknown value of the tuning parameter λ_n. Hence, we will prove the theoretical

result for the tuning parameter λ_n selected by generalized information criterion(GIC).

Bibliography

Peter J Bickel, Ya'acov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

T Tony Cai, Lie Wang, and Guangwu Xu. Shifting inequality and recovery of sparse signals. *IEEE Transactions on Signal Processing*, 58(3):1300–1308, 2009a.

T Tony Cai, Guangwu Xu, and Jun Zhang. On recovery of sparse signals via l1 minimization. *IEEE Transactions on Information Theory*, 55(7):3388–3397, 2009b.

Tony Tony Cai, Lie Wang, and Guangwu Xu. Stable recovery of sparse signals and an oracle inequality. *IEEE Transactions on Information Theory*, 56(7):3516–3522, 2010.

E Candes, J Romberg, and T Tao. Stable signal recovery from

imcomplete and inaccurate information. *Commun. Pure Appl. Math*, 59:1207–1233, 2005.

Emmanuel Candes and Terence Tao. Decoding by linear programming. *arXiv preprint math/0502327*, 2005.

Emmanuel Candes, Terence Tao, et al. The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351, 2007.

Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2005.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2): 407–499, 2004.

Jianqing Fan and Runze Li. Variable selection via nonconcave

penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

Hai-Hui Huang, Xiao-Ying Liu, and Yong Liang. Feature selection and cancer classification via sparse logistic regression with the hybrid l1/2+ 2 regularization. *PloS one*, 11(5):e0149675, 2016.

Jinseog Kim, Yuwon Kim, and Yongdai Kim. A gradient-based optimization algorithm for lasso. *Journal of Computational and Graphical Statistics*, 17(4):994–1009, 2008a.

Yongdai Kim and Sunghoon Kwon. Global optimality of nonconvex penalized estimators. *Biometrika*, 99(2):315–325, 2012.

Yongdai Kim, Hosik Choi, and Hee-Seok Oh. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673, 2008b.

Sunghoon Kwon and Yongdai Kim. Large sample properties of the scad-penalized maximum likelihood estimation on high dimensions. *Statistica Sinica*, pages 629–653, 2012.

Sangin Lee, Sunghoon Kwon, and Yongdai Kim. A modified local quadratic approximation algorithm for penalized optimization problems. *Computational Statistics & Data Analysis*, 94:275–286, 2016.

Mee Young Park and Trevor Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over lq-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

Saharon Rosset, L Saul, Y Weiss, and L Bottou. Tracking curved regularized optimization solution paths. Citeseer, 2004.

Robert Tibshirani. Regression shrinkage and selection via the

lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Sara A Van de Geer et al. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

Lan Wang, Yongdai Kim, and Runze Li. Calibrating non-convex penalized regression in ultra-high dimension. *Annals of statistics*, 41(5):2505, 2013.

Fei Ye and Cun-Hui Zhang. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *Journal of Machine Learning Research*, 11(Dec):3519–3540, 2010.

Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. An improved glmnet for l1-regularized logistic regression. *Journal of Machine Learning Research*, 13(Jun):1999–2030, 2012.

Cun-Hui Zhang et al. Nearly unbiased variable selection under

minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Peng Zhao and Bin Yu. Boosted lasso. Technical report, CALIFORNIA UNIV BERKELEY DEPT OF STATISTICS, 2004.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

국문초록

고차원 선형회귀분석에서 벌점화 회귀 방법은 추정과 변수선택을 동시에 하는 방법이다. 라소는 벌점화 회귀 방법의 한가지로, 그 해를 구하기 쉽다는 장점이 있으나 변수선택 일치성을 만족하기 어렵다. MCP와 SCAD 등과 같은 비볼록 벌점화 회귀 방법은 변수선택 일치성을 포함한 신의 성질을 가진다. 그러나 비볼록 벌점화 회귀에서 전역 최적해의 직접적인 계산이 어려워 신의 추정량을 구하기가 어렵다. 한편, 조정된 CCCP 알고리즘으로 구한 유일한 국소최소해는 신의 추정량이 된다는 이론적 사실이 알려져있다. 본 학위논문에서는 로지스틱 모형에 대한 조정된 CCCP 알고리즘을 제안한다. 그리고 로지스틱 모형에서, 조정된 CCCP 알고리즘으로 계산된 해가 1로 향해가는 확률로 신의 추정량이 됨을 증명한다. 로지스틱모형에서는 손실함수가 2차함수가 아니기 때문에, MLQA-CCCP 알고리즘으로 그 해를 반복적으로 계산하였다. 또한, 로지스틱 손실함수를 확장하여 Huber 손실함수에서도 같은 결과가 성립함을 증명한다. 본 학위

논문의 수치 실험들은 이론적 결과들을 뒷받침한다.

주요어: 고차원 회귀분석, 벌점화 회귀, 로지스틱 손실함수, Huber 손실함수, 변수 선택, 신의 추정량, MCP, SCAD

학 번: 2012–20238