2018

THE FAMILY OF CONDITIONAL PENALIZED METHODS WITH THEIR APPLICATION IN SUFFICIENT VARIABLE SELECTION

Jin Xie
University of Kentucky, ginjinx@gmail.com
Author ORCID Identifier:
https://orcid.org/0000-0003-4818-5959
Digital Object Identifier: https://doi.org/10.13023/etd.2018.484

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Jin Xie, Student

Dr. Xiangrong Yin, Major Professor

Dr. Constance Wood, Director of Graduate Studies

THE FAMILY OF CONDITIONAL PENALIZED METHODS WITH THEIR APPLICATION IN SUFFICIENT VARIABLE SELECTION

DISSERTATION

A dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of
Philosophy in the College of Arts and Sciences
at the University of Kentucky

By
Jin Xie
Lexington, Kentucky
Director: Dr. Xiangrong Yin, Professor of Statistics
Lexington, Kentucky

2018

ABSTRACT OF DISSERTATION

THE FAMILY OF CONDITIONAL PENALIZED METHODS WITH THEIR APPLICATION IN SUFFICIENT VARIABLE SELECTION

When scientists know in advance that some features (variables) are important in modeling a data, then these important features should be kept in the model. How can we utilize this prior information to effectively find other important features? This dissertation is to provide a solution, using such prior information. We propose the Conditional Adaptive Lasso (CAL) estimates to exploit this knowledge. By choosing a meaningful conditioning set (prior information), CAL shows better performance in both variable selection and model estimation. We then extend to the linear model setup to the generalized linear models (GLM). Instead of least squares, we consider the likelihood function with L_1 penalty. We proposed for Generalized Conditional Adaptive Lasso (GCAL) for GLMs. We further extend the method for any penalty terms that satisfy certain regularity conditions, namely Conditionally Penalized Estimate (CPE). Asymptotic and oracle properties are showed. Four corresponding sufficient variable screening algorithms are proposed. Simulation examples are evaluated for our method with comparisons with existing methods. GCAL is also evaluated with a read data set on leukemia.

KEYWORDS: Generalized Conditional Adaptive Lasso, High-dimensional Data, Variable Screening, Variable Selection

Author's signature: Jin Xie

Date: December 14, 2018

THE FAMILY OF CONDITIONAL PENALIZED METHODS WITH THEIR APPLICATION IN SUFFICIENT VARIABLE SELECTION

By
Jin Xie

Director of Dissertation: Dr. Xiangrong Yin

Director of Graduate Studies: Dr. Constance Wood

Date: December 14, 2018

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. Xiangrong Yin. I have been very fortunate to be one of his students. Without his dedicated guidance, this dissertation could never be finished. His passion, diligence and enthusiasm on statistics infect every student including me and motivate us to better our research. His persistent guidance, endless support and encouragement have helped me overcome many obstacles throughout my entire Ph.D. learning process.

I greatly appreciate the help of all my dissertation committee members: Dr. Arnold Stromberg, Dr. Solomon Harrar and Dr. Katherine Thompson and Dr. Chi Wang. Each of them have provided lots of insights, help and support for my research.

I would like to thank our department chair, Dr. Arnold Stromberg, for supporting me to attend conferences and meetings. His thought that he supports whatever is best for the students infects me. His meticulous care, concern and support on every student in the department creates not only the best learning/research environment, but also a comfortable living environment here, especially for foreign students.

I would like to thank Dr. Kristen McQuerry, the precious experience at the Applied Statistics Laboratory has a significant impact on my career.

I am thankful to Department of Statistics for the consistent support during my years at UKY, and to all the faculty, staff, peers and friends, who are always ready to help and are pleasant to work with.

Finally, special thanks to my parents Zhenjie Xie and Ping Wu for their unconditional love and endless support.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1 Introduction

1.1 Big Data and High Dimensional Data

Today, the cost of collecting data has become unimaginably cheap. A researcher can easily gather huge amount of data from many different areas such as internet traffic, financial market, DNA microarrays and etc. Data collection can be relentless. Stock market, traffic camera or internet browsing content can generate incredible amount of data within a second. The big data is so large that it becomes very difficult to process or analyze using traditional statistical models. The big data is different from the traditional data set not only in its massive volume, but also in its high dimensionality. Typical statistical methods fail to work appropriately on high-dimensional data. For example, in gene expression data analysis, we cannot even run a linear regression for cancer status on genes since the number of predictors is much larger than the sample size. How to appropriately analyze the high dimensional data has been a huge challenge for modern data scientists. There are two recently developed main stream techniques to deal with such problems—penalization and screening. In this dissertation, we propose new methods/tools to deal with both penalization and screening methods for high-dimensional data analysis.

1.2 Penalization and Screening Methods

Penalization methods typically add a penalty term after certain target function such as negative log-likelihood or residual sum of squares (RSS) etc. Akaike (1973, 1974) propose the Akaike Information Criterion (AIC) which uses the number of predictor

as the penalty term. The definition is

$$\mathrm{AIC}(\mathcal{M}) = -2 \log \mathrm{Lik}(\mathcal{M}) + 2 \cdot p(\mathcal{M}),$$

where $\mathrm{Lik}(\mathcal{M})$ is the likelihood function of the parameters in model $\mathcal{M}$ and $p(\mathcal{M})$ is the number of predictors in $\mathcal{M}$ or the degrees of freedom used up by the model. Along the line, Schwarz et al. (1978) propose the Bayesian Information Criterion (BIC) from the perspective of Bayesian approach. The BIC is defined as

$$\mathrm{BIC}(\mathcal{M}) = -2 \log \mathrm{Lik}(\mathcal{M}) + \log n \cdot p(\mathcal{M}).$$

BIC is similar to AIC except for the $\log n$ penalty coefficient instead of the fixed coefficient 2 in the AIC formula. In fact, AIC and BIC can be viewed as a penalized likelihood methods, where $k \cdot p(\mathcal{M})$ is the penalty term for some $k = 2$ or $\log n$. Many other traditional methods can also be viewed as penalized likelihood methods with different choices of penalty term such as Mallow's Cp (Mallows, 1973), risk inflation criterion (Foster and George, 1994) and Residual Information Criterion (Shi and Tsai, 2002).

More recently, researchers start using L_q ($q \geq 0$) regularizations on coefficients as the penalty term. Best subset selection is actually a L_0 regularization problem. It has nice properties and performance but is computationally inefficient. Especially dealing with high-dimensional data, the best subset selection is computationally infeasible. Frank and Friedman (1993) propose the bridge regression, which is a L_q penalized regression method. Ridge regression is a L_2 regularization problem with closed form solutions. Tibshirani (1996) propose the least absolute shrinkage and selection operator (Lasso) which uses a L_1 regularization as the penalty term. Elastic

net (Zou and Hastie, 2005) uses a linear combination of L_1 and L_2 regularization as the penalty terms. Particularly, Lasso has become a very popular technique since it has the shrinkage ability and computation ease. Due to the inconsistency drawback of Lasso, Zou (2006) propose the Adaptive Lasso which enjoys the oracle properties. The non-negative garrote (Breiman et al., 1996) is another shrinkage method and can be viewed as a special case of Adaptive Lasso. Along the line, Yuan and Lin (2006) introduce the group lasso to enable group shrinkage in L_1 regularization problems. Fan and Li (2001) propose the smoothly clipped absolute deviation (SCAD) penalization, which also enjoys the oracle properties. New comers such as Least Angle Regression (Efron et al., 2004) also shows nice behaviors in both estimation accuracy and computational speed.

Variable screening is another important track of high-dimensional data analysis methods. There is a clear distinction between penalization and screening methods. Penalization method simultaneously select and estimate parameters by solving an optimization problem. The most common variable screening techniques are stepwise algorithms such as forward, backward and bidirectional selections. Fan and Lv (2008) propose the sure independence screening (SIS), which exploits the marginal correlation between each predictor and the response. It is computationally efficient especially for ultra-high dimensional problems. Combining with other penalization method such as Lasso, SCAD, Dantzig selector etc., SIS has become a powerful tool in variable screening. For example, with ultra-high dimensional data, we could use SIS to roughly select $\log n/n$ variables at first and then use other techniques to perform a finer selection based on the reduced selection set. If we use distance correlation (Székely et al., 2007) instead of the Pearson correlation, it becomes DC-SIS (Li et al., 2012). Recently, Barut et al. (2016) propose the conditional sure independence screening (CSIS) method which uses a prior knowledge—a conditioning set of

pre-selected variables (denoted by $\mathcal{C}$). With the help of an appropriate $\mathcal{C}$, CSIS can find other important variables where SIS fails to discover. CSIS has also been showed to enjoy the sure screening properties. Motivated by CSIS, in this dissertation, we exploit the prior information in the penalization method such as Lasso, SCAD and etc.

1.3 Overview of the Dissertation

This dissertation is organized as the following. In Chapter 2, we propose the Conditional Adaptive Lasso (CAL) estimates using the prior information in the original Adaptive Lasso. With the prior information, CAL has a better variable selection result than the original Adaptive Lasso. We also demonstrate that CAL enjoys the oracle properties. A sufficient variable screening method based on CAL is proposed in Chapter 2 as well, namely Sufficient Conditional Adaptive Lasso Variable Screening (SCAL-VS) and Conditioning Set Sufficient Conditional Adaptive Lasso Variable Screening (CS-SCAL-VS) algorithms. In Chapter 3, we further extend CAL from the linear setup to a more general case, that is the generalized linear models (GLM). Generalized Conditional Adaptive Lasso (GCAL) is proposed for generalized linear models. Similarly, GCAL in the generalized linear models has also been demonstrated to enjoy the oracle properties. A corresponding sufficient variable screening algorithm for the generalized linear models is proposed in Chapter 3, that is Sufficient Conditional Adaptive Lasso Variable Screening for GLM (SCAL-VS-G) and Conditioning Set Sufficient Conditional Adaptive Lasso Variable Screening for GLM (CS-SCAL-VS-G). In Chapter 4, we further extend the idea of GCAL to any penalty function satisfying certain conditions. Conditional Penalized Estimate (CPE) is proposed. We then prove the oracle properties of CPE. We propose a suffi-

cient variable screening algorithm, that is Sufficient Conditional Penalized Estimate Variable Screening (SCPE-VS) and Conditioning Set Sufficient Conditional Penalized Estimate Variable Screening (CS-SCPE-VS). Simulation studies and real data are performed to show the appealing properties each method in each chapter. The idea of CPE can be naturally extended to other penalized likelihood problems as new research direction, for instance, survival models or longitudinal data analysis. Proofs of the theorems are deferred to the appendices at the end of each chapter. In summary, this dissertation provides a novel system of theory/methods for conditional penalized estimates, using a loss function and penalty term with a conditional predictor set, leading to a new research direction of statistical modeling and data analysis.

Chapter 2 The Conditional Adaptive Lasso and Its Sufficient Variable Selection Algorithm

2.1 Introduction

More and more massive datasets are coming into the research fields such as genomics and finance. These are the so-called "high-dimensional" data since usually the number of predictors is much larger than the number of observations. One of the most important and difficult tasks for statisticians is to recognize the true active variables from the numerous predictors, especially when the set of truly active variables is small. In such a case, sparse estimates or variable selection methods are very useful.

Penalized approach such as Lasso (Tibshirani, 1996) has been a very popular technique in variable selection and sparse estimates for high-dimensional data. For a given centered continuous response vector $\mathbf{y}$ and an $n \times p$ column-standardized design matrix $\mathbf{X}$, consider the classic linear regression problem,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^{*} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{\beta}^{*} = (\beta_1^*, \ldots, \beta_p^*)^T$ is the true coefficient vector and the error $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2\mathbf{I})$. The Lasso estimates are defined as

$$\hat{\boldsymbol{\beta}}^{(n)}(\mathrm{L}asso) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.2}$$

where λ is a non-negative regularization parameter. The ℓ_1 penalty term in (2.2) is the key for the success of Lasso. The fused Lasso (Tibshirani et al., 2005) encourages

to penalize on the ℓ_1-norm of both the coefficients and their successive difference. Zou and Hastie (2005) then propose the elastic net, where the penalty term is a linear combination of ℓ_1 and ℓ_2 penalties. Unfortunately, Lasso doesn't satisfy the so-called oracle property. The oracle properties state that the estimated coefficients must have asymptotic normality. SCAD (Fan and Lv, 2011) is firstly introduced to enjoy the oracle properties. To remedy the inconsistency of Lasso, Zou (2006) introduces the Adaptive Lasso. It adds a weight parameter in front of each ℓ_1 penalized coefficient. With the added weight, the Adaptive Lasso has been shown to not only enjoy the oracle property, but also that the probability of the non-zero Adaptive Lasso estimates containing the true active set tends to 1 asymtotically. Another generalization of the Lasso is the group Lasso (Yuan and Lin, 2006). It enables Lasso to penalize grouped variables together. Furthermore, the graphical Lasso (Yuan and Lin, 2007; Friedman et al., 2008) makes it able to penalize on the log-likelihood and the inverse covariance matrix. Along the line, Dantzig Selector (Candes and Tao, 2007) solves the ℓ_1-penalization problem with more restrictions on residuals.

The Lasso solves the variable selection problem by optimizing the penalized target function. There has been another track of variable selection technique—screening approach. Fan and Lv (2008) propose the sure independence screening (SIS), which utilizes the marginal correlation, along the line, ISIS is the iterative version of SIS. The idea is further developed on generalized linear models by Fan et al. (2009). Fan and Song (2010) shows the theoretical properties enjoyed by SIS for generalized linear models. Along the line, Li et al. (2012) propose DC-SIS, which is a distance correlation based SIS algorithm. One of the drawbacks of SIS is that it can screen out those variables which have a big impact on response but are weakly correlated with the response. With this background, Barut et al. (2016) propose the conditional sure independence screening (CSIS). It becomes SIS when the conditioning

set is empty. One of the most important advantages of CSIS is to utilize the prior information, namely the conditioning set, say $\mathbf{X}_{\mathcal{C}}$. Conditional screening recruits important additional variables based on the conditional set of variables.

Conditional Adaptive Lasso (CAL) is a natural extension from Adaptive Lasso. It is equivalent to Adaptive Lasso when the conditioning set is empty. How we can exploit the known prior information to improve the Adaptive Lasso estimates is the main logic behind. Indeed, as we will show later, the CAL outperforms Adaptive Lasso and Lasso.

Note that Lasso or Adaptive Lasso estimates the coefficient simultaneously, which may not be sufficient when dealing with ultra-high dimensional data. For large p small n data, we develop the Sufficient Conditional Adaptive Lasso (SCAL) algorithms based on the idea of CAL, utilizing some fitted information, then to deal with the problem of non-sufficiency. Let $\mathbf{X}_{\mathcal{D}}$ be the design matrix excluding the conditioning set. When $p > n$, we decompose the design matrix $\mathbf{X}_{\mathcal{D}}$ into several sub-design matrix $\mathbf{X}_i$ of size p_i ($\sum_i p_i = p$), such that $p_i < p$ for every sub-design matrix $\mathbf{X}_i$. Once we give initial estimates to $\boldsymbol{\beta}_{\mathcal{C}}$, we go through each sub-design matrix $\mathbf{X}_i$ to fit an Adaptive Lasso using the residuals from all previous pieces. By subsetting the design matrix, our strategy is to sufficiently solve the problem within each piece and iterate through all pieces in turns. In addition, combining CAL and SCAL, we develop a sufficient variable screening procedure.

The rest of the chapter is organized as follows. In Section 2.2, we introduce the Conditional Adaptive Lasso (CAL) and prove its oracle properties. The Sufficient Conditional Adaptive Lasso Variable Selection (SCAL-VS) algorithm and the Conditioning Set Sufficient Conditional Adaptive Lasso Variable Selection (CS-SCAL-VS) are proposed in Section 2.3. We use numerical studies to examine the performance of CAL, SCAL-VS and CS-SCAL-VS in Section 2.4. We defer the details of the proofs

to an Appendix.

2.2 Conditional Adaptive Lasso

Definition

Suppose that $\hat{\boldsymbol{\beta}}^{(n)}$ is a root-n-consistent estimator of $\boldsymbol{\beta}^*$. For example, we can use the ordinary least square solution as $\hat{\boldsymbol{\beta}}^{(n)}(\text{ols})$. Let $\gamma > 0$, and define the weight vector as $\hat{\boldsymbol{\omega}} = 1/\left|\hat{\boldsymbol{\beta}}^{(n)}\right|^{\gamma}$. The adaptive Lasso estimates, $\hat{\boldsymbol{\beta}}^{(n)}(\text{adaLasso})$ (Zou, 2006), are given by

$$\hat{\boldsymbol{\beta}}^{(n)}(adaLasso) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\beta_j|, \tag{2.3}$$

where λ_n varies with n.

Without loss of generality, let $\mathcal{C}$ be the index set of the first q conditional variables, that is $\mathcal{C} = \{1, 2, \ldots, q\}$. Let $\mathcal{D}$ be the index set of the remaining $d = p - q$ variables, that is $\mathcal{D} = \{q+1, q+2, \ldots, p\}$. And we will use the notation:

$$\boldsymbol{\beta}_{\mathcal{C}} = (\beta_1, \ldots, \beta_q)^T \in \mathbf{R}^q \qquad \text{and} \qquad \boldsymbol{\beta}_{\mathcal{D}} = (\beta_{q+1}, \ldots, \beta_p)^T \in \mathbf{R}^d.$$

The covariates have been standardized so that $E(X_j) = 0$ and $E(X_j^2) = 1$, for $j = 1, 2, \ldots, p$. Let $\mathcal{A} = \{j \in \mathcal{D} : \beta_j^* \neq 0\} = \{q+1, q+2, \ldots, q+s\}$, that is, the first s variables in $\mathcal{D}$ are active and they constitute the set $\mathcal{A}$. Our setup is similar to the one in Knight and Fu (2000) but a further finer structural assumption, when

$n \to \infty$,

$$\frac{1}{n}\mathbf{X}^T\mathbf{X} = \frac{1}{n}\begin{bmatrix}\mathbf{X}_{\mathcal{C}} & \mathbf{X}_{\mathcal{D}}\end{bmatrix}^T\begin{bmatrix}\mathbf{X}_{\mathcal{C}} & \mathbf{X}_{\mathcal{D}}\end{bmatrix} = \frac{1}{n}\begin{bmatrix}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}} & \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} \\ \mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}} & \mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\end{bmatrix} \to \begin{bmatrix}\Sigma_{\mathcal{C}\mathcal{C}} & \Sigma_{\mathcal{C}\mathcal{D}} \\ \Sigma_{\mathcal{D}\mathcal{C}} & \Sigma_{\mathcal{D}\mathcal{D}}\end{bmatrix}, \tag{2.4}$$

$$\Sigma_{\mathcal{C}\mathcal{D}} = \begin{bmatrix}\Sigma_{\mathcal{C}\mathcal{D}1} & \Sigma_{\mathcal{C}\mathcal{D}2}\end{bmatrix} \qquad and \qquad \Sigma_{\mathcal{D}\mathcal{D}} = \begin{bmatrix}\Sigma_{ss} & \Sigma_{s(d-s)} \\ \Sigma_{(d-s)s} & \Sigma_{(d-s)(d-s)}\end{bmatrix}. \tag{2.5}$$

where $\Sigma_{\mathcal{C}\mathcal{C}}$ is a $q \times q$ matrix, $\Sigma_{\mathcal{C}\mathcal{D}}$ is a $q \times d$ matrix, $\Sigma_{\mathcal{D}\mathcal{D}}$ is a $d \times d$ matrix. Σ_{ss}, $\Sigma_{s(d-s)}$ and $\Sigma_{(d-s)(d-s)}$ are $s \times s$, $s \times (d-s)$ and $(d-s) \times (d-s)$ matrices, respectively. $\Sigma_{\mathcal{C}\mathcal{D}1}$ is a $q \times s$ matrix and $\Sigma_{\mathcal{C}\mathcal{D}2}$ is a $q \times (d-s)$ matrix.

If a set of variables $\mathbf{X}_{\mathcal{C}}$ is given aforehand, we would like to solve for other important variables from the remaining variables, namely $\mathbf{X}_{\mathcal{D}}$, to help better explain the response variable $\mathbf{y}$. Therefore, we propose the CAL estimates as below.

Definition 1. *Suppose that $\hat{\boldsymbol{\beta}}^{(n)}$ is a root-n-consistent estimator of $\boldsymbol{\beta}^*$; for example, we can use $\hat{\boldsymbol{\beta}}^{(n)}(\mathrm{o}ls)$. Pick a $\gamma > 0$, and define the weight vector as $\hat{\boldsymbol{w}} = 1/|\hat{\boldsymbol{\beta}}^{(n)}|^\gamma$. The CAL estimates, $\hat{\boldsymbol{\beta}}^{(n)}(\mathrm{c}al)$, are defined as*

$$\hat{\boldsymbol{\beta}}^{(n)}(\mathrm{c}al) = \underset{\boldsymbol{\beta}_{\mathcal{D}}}{\operatorname{argmin}} \left\| \mathbf{y} - \mathbf{X}_{\mathcal{C}}\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} - \mathbf{X}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}} \right\|^2 + \lambda_n \sum_{j \in \mathcal{D}} \hat{w}_j |\beta_j|, \tag{2.6}$$

where $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$ is the first q elements of the ordinary least square estimates $\hat{\boldsymbol{\beta}}^{(n)}(\mathrm{o}ls)$ and λ_n varies with n.

Note that, in the above definition, the key is the estimate of $\boldsymbol{\beta}_{\mathcal{C}}$, namely $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$. There are many ways to get $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$. For example, we could use methods such as linear regression, ridge regression, Adaptive Lasso and SCAD etc, to estimate on all variables and then only take out the coefficients of the conditional set. We could also

use variable screening methods such as SIS, to select the conditional set at first and then fit a linear regression on the conditional set to get $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$. In the above definition, we require $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$ to be ordinary least square estimates in order to enjoy the asymptotic property. But in reality, the algorithm can work with different estimates of $\boldsymbol{\beta}_{\mathcal{C}}$.

In Definition 1, we try to give an estimate of $\boldsymbol{\beta}_{\mathcal{C}}$ at first. However, we could also estimate $\boldsymbol{\beta}_{\mathcal{C}}$ and $\boldsymbol{\beta}_{\mathcal{D}}$ simultaneously without using any estimates for $\boldsymbol{\beta}_{\mathcal{C}}$. All we need to do is not to penalize on the conditional set. This leads us to propose another simultaneous approach for the CAL estimates.

Definition 2. *Suppose that $\hat{\boldsymbol{\beta}}^{(n)}$ is a root-n-consistent estimator of $\boldsymbol{\beta}^*$; for example, we can use $\hat{\boldsymbol{\beta}}^{(n)}(ols)$. Pick a $\gamma > 0$, and define the weight vector as $\hat{\boldsymbol{w}} = 1/|\hat{\boldsymbol{\beta}}^{(n)}|^{\gamma}$, the estimates of the conditional variables, $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$, and the CAL estimates, $\hat{\boldsymbol{\beta}}^{(n)}(cal)$, are defined as*

$$\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}, \hat{\boldsymbol{\beta}}^{(n)}(cal) = \underset{\boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\beta}_{\mathcal{D}}}{\operatorname{argmin}} \left\| \mathbf{y} - \mathbf{X}_{\mathcal{C}}\boldsymbol{\beta}_{\mathcal{C}} - \mathbf{X}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}} \right\|^2 + \lambda_n \sum_{j \in \mathcal{D}} \hat{w}_j |\beta_j|, \tag{2.7}$$

where λ_n varies with n.

Actually, Definition 2 is the same with Adaptive Lasso except that we do not penalize on the conditional set $\mathcal{C}$. In this chapter, we mainly work with Definition 1 since it performs better. We give Definition 2 to show that if you don't have a prior estimates on the conditional set, you could simultaneously estimate $\boldsymbol{\beta}_{\mathcal{C}}$ and $\boldsymbol{\beta}_{\mathcal{D}}$.

Oracle Properties

In this section, we will show that under a proper choice of λ_n, the CAL estimates enjoy the oracle properties.

Theorem 1. *Suppose that* $\lambda_n/\sqrt{n} \to 0$ *and* $\lambda_n n^{(\gamma-1)/2} \to \infty$. *Let* $\mathcal{A}_n = \{j \in \mathcal{D} : \hat{\boldsymbol{\beta}}_j^{(n)}(cal) \neq 0\}$, *that is, the estimated active set. The CAL estimates under Definition 1 must satisfy the following:*

1. *Consistency in variable selection:* $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$;
2. *Asymptotic normality:* $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(n)}(\mathrm{cal}) - \boldsymbol{\beta}_{\mathcal{A}}^*) \to_d \mathrm{N}(\mathbf{0}, \sigma^2\Sigma_{*\mathcal{A}})$.

where $\Sigma_{*\mathcal{A}}$ *is the upper left* $s \times s$ *corner of* $\Sigma_{*\mathcal{D}}$, *and*

$$\begin{aligned}\Sigma_{*\mathcal{D}} = \Sigma_{\mathcal{D}\mathcal{D}} - 2\Sigma_{\mathcal{D}\mathcal{C}}\Sigma_{11}\Sigma_{\mathcal{C}\mathcal{D}} - 2\Sigma_{\mathcal{D}\mathcal{D}}\Sigma_{12}\Sigma_{\mathcal{D}\mathcal{D}} \\ + \Sigma_{\mathcal{D}\mathcal{C}}(\Sigma_{11}\Sigma_{\mathcal{C}\mathcal{C}}\Sigma_{11} + \Sigma_{11}\Sigma_{\mathcal{C}\mathcal{D}}\Sigma_{12} + \Sigma_{12}\Sigma_{\mathcal{D}\mathcal{C}}\Sigma_{11} + \Sigma_{12}\Sigma_{\mathcal{D}\mathcal{D}}\Sigma_{12})\Sigma_{\mathcal{C}\mathcal{D}},\end{aligned}$$

where

$$\begin{aligned}\Sigma_{11} &= \Sigma_{\mathcal{C}\mathcal{C}}^{-1} + \Sigma_{\mathcal{C}\mathcal{C}}^{-1}\Sigma_{\mathcal{C}\mathcal{D}}\Sigma_{\mathcal{D}|\mathcal{C}}^{-1}\Sigma_{\mathcal{D}\mathcal{C}}\Sigma_{\mathcal{C}\mathcal{C}}^{-1}, \\ \Sigma_{12} &= -\Sigma_{\mathcal{C}|\mathcal{D}}^{-1}\Sigma_{\mathcal{C}\mathcal{D}}\Sigma_{\mathcal{D}\mathcal{D}}^{-1}, \\ \Sigma_{\mathcal{C}|\mathcal{D}} &= \Sigma_{\mathcal{C}\mathcal{C}} - \Sigma_{\mathcal{C}\mathcal{D}}\Sigma_{\mathcal{D}\mathcal{D}}^{-1}\Sigma_{\mathcal{D}\mathcal{C}}, \\ \Sigma_{\mathcal{D}|\mathcal{C}} &= \Sigma_{\mathcal{D}\mathcal{D}} - \Sigma_{\mathcal{D}\mathcal{C}}\Sigma_{\mathcal{C}\mathcal{C}}^{-1}\Sigma_{\mathcal{C}\mathcal{D}}.\end{aligned}$$

Theorem 1 shows that CAL enjoys the oracle properties. The proof of Theorem 1 is given in the Appendix. The oracle properties are also enjoyed by the simultaneous version of CAL under Definition 2 as stated by the following theorem.

Theorem 2. *Suppose that* $\lambda_n/\sqrt{n} \to 0$ *and* $\lambda_n n^{(\gamma-1)/2} \to \infty$. *Let* $\mathcal{A}_n = \{j \in \mathcal{D} : \hat{\beta}^{CAL_j} \neq 0\}$. *The CAL estimates in Definition 2 must satisfy the following:*

1. *Consistency in variable selection:* $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$;

2. Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{A}}(cal) - \boldsymbol{\beta}^{*}_{\mathcal{A}}) \to_d \mathrm{N}(\mathbf{0}, \sigma^2 \Sigma_{*\mathcal{A}})$.

The proof of Theorem 2 is also deferred to the Appendix.

2.3 Sufficient Conditional Adaptive Lasso

For ultra-high dimensional data analysis, one of the biggest challenges is $p >> n$. It's due to the lack of enough information from the sample, as the sample covariance matrix is singular, leading to "insufficient" analyses. In this section, we propose a method to overcome this problem "sufficiently". We separate the predictors into several smaller sets, such that within each set, the number of predictors is smaller than the sample size. Then the large p small n problem can be solved sufficiently and sequentially through each set. To help solve the problem, we first order all the predictors based on the strength of marginal and conditional correlations. Similarly with Li et al. (2012), we use distance correlation (Székely et al., 2007) to calculate the marginal correlations between each X_i and Y and correlations between X_i and X_j for $i \neq j$ conditioning on Y. We first propose a variable screening method—SCAL Variable Selection Algorithm. The algorithms are described below.

SCAL Variable Selection Algorithm (SCAL-VS)

0. First calculate two rankings–marginal distance correlation rankings between X_i and Y and in-between distance correlation rankings between X_i and X_j $(i \neq j)$ conditioning on Y. Combine two rankings by taking out the highest s_0 variables from the in-between correlation rankings and putting them on top of the marginal rankings. Update the design matrix $\mathbf{X}$ by reordering the columns based on the combined rankings. Without loss of generality, we will always use the ordered design matrix hereafter.

1. Separate the ordered predictors into to several sets sequentially, such that each set contains $\lfloor \delta n \rfloor$ $(0 < \delta < 1)$ variables except the last set. The last set has whatever variables left (less than $\lfloor \delta n \rfloor$). Let $\mathbf{X}_1, \ldots, \mathbf{X}_k$ be the separated k sub-design matrices with $p_1, \ldots, p_k$ $(p = \sum_{i=i}^{k} p_i)$ number of predictors, respectively, and $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_k$ be the estimated coefficients corresponding to sub-design matrices $\mathbf{X}_1, \ldots, \mathbf{X}_k$, respectively.

2. Get initial estimates through a sequence of linear regressions. Let $\hat{\boldsymbol{\beta}}_1^{(0)}, \ldots, \hat{\boldsymbol{\beta}}_k^{(0)}$ all be zeros. For m^{th} iteration, set $\hat{\boldsymbol{\beta}}_1^{(m)}, \ldots, \hat{\boldsymbol{\beta}}_k^{(m)}$ equal to $\hat{\boldsymbol{\beta}}_1^{(m-1)}, \ldots, \hat{\boldsymbol{\beta}}_k^{(m-1)}$, respectively. Regress the current residuals, $\mathbf{y} - \sum_{i \neq 1} \mathbf{X}_i \hat{\boldsymbol{\beta}}_i^{(m)}$, on $\mathbf{X}_1$ and update $\hat{\boldsymbol{\beta}}_1^{(m)}$ using the estimated coefficients. Next, regress the current residuals, $\mathbf{y} - \sum_{i \neq 2} \mathbf{X}_i \hat{\boldsymbol{\beta}}_i^{(m)}$, on $\mathbf{X}_2$ and update $\hat{\boldsymbol{\beta}}_2^{(m)}$ using the estimated coefficients. Keep doing this until m_0^{th} iteration, when $\hat{\boldsymbol{\beta}}^{(m_0)}$ and $\hat{\boldsymbol{\beta}}^{(m_0-1)}$ are close enough by some criterion, e.g. ℓ_2-norm. Here, $\hat{\boldsymbol{\beta}}^{(m_0)}$ indicates the estimated coefficient vector of all ordered predictors.

3. Set $\hat{\boldsymbol{\beta}}^{(0)}(scal)$ equal to $\hat{\boldsymbol{\beta}}^{(m_0)}$. For j^{th} iteration, set $\hat{\boldsymbol{\beta}}_1^{(j)}(scal), \ldots, \hat{\boldsymbol{\beta}}_k^{(j)}(scal)$ equal to $\hat{\boldsymbol{\beta}}_1^{(j-1)}(scal), \ldots, \hat{\boldsymbol{\beta}}_k^{(j-1)}(scal)$, respectively. For each fixed λ_n, from the current residuals, $\mathbf{y} - \sum_{i \neq 1} \mathbf{X}_i \hat{\boldsymbol{\beta}}_i^{(j)}(scal)$, fit an Adaptive Lasso model on $\mathbf{X}_1$ with weights $1/|\hat{\boldsymbol{\beta}}_1^{(m_0)}|^{\gamma}$. Update $\hat{\boldsymbol{\beta}}_1^{(j)}(scal)$ using the estimated coefficients. From the current residuals, $\mathbf{y} - \sum_{i \neq 2} \mathbf{X}_i \hat{\boldsymbol{\beta}}_i^{(j)}(scal)$, fit an Adaptive Lasso model on $\mathbf{X}_2$ with weights $1/|\hat{\boldsymbol{\beta}}_2^{(m_0)}|^{\gamma}$. Update $\hat{\boldsymbol{\beta}}_2^{(j)}(scal)$ using the estimated coefficients. Repeat until $\mathbf{X}_k$ is fitted. Use RIC (Shi and Tsai, 2002) to forcibly select less than n variables combining all the pieces. Remove those variables with zero estimated coefficients in $\mathbf{X}_1$ from $\mathbf{X}_1$. Combine $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$ to form the newly updated $\mathbf{X}$ which has fewer columns.

4. Repeat Step 0 to Step 3 until there's no zero estimated coefficients in current $\mathbf{X}_1$ or current $\mathbf{X}_1$ has less than $\lfloor \delta n \rfloor$ columns. Perform an Adaptive Lasso fit with original $\mathbf{y}$ on current $\mathbf{X}_1$ with corresponding weights in $1/|\hat{\boldsymbol{\beta}}^{(m_0)}|^{\gamma}$. Let the $\mathbf{X}_{\mathrm{D}eleted}$ be the design matrix only containing the deleted variable column. Order the columns of $\mathbf{X}_{\mathrm{D}eleted}$ as described in Step 0. Update the design matrix $\mathbf{X}$ by putting the ordered columns in $\mathbf{X}_{\mathrm{D}eleted}$ after the remaining (non-deleted) variable columns in $\mathbf{X}$. Return the column number (rank) of the current $\mathbf{X}$ as the screening ranking of all the variables.

Conditioning Set SCAL Variable Selection Algorithm (CS-SCAL-VS)

The Conditioning Set SCAL Variable Selection Algorithm (CS-SCAL-VS) is almost the same as SCAL-VS except that we incorporate a pre-known set of variables—conditioning set $\mathbf{X}_{\mathcal{C}}$, in the model. Then the remaining variables have a corresponding design matrix $\mathbf{X}_{\mathcal{D}}$. Perform the same Steps 0-4 on $\mathbf{X}_{\mathcal{D}}$ as in SCAL-VS except that before Step 2 and Step 3, first regress the current residuals on $\mathbf{X}_{\mathcal{C}}$ and update $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(m)}$ and $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(j)}(scal)$ correspondingly.

The main logic of SCAL-VS and CS-SCAL-VS is to exploit the idea of CAL iteratively. When estimating $\boldsymbol{\beta}_i$, we treat $\{\mathbf{X}_1, \ldots, \mathbf{X}_k\} \backslash \mathbf{X}_i$—all sub-design matrices except $\mathbf{X}_i$, as the conditional set in CAL and fit an Adaptive Lasso model on $\mathbf{X}_i$. With each p_i less than n, we sufficiently estimate the coefficients separately and sequentially. We decide to use distance correlation in our computation since it's much faster to calculate the conditional correlations between X_i and X_j $(i \neq j)$ and can deal with more complicated situations such as multivariate $\mathbf{y}$ circumstances. The key of SCAL-VS and CS-SCAL-VS algorithm relies on predictor segmentation and deleting scheme. Usually, the more important a variable is, the higher rank or earlier ranking position it would have. It will benefit the estimation if important variables

are ordered at front. However, in our numerical studies, we find that some marginally important variables often appear at the tail of the marginal ranking. This inspires us to introduce the deleting scheme. Deleting scheme is a remedy to the ordering process. It enables those non-important variables to be dropped dynamically as well as important variables to enter $\mathbf{X}_1$ and be picked up at a later time. By forcing to select less than n variables in Step 3, we reinforce the importance of $\mathbf{X}_1$. This will make the deleted variables in $\mathbf{X}_1$ be truly non-important ones. The parameter s_0 is chosen based on the understanding of the dataset. Normally, we don't consider a dataset to contain many conditionally correlated variables. Thus, s_0 is set equal to 3 through out all our numerical studies. And one may adjust s_0 to a different value if more conditionally correlated variables are preferred. The size of the segmented set is also important. We find that $\delta = 85\%$ is the best cutoff for segmenting the predictors. For the weights in Adaptive Lasso, we find that $\gamma = 0.2$ is the best value for our algorithms.

As we will show in the next section, we compare SCAL-VS versus SIS and other variable selection techniques. Similarly, we compare CS-SCAL-VS with CSIS technique. To compare with SIS, using the same criteria in Fan and Lv (2008), we choose the first $n-1$ variables selected by SCAL-VS and report the proportion of containing all important variables. To compare with CSIS, we use the same criteria in Barut et al. (2016), that is to report the median minimum model size (MMMS).

2.4 Numerical Studies

In this section, we compare the SCAL-VS algorithms with existing methods such as SIS, ISIS, CSIS, Lasso and Adaptive Lasso etc. We use R packages **SIS** to calculate SIS and ISIS. R package **glmnet** is used to calculate Lasso and adaptive Lasso. We

also write our own code of CAL, SCAL-VS, SIS, ISIS and CSIS in R.

For *Example 1*, *Example 2* and *Example 3*, we apply our algorithm, namely SCAL-VS, comparing with SIS, ISIS and Lasso. Similarly with SIS and ISIS, to make fair comparison, we select $n-1$ variables and report the accuracy as a percentage of including the truly active set of the important variables. For *Example 3*, when we compare CS-SCAL-VS with CSIS, we use the median of minimum model size (MMMS). In *Example 4*, we use $||\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}^*||^2$ as the evaluation metric.

Simulation Study

Following Fan and Lv (2008), we construct the first 3 examples as below. $X_1, \ldots, X_p$ are p predictors and the error term $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2\mathbf{I})$ is independent of the predictors. $X_1, \ldots, X_p$ with sample size n are drawn from a multivariate normal distribution $\mathrm{N}(0, \Sigma)$, where the covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$ has entries $\sigma_{ii} = 1$, $i = 1, \ldots, p$ and $\sigma_{ij} = \rho$, $i \neq j$. We consider different combinations of (n, p, ρ) with $p = 100, 1000$, $n = 20, 50, 70$, and $\rho = 0, 0.1, 0.5, 0.9$. For each combination, we perform 200 simulations on each example. We report the proportion of containing all important variables in the first $n-1$ variables in our SCAL-VS ranking.

All three examples below have similar settings except that in *Example 2*, we consider $\rho = 0.5$ and introduce X_4, where X_4 is designed to be marginally uncorrelated with Y. In *Example 3*, we introduced X_5 to make it have a small correlation with the response and X_5 has the same proportion of contribution to the response as the noise ε does, but X_5 has even weaker marginal correlation with Y than $X_6, \ldots, X_p$. The three examples are as below.

Example 1:

$$Y = 5X_1 + 5X_2 + 5X_3 + \varepsilon, \tag{2.8}$$

Example 2:

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\rho^{1/2}X_4 + \varepsilon, \tag{2.9}$$

Example 3:

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\rho^{1/2}X_4 + X_5 + \varepsilon. \tag{2.10}$$

Table 2.1 shows the results of *Example 1.* And we can see that SCAL-VS outperform SIS, Lasso and ISIS in all cases. Even in severe cases, such that $p = 1000, n = 20, \rho = 0.9$, SCAL-VS can still capture all the truly important variables 70% times.

Table 2.1: *Example 1.* Accuracy of SIS, Lasso, ISIS and SCAL-VS.

			Results for the following values of ρ:			
p	n	Method	ρ=0	ρ=0.1	ρ=0.5	ρ=0.9
100	20	SIS	0.755	0.855	0.690	0.670
		Lasso	0.970	0.990	0.985	0.870
		ISIS	1	1	1	1
		ISIS_test	0.913	0.897	0.867	0.720
		SCAL-VS	0.970	0.990	0.960	0.950
	50	SIS	1	1	1	1
		Lasso	1	1	1	1
		ISIS	1	1	1	1
		ISIS_test	1	1	1	1
		SCAL-VS	1	1	1	1
1000	20	SIS	0.205	0.255	0.145	0.085
		Lasso	0.340	0.555	0.556	0.220
		ISIS	1	1	1	1
		ISIS_test	0.524	0.517	0.425	0.262
		SCAL-VS	0.600	0.630	0.770	0.700
	50	SIS	0.990	0.960	0.870	0.860
		Lasso	1	1	1	1
		ISIS	1	1	1	1
		ISIS_test	1	1	0.997	0.993
		SCAL-VS	1	1	1	1
	70	SIS	1	0.995	0.970	0.970
		Lasso	1	1	1	1
		ISIS	1	1	1	1
		ISIS_test	1	1	1	1
		SCAL-VS	1	1	1	1

Table 2.2: *Example 2.* Accuracy of SIS, Lasso, ISIS and SCAL-VS.

p	Method	$n = 20$	$n = 50$	$n = 70$
100	SIS	0.025	0.490	0.740
	Lasso	0.000	0.360	0.915
	ISIS	0.425	0.925	0.990
	ISIS_test	0.495	0.987	1
	SCAL-VS	0.840	1	1
1000	SIS	0.000	0.000	0.000
	Lasso	0.000	0.000	0.000
	ISIS	0.030	0.990	0.995
	ISIS_test	0.194	0.890	0.997
	SCAL-VS	0.180	0.970	1

Again, X_4 is designed to be marginally uncorrelated with Y in *Example 2*. In Table 2.2, SCAL-VS outperforms SIS, Lasso and ISIS in all cases except $p = 1000, n = 20$ case (nearly the same in this case). One of the reasons that SIS, Lasso and ISIS fail in extreme cases such as $p = 100, 1000, n = 20$ is due to the lack of information provided by the sample. When $p = 100, 1000, n = 20$, most of the time, even marginally correlated variables X_1, X_2, X_3 can have a small correlation with Y in the real sample. However, SCAL-VS can still capture select the marginally uncorrelated variable X_4 and other variables simultaneously.

Table 2.3: *Example 3.* Accuracy of SIS, Lasso, ISIS and SCAL-VS.

p	Method	$n = 20$	$n = 50$	$n = 70$
100	SIS	0.000	0.285	0.645
	Lasso	0.000	0.310	0.890
	ISIS	0.000	0.430	0.850
	SCAL-VS	0.550	1	1
1000	SIS	0.000	0.000	0.000
	Lasso	0.000	0.000	0.000
	ISIS	0.000	0.000	0.000
	SCAL-VS	0.020	0.880	0.960

Table 2.3 shows the result of *Example 3*. With the variable X_5 having the same

variance as noise, SIS, Lasso and ISIS fail to capture all truly important variables in all $p = 1000$ cases. But SCAL-VS can still capture the true model in $n = 50$ and $n = 70$ cases 88% times and 96% times, respectively.

Example 4:

$$Y = 3X_1 + 3X_2 + 3X_3 + 3X_4 + 3X_5 - 7.5X_6 + \varepsilon, \tag{2.11}$$

where $\rho = 0.5$ is the correlation between X_i and X_j for $i \neq j$. Similarly, *Example 3* and *Example 4*, X_6 is marginally uncorrelated with Y. We choose $n = 20, 50, 70$ and $p = 100, 1000$, respectively. This is the same example as in Barut et al. (2016). We report the median of minimum model size (MMMS) and its standard deviation as comparison with CSIS method, based on 200 simulations.

Table 2.4: *Example 4.* MMMS of CS-SCAL-VS and CSIS.

Conditioning Set	$p = 100$				$p = 1000$			
	$n = 20$		$n = 50$		$n = 50$		$n = 70$	
	CS-SCAL-VS	CSIS	CS-SCAL-VS	CSIS	CS-SCAL-VS	CSIS	CS-SCAL-VS	CSIS
X_1	65 (33)	56 (25)	6 (18)	27 (24)	47 (380)	186 (193)	6 (181)	117 (226)
X_1, X_2	61 (34)	53 (28)	6 (13)	17 (23)	7 (413)	119 (222)	6 (215)	57 (166)
X_1, X_2, X_3	68 (35)	56 (27)	6 (9)	29 (27)	6 (335)	236 (302)	6 (211)	98 (271)
X_1, X_2, X_3, X_4	44 (37)	42 (27)	6 (13)	31 (28)	6 (352)	250 (300)	6 (215)	164 (263)
X_1, X_2, X_3, X_4, X_5	6 (23)	6 (1)	6 (2)	6 (0)	6 (4)	6 (0)	6 (2)	6 (0)

In Table 2.4, CS-SCAL-VS outperforms CSIS in many cases. One exception is $n = 20$ case, where CS-SCAL-VS performs relatively comparable with CSIS. Actually, both methods do not perform well in this case. We also note that CSIS performs slightly better when the conditioning set is $\{X_1, X_2, X_3, X_4, X_5\}$. But in other cases, CS-SCAL-VS outperforms CSIS dramatically. Most of the time, CS-SCAL-VS only need 6 variables to cover the true model.

We now want to compare CAL with Lasso. Since CAL needs a conditional set, in order to make a fair comparison, we propose several revised adaptive Lasso algorithm to overcome this unfairness.

- For ALasso_orig, we first fit Adaptive Lasso on all X_i's and then fit linear regression on those variables with nonzero estimated coefficients.

- For ALasso_orig+cond, similarly with ALasso_orig, but we then fit linear regression on nonzero variables plus the variables in the conditioning set.

- For ALasso_cond, we first fit Adaptive Lasso on $\mathbf{X}_{\mathcal{D}}$, and then fit linear regression on nonzero variables plus the variables in the conditioning set.

In *Example 5*, we use the same settings as in *Example 4*. But we vary the conditioning set and choose a high collinearity parameter, i.e. $\rho = 0.8$. We already show that SCAL algorithm can do well in variable screening. In this example, we try to show that if we use the selected variables from CAL to fit a linear model, we can have a good modeling fitting. We compare our CAL estimates with 3 revised adaptive Lasso methods. For CAL, we first fit CAL algorithm, and then fit nonzero linear regression on nonzero variables. We report the ℓ_2-norm of the difference between the true coefficients and the estimated coefficients from the model fitting, namely, $||\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}^*||^2$. The simulation is performed on 200 data sets.

In Table 2.5, CAL outperforms all three Adaptive Lasso methods. Even if we force to add back the condition set to the original Adaptive Lasso selection results, i.e. ALasso_orig+cond, CAL still shows a better performance. It shows that CAL can correctly utilize the prior information. We can also observe that with more correct prior information, CAL performs better. Containing the nonimportant variables in the conditioning set will somehow harm the analysis. But this hurt is not very severe in $p = 1000, n = 50$ case, which suggests that when n is larger, containing the nonimportant variables in the conditioning set should be fine.

Table 2.5: *Example 5.* $||\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}^*||^2$ of CAL and 3 revised adaptive Lasso.

Conditioning Set	$p = 100, n = 20, \rho = 0.8$			
	ALasso_orig	ALasso_orig+cond	ALasso_cond	CAL
X_1	7.501 (2.373)	7.448 (2.812)	7.790 (3.207)	7.023 (2.303)
$X_1 - X_2$	7.379 (2.404)	6.970 (2.938)	8.115 (2.733)	6.187 (2.822)
$X_1 - X_3$	7.071 (2.466)	6.799 (4.439)	9.727 (7.475)	5.231 (2.696)
$X_1 - X_4$	8.393 (11.39)	6.243 (3.519)	8.614 (6.317)	3.835 (2.492)
X_1 and X_{11}	7.725 (2.121)	7.718 (2.524)	20.482 (87.322)	7.506 (2.388)
$X_1 - X_2$ and $X_{11} - X_{12}$	7.427 (2.122)	8.223 (5.150)	19.296 (21.538)	6.890 (2.659)
$X_1 - X_3$ and $X_{11} - X_{13}$	7.733 (2.268)	8.376 (3.850)	20.861 (14.597)	6.789 (3.015)
$X_1 - X_4$ and $X_{11} - X_{14}$	7.297 (2.432)	9.808 (6.664)	48.053 (88.713)	6.333 (3.756)
Conditioning Set	$p = 1000, n = 50, \rho = 0.8$			
	ALasso_orig	ALasso_orig+cond	ALasso_cond	CAL
X_1	4.700 (1.870)	4.596 (1.772)	5.745 (1.785)	3.602 (1.640)
$X_1 - X_2$	4.600 (1.682)	4.427 (1.620)	6.066 (1.789)	2.918 (1.203)
$X_1 - X_3$	4.512 (1.717)	4.242 (1.524)	6.521 (2.261)	2.598 (1.041)
$X_1 - X_4$	4.504 (1.799)	3.956 (1.494)	6.542 (4.320)	1.987 (1.031)
X_1 and X_{11}	4.375 (1.823)	4.360 (1.766)	9.122 (3.450)	3.648 (1.509)
$X_1 - X_2$ and $X_{11} - X_{12}$	4.475 (1.815)	4.393 (1.706)	13.349 (12.178)	2.963 (1.116)
$X_1 - X_3$ and $X_{11} - X_{13}$	4.640 (1.860)	4.785 (1.949)	21.232 (10.736)	2.788 (1.133)
$X_1 - X_4$ and $X_{11} - X_{14}$	4.718 (1.866)	4.430 (2.022)	26.483 (29.299)	2.465 (0.967)

2.5 Discussion

In this chapter, we propose the CAL for linear models. It enjoys nice properties, that is the oracle properties. With useful prior information, CAL improves the original Adaptive Lasso and shows a better result in both variable selection and model estimation. Numerical studies for different settings shows the performance of CAL, SCAL-VS and CS-SCAL-VS with comparisons versus other existing methods.

Chapter 3 The Generalized Conditional Adaptive Lasso and Sufficient Variable Selection

3.1 Introduction

Xie and Yin (2018) propose Conditional Adaptive Lasso (CAL) and its extension, Sufficient Conditional Adaptive Lasso Variable Selection (SCAL-VS) algorithm, which are powerful tools dealing with variable selection problems on ultra-high dimensional data with linear model. Under the linear model settings, a well selected conditioning set can dramatically help select other important variables and reduce the false positive selections even when covariates are highly correlated. The Generalized Conditional Adaptive Lasso (GCAL) estimates and Sufficient Conditional Adaptive Lasso Variable Selection for Generalized Linear Models (SCAL-VS-G) algorithm are proposed in this chapter as extensions of CAL estimates and SCAL-VS algorithm to a more general environment, namely, the generalized linear model settings. GCAL enjoys the oracle properties and can outperform several current methods, as shown in numerical studies.

3.2 Generalized Conditional Adaptive Lasso

Lasso (Tibshirani, 1996) has been applied to several different models such as linear model, generalized linear model, cox proportional model and poisson model etc. The Adaptive Lasso proposed by Zou (2006) has been shown to enjoy the oracle properties in both linear models and generalized linear models (GLMs). In this chapter, we further extend our CAL and SCAL-VS to GLMs. And we will show that

GCAL also enjoys the oracle properties.

We adopt the same setting of GLM as in McCullagh and Nelder (1989). For a canonical parameter $\theta = \mathbf{x}^T\boldsymbol{\beta}^*$, generalized linear models consider that the generic density belongs to an exponential family

$$f(y|\mathbf{x}, \theta) = h(y)\exp(y\theta - \phi(\theta)). \tag{3.1}$$

Without loss of generality, let $\mathcal{C}$ be the index set of the first q conditional variables, that is $\mathcal{C} = \{1, 2, \ldots, q\}$. Let $\mathcal{D}$ be the index set of the remaining $d = p - q$ variables, that is $\mathcal{D} = \{q+1, q+2, \ldots, p\}$. And we will use the notation:

$$\boldsymbol{\beta}_{\mathcal{C}} = (\beta_1, \ldots, \beta_q)^T \in \mathbf{R}^q \qquad \text{and} \qquad \boldsymbol{\beta}_{\mathcal{D}} = (\beta_{q+1}, \ldots, \beta_p)^T \in \mathbf{R}^d.$$

The covariates have been standardized so that $E(X_j) = 0$ and $E(X_j^2) = 1$, for $j = 1, 2, \ldots, p$. Let $\mathcal{A} = \{j \in \mathcal{D} : \beta_j^* \neq 0\} = \{q+1, q+2, \ldots, q+s\}$, that is, the first s variables in $\mathcal{D}$ are active and they constitute the set $\mathcal{A}$. Let $\mathcal{A}^c = \{j \in \mathcal{D} : j \notin \mathcal{A}\}$. Suppose that $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the response vector and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, $i = 1, \ldots, n$, is the ith sample. Let $\mathbf{X}$ be the design matrix. We also have the following assumption about the Fisher information matrix,

$$\mathbf{I}(\boldsymbol{\beta}^*) = \begin{bmatrix} \mathbf{I}_{\mathcal{C}\mathcal{C}} & \mathbf{I}_{\mathcal{C}\mathcal{D}} \\ \mathbf{I}_{\mathcal{D}\mathcal{C}} & \mathbf{I}_{\mathcal{D}\mathcal{D}} \end{bmatrix}, \tag{3.2}$$

$$\mathbf{I}_{\mathcal{C}\mathcal{D}} = \begin{bmatrix} \mathbf{I}_{\mathcal{C}\mathcal{D}1} & \mathbf{I}_{\mathcal{C}\mathcal{D}2} \end{bmatrix} \qquad \text{and} \qquad \mathbf{I}_{\mathcal{D}\mathcal{D}} = \begin{bmatrix} \mathbf{I}_{ss} & \mathbf{I}_{s(d-s)} \\ \mathbf{I}_{(d-s)s} & \mathbf{I}_{(d-s)(d-s)} \end{bmatrix}. \tag{3.3}$$

where $\mathbf{I}_{\mathcal{C}\mathcal{C}}$ is a $q \times q$ matrix, $\mathbf{I}_{\mathcal{C}\mathcal{D}}$ is a $q \times d$ matrix, $\mathbf{I}_{\mathcal{D}\mathcal{D}}$ is a $d \times d$ matrix. $\mathbf{I}_{ss}$, $\mathbf{I}_{s(d-s)}$ and $\mathbf{I}_{(d-s)(d-s)}$ are $s \times s$, $s \times (d-s)$ and $(d-s) \times (d-s)$ matrices, respectively. $\mathbf{I}_{\mathcal{C}\mathcal{D}1}$

is a $q \times s$ matrix and $\mathbf{I}_{\mathcal{C}\mathcal{D}2}$ is a $q \times (d-s)$ matrix. With the above setup, we now give the definition of GCAL below.

Definition 3. *Suppose that $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$ is a root-n-consistent estimator of $\boldsymbol{\beta}_{\mathcal{C}}^{*}$. We assume that $\hat{\boldsymbol{\beta}}(\text{mle})$ is the maximum likelihood estimates in the GLM. The weight vector is constructed as $\hat{\boldsymbol{\omega}} = 1/|\hat{\boldsymbol{\beta}}(\text{mle})|^{\gamma}$ for some $\gamma > 0$. The GCAL estimates $\hat{\boldsymbol{\beta}}^{(n)}(\text{gcal})$ are given by*

$$\hat{\boldsymbol{\beta}}^{(n)}(\text{gcal}) = \underset{\boldsymbol{\beta}_{\mathcal{D}}}{\text{argmin}} \sum_{i=1}^{n} \Big(-y_i \big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}} \big) + \phi \big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}} \big) \Big) + \lambda_n \sum_{j \in \mathcal{D}} \hat{w}_j |\beta_j|, \tag{3.4}$$

where λ_n varies with n.

For logistic regression, Definition 3 becomes

$$\hat{\boldsymbol{\beta}}^{(n)}(\text{logistic}) = \text{argmin}_{\boldsymbol{\beta}_{\mathcal{D}}} \sum_{i=1}^{n} \Big(- y_i \big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}} \big) + \log \big(1 + e^{\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}}} \big) \Big) + \lambda_n \sum_{j \in \mathcal{D}} \hat{w}_j |\beta_j|. \tag{3.5}$$

In Poisson log-linear regression models, Definition 3 can be written as

$$\hat{\boldsymbol{\beta}}^{(n)}(\text{poisson}) = \underset{\boldsymbol{\beta}_{\mathcal{D}}}{\text{argmin}} \sum_{i=1}^{n} \Big(- y_i \big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}} \big) + e^{\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}}} \Big) + \lambda_n \sum_{j \in \mathcal{D}} \hat{w}_j |\beta_j|. \tag{3.6}$$

We prove that GCAL estimates enjoy the oracle properties as below.

Theorem 3. *Suppose that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$. Let $\mathcal{A}_n = \{j \in \mathcal{D} : \hat{\boldsymbol{\beta}}_j^{(n)}(\text{gcal}) \neq 0\}$, that is, the estimated active set. The GCAL estimates under Definition 3 must satisfy the following:*

1. *Consistency in variable selection:* $\lim_{n \to \infty} P(\mathcal{A}_n = \mathcal{A}) = 1$;
2. *Asymptotic normality:* $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(n)}(\text{gcal}) - \boldsymbol{\beta}_{\mathcal{A}}^{*}) \to_d \text{N}(\mathbf{0}, \mathbf{I}_{ss}^{-1})$.

The detail of the proof of Theorem 3 is deferred to the Appendix.

Theorem 3 shows GCAL estimates enjoy the oracle properties. Similar with CAL, we can simultaneously estimate $\boldsymbol{\beta}_{\mathcal{C}}$ and $\boldsymbol{\beta}_{\mathcal{D}}$ in GCAL if we do not use the previously estimated $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$ as the values of $\boldsymbol{\beta}_{\mathcal{C}}$. The definition can be found below.

Definition 4. *Suppose that $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$ is a root-n-consistent estimator of $\boldsymbol{\beta}_{\mathcal{C}}^{*}$. We assume that $\hat{\boldsymbol{\beta}}(\text{mle})$ is the maximum likelihood estimates in the GLM. The weight vector is constructed as $\hat{\boldsymbol{\omega}} = 1/|\hat{\boldsymbol{\beta}}(\text{mle})|^{\gamma}$ for some $\gamma > 0$. The GCAL estimates $\hat{\boldsymbol{\beta}}^{(n)}(\text{gcal})$ are given by*

$$\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}, \hat{\boldsymbol{\beta}}^{(n)}(\text{gcal}) = \underset{\boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\beta}_{\mathcal{D}}}{\operatorname{argmin}} \sum_{i=1}^{n} \Big(-y_i \big(\mathbf{x}_{i\mathcal{C}}^T \boldsymbol{\beta}_{\mathcal{C}} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}} \big) + \phi \big(\mathbf{x}_{i\mathcal{C}}^T \boldsymbol{\beta}_{\mathcal{C}} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}} \big) \Big) + \lambda_n \sum_{j \in \mathcal{D}} \hat{w}_j |\beta_j|, \tag{3.7}$$

where λ_n varies with n.

We also show that GCAL under Definition 4 enjoy the oracle properties.

Theorem 4. *Suppose that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$. Let $\mathcal{A}_n = \{j \in \mathcal{D} : \hat{\boldsymbol{\beta}}_j^{(n)}(\text{gcal}) \neq 0\}$, that is, the estimated active set. The GCAL estimates under Definition 4 must satisfy the following:*

1. *Consistency in variable selection:* $\lim_{n \to \infty} P(\mathcal{A}_n = \mathcal{A}) = 1$;
2. *Asymptotic normality:* $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(n)}(\text{gcal}) - \boldsymbol{\beta}_{\mathcal{A}}^{*}) \to_d \mathrm{N}(\mathbf{0}, \mathbf{I}_{ss}^{-1})$.

The proof of Theorem 4 is also deferred to the Appendix.

3.3 Numerical Optimization for GCAL

To calculate the entire regularization path for the lasso, people have proposed many different algorithms. One of the most efficient algorithms to compute the regular-

ization path for the lasso in linear models is proposed by Efron et al. (2004). Their algorithm is based on the fact that the coefficients behave piece-wise linearly along the regularization path. Rosset and Zhu (2007) rigorously discuss the conditions under which the piece-wise linearity exists.

Cyclical coordinate descent optimization method has been proposed for a while. Tseng (2001) has discussed the convergence criteria of coordinate descent algorithm for nondifferentiable minimization problems. The coordinate descent algorithm has not received full appreciation until recent time. In the recent re-visit by Friedman et al. (2007), they proposed to use the current estimates as warm-up starts for the next smaller value of regularization parameters along the regularization path. This strategy turns out to be extremely efficient. Friedman et al. (2010) has extended the coordinate descent algorithm for the lasso to the generalized linear models. The widely used R package **glmnet** (Friedman et al., 2016) is based on this coordinate descent algorithm. Our computations for the GCAL and GCAL-VS also exploit the coordinate descent algorithm. We alter the source C code in the R package **ncvreg** (Breheny and Huang, 2011) to form our own algorithm for the GCAL and GCAL-VS.

Suppose we have n observations and we standardize the column of the design matrix $\mathbf{X}$ such that $\sum_{i=1}^{n} x_{ij} = 0$ and $\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, \ldots, p$. For a typical linear situation, we are trying to solve the following problem

$$\min_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{p+1}} R_\lambda(\beta_0, \boldsymbol{\beta}) = \min_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{p+1}} \left[\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\hat{w}_j|\beta_j|\right]. \quad (3.8)$$

For a coordinate descent step, suppose we have current estimates $\tilde{\beta}_0$ and $\tilde{\beta}_\ell$ for $(\ell \neq j)$, we would like to partially optimize (3.8) with respect to β_j. We need to calculate the gradient at $\beta_j = \tilde{\beta}_j$, which only exists if $\tilde{\beta}_j \neq 0$. For $\tilde{\beta}_j > 0$, the partial

derivative is

$$\left.\frac{\partial R_\lambda}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} = -\sum_{i=1}^{n} x_{ij}(y_i - \tilde{\beta}_0 - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}}) + \lambda\hat{w}_j. \tag{3.9}$$

Similar partial derivative can be calculated if $\tilde{\beta}_j < 0$. Then the coordinate-wise update has the following form

$$\tilde{\beta}_j \leftarrow S\Big(\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda\hat{w}_j\Big), \tag{3.10}$$

where

- $\tilde{y}_i^{(j)} = \tilde{\beta}_0 + \sum_{\ell\neq j} x_{i\ell}\tilde{\beta}_\ell$ is the i^{th} fitted value excluding the contribution of x_{ij}. Hence, $y_i - \tilde{y}_i^{(j)}$ is the i^{th} partial residual for fitting β_j.
- $S(z,\gamma)$ is the soft-thresholding operator

$$S(z,\gamma) = \text{sign}(z)(|z| - \gamma)_+ = \begin{cases} z-\gamma & \text{if } z>0 \text{ and } \gamma<|z| \\ z+\gamma & \text{if } z<0 \text{ and } \gamma<|z| \\ 0 & \text{otherwise.} \end{cases} \tag{3.11}$$

If the observation weight w_i $(i = 1, \ldots, n)$ is assigned, then the update naturally becomes

$$\tilde{\beta}_j \leftarrow \frac{S\Big(\sum_{i=1}^{n} w_i x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda\hat{w}_j\Big)}{\sum_{i=1}^{n} w_i x_{ij}^2}, \tag{3.12}$$

which is used to solve the iteratively reweighted least squares (IRLS) problems.

For penalized generalized linear model problems, we here use the penalized logistic regression as an example to illustrate the coordinate descent algorithm. For the

penalized logistic regression, we are trying to solve the problem,

$$\min_{\boldsymbol{\beta}} \Big\{ \sum_{i=1}^{n} \Big[- y_i\big(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}\big) + \log\big(1 + e^{\beta_0+\mathbf{x}_i^T\boldsymbol{\beta}}\big)\Big] + \lambda \sum_{j=1}^{p} \hat{w}_j|\beta_j| \Big\}. \tag{3.13}$$

The negative log-likelihood can be written as

$$\ell(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \Big[- y_i\big(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}\big) + \log\big(1 + e^{\beta_0+\mathbf{x}_i^T\boldsymbol{\beta}}\big)\Big], \tag{3.14}$$

a convex function of β_0 and $\boldsymbol{\beta}$. We are quite familiar that minimizing (3.14), the unpenalized negative log-likelihood, can be done with the iteratively reweighted least squares (IRLS) algorithm, a Newton method. Once we apply the current estimates to (3.14), we can get a quadratic approximation of the negative log-likelihood using Taylor expansion,

$$\ell_Q(\beta_0, \boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n} w_i(z_i - \beta_0 - \mathbf{x}_i^T\boldsymbol{\beta})^2 + C(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}), \tag{3.15}$$

where

- $z_i = \tilde{\beta}_0 + \mathbf{x}_i^T\tilde{\boldsymbol{\beta}} + \dfrac{y_i - \tilde{p}(\mathbf{x}_i)}{\tilde{p}(\mathbf{x}_i)(1-\tilde{p}(\mathbf{x}_i))}$ is the working response.
- $w_i = \tilde{p}(\mathbf{x}_i)(1-\tilde{p}(\mathbf{x}_i))$ is the weight.

and $\tilde{p}(\mathbf{x}_i) = 1/(1+\exp(-\tilde{\beta}_0 - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}}))$ is evaluated at current estimates. $C(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}})$ is a constant. Then the optimization of (3.13) can be solved by optimizing

$$\min_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{p+1}} \Big[\ell_Q(\beta_0, \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \hat{w}_j|\beta_j|\Big]. \tag{3.16}$$

Then the coordinate descent algorithm can be applied here.

Based on the coordinate descent algorithm, we propose our numerical optimization algorithm for the GCAL.

Numerical Optimization Algorithm for GCAL (Definition 3)

1. Let $\tilde{\boldsymbol{\beta}}_{\mathcal{C}} = \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$. Using coordinate descent algorithm, iteratively use quadratic approximation and IRLS to compute $\hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}$ until convergence.

Numerical Optimization Algorithm for GCAL (Definition 4)

0. Initialize $\tilde{\boldsymbol{\beta}}_{\mathcal{D}} = \mathbf{0}$.

1. Given $\tilde{\boldsymbol{\beta}}_{\mathcal{D}}$, iteratively use quadratic approximation to compute $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}(\mathrm{m}le)$, where $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}(\mathrm{m}le)$ is the marginal MLE. Let $\tilde{\boldsymbol{\beta}}_{\mathcal{C}} = \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}(\mathrm{m}le)$.

2. Given $\tilde{\boldsymbol{\beta}}_{\mathcal{C}}$, using coordinate descent algorithm, iteratively use quadratic approximation and IRLS to compute $\hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}$ until convergence. Let $\tilde{\boldsymbol{\beta}}_{\mathcal{D}} = \hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}(\mathrm{m}le)$.

3. Repeat step 1 and 2 until convergence.

The numerical optimization algorithm of GCAL is slightly different from the naive coordinate descent for the Lasso in generalized linear models. The only difference is that we compute the marginal MLE on the conditional set first. Then we iteratively apply quadratic approximation, IRLS and coordinate descent to compute other estimated coefficients. And we iterate between the conditional set and non-conditional set until convergence.

3.4 Sufficient Conditional Adaptive Lasso for Generalized Linear Models

The Sufficient Conditional Adaptive Lasso Variable Screening algorithm for Generalized linear models (SCAL-VS-G) is very similar to the one in the linear case.

SCAL Variable Screening Algorithm for Generalized Linear Models (SCAL-VS-G)

0. First calculate two rankings–marginal distance correlation rankings between X_i and Y and in-between distance correlation rankings between X_i and X_j $(i \neq j)$ conditioning on Y. Combine two rankings by taking out the highest s_0 variables from the in-between correlation rankings and putting them on top of the marginal rankings. Update the design matrix $\mathbf{X}$ by reordering the columns based on the combined rankings. Without loss of generality, we will always use the ordered design matrix hereafter.

1. Separate the ordered predictors into to several sets sequentially, such that each set contains $\lfloor \delta n \rfloor$ $(0 < \delta < 1)$ variables except the last set. The last set has whatever variables left (less than $\lfloor \delta n \rfloor$). Let $\mathbf{X}_1, \ldots, \mathbf{X}_k$ be the separated k sub-design matrices with $p_1, \ldots, p_k$ $(p = \sum_{i=i}^{k} p_i)$ number of predictors, respectively, and $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_k$ be the estimated coefficients corresponding to sub-design matrices $\mathbf{X}_1, \ldots, \mathbf{X}_k$, respectively.

2. Get initial estimates through a sequence of generalized linear regressions. Let $\hat{\boldsymbol{\beta}}_1^{(0)}, \ldots, \hat{\boldsymbol{\beta}}_k^{(0)}$ all be zeros. For m^{th} iteration, set $\hat{\boldsymbol{\beta}}_1^{(m)}, \ldots, \hat{\boldsymbol{\beta}}_k^{(m)}$ equal to $\hat{\boldsymbol{\beta}}_1^{(m-1)}$, $\ldots$, $\hat{\boldsymbol{\beta}}_k^{(m-1)}$, respectively. Fit a likelihood based GLM using $\sum_{i \neq 1} \mathbf{X}_i \hat{\boldsymbol{\beta}}_i^{(m)}$ as the given canonical on predictor $\mathbf{X}_1$. Update $\hat{\boldsymbol{\beta}}_1^{(m)}$ using the estimated coefficients. Next, fit an likelihood based GLM using $\sum_{i \neq 2} \mathbf{X}_i \hat{\boldsymbol{\beta}}_i^{(m)}$ as the given canonical on predictor $\mathbf{X}_2$. Update $\hat{\boldsymbol{\beta}}_2^{(m)}$ using the estimated coefficients. Keep doing this until m_0^{th} iteration, when $\hat{\boldsymbol{\beta}}^{(m_0)}$ and $\hat{\boldsymbol{\beta}}^{(m_0-1)}$ are close enough by some criterion, e.g. ℓ_2-norm. Here, $\hat{\boldsymbol{\beta}}^{(m_0)}$ indicates the estimated coefficient vector of all ordered predictors.

3. Set $\hat{\boldsymbol{\beta}}^{(0)}(\text{scal})$ equal to $\hat{\boldsymbol{\beta}}^{(m_0)}$. For j^{th} iteration, set $\hat{\boldsymbol{\beta}}_1^{(j)}(\text{scal}), \ldots, \hat{\boldsymbol{\beta}}_k^{(j)}(\text{scal})$ equal to $\hat{\boldsymbol{\beta}}_1^{(j-1)}(\text{scal}), \ldots, \hat{\boldsymbol{\beta}}_k^{(j-1)}(\text{scal})$, respectively. For each fixed λ_n, fit an adaptive Lasso using $\sum_{i \neq 1} \mathbf{X}_i \hat{\boldsymbol{\beta}}_i^{(j)}(\text{scal})$ as the given canonical on predictor $\mathbf{X}_1$ with weights $1/|\hat{\boldsymbol{\beta}}_1^{(m_0)}|^{\gamma}$. Update $\hat{\boldsymbol{\beta}}_1^{(j)}(\text{scal})$ using the estimated coefficients. Fit an Adaptive Lasso using $\sum_{i \neq 2} \mathbf{X}_i \hat{\boldsymbol{\beta}}_i^{(j)}(\text{scal})$ as the given canonical on predictor $\mathbf{X}_2$ with weights $1/|\hat{\boldsymbol{\beta}}_2^{(m_0)}|^{\gamma}$. Update $\hat{\boldsymbol{\beta}}_2^{(j)}(\text{scal})$ using the estimated coefficients. Repeat until $\mathbf{X}_k$ is fitted. Use AIC to select best model with the corresponding j^{th} value. Remove those variables with zero estimated coefficients in $\mathbf{X}_1$ from $\mathbf{X}_1$. Combine $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$ to form the newly updated $\mathbf{X}$ which has fewer columns.

4. Repeat Step 0 to Step 3 for a fixed number of steps or stop until there's no zero estimated coefficients in current $\mathbf{X}_1$ or current $\mathbf{X}_1$ has less than $\lfloor \delta n \rfloor$ columns. Fit one last adaptive Lasso generalized linear model with original $\mathbf{y}$ on current remaining variables with corresponding weights in $1/|\hat{\boldsymbol{\beta}}^{(m_0)}|^{\gamma}$, resulting $\mathbf{X}_{\text{Remained}}$ being the only variables with non-zero coefficients. Let the $\mathbf{X}_{\text{Deleted}}$ be the design matrix containing all other deleted variable column. Order the columns of $\mathbf{X}_{\text{Deleted}}$ as described in Step 0. Update the design matrix $\mathbf{X}$ by putting the ordered columns in $\mathbf{X}_{\text{Deleted}}$ after the remaining (non-deleted) variable columns in $\mathbf{X}$. Return a full screening ranking of all the variables, that is the column number in $\mathbf{X}$.

Conditioning Set SCAL Variable Selection Algorithm for Generalized Linear Models (CS-SCAL-VS-G)

The Conditioning Set SCAL Variable Selection Algorithm for Generalized Linear Models (CS-SCAL-VS-G) is almost the same as SCAL-VS-G except that we incorporate a pre-known set of variables—conditioning

set $\mathbf{X}_{\mathcal{C}}$, in the model. Then the remaining variables have a corresponding design matrix $\mathbf{X}_{\mathcal{D}}$. Perform the same Steps 0-4 on $\mathbf{X}_{\mathcal{D}}$ as in SCAL-VS-G except that before Step 2 and Step 3, first fit a generalized linear model using $\mathbf{X}_{\mathcal{D}}\hat{\boldsymbol{\beta}}_{\mathcal{D}}$ as the given canonical on predictor $\mathbf{X}_{\mathcal{C}}$ and update $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(m)}$ and $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(j)}$(scal), correspondingly.

3.5 Numerical Studies

In our numerical studies, we use R package **ncvreg** for adaptive Lasso. GCAL and CS-SCAL-VS-G are written in C and R codes based on the source code of **ncvreg** package. We write our own CSIS code in R for all the simulations. All the codes are available upon request.

We use several different metrics to evaluate the performance of GCAL and CS-SCAL-VS-G. True positive rate (TPR) is the proportion of successfully selecting all important variables in the model. We use TPR as the evaluation metric in *Example 1*. Minimum model size is the largest variable rank among all the important variables. We report the median of minimum model size (MMMS) in *Example 2* and *Example 3*.

Simulations

In this section, simulation data are given by iid copies of (X^T, Y), where the conditional distribution of Y given $X = x$ is a binomial distribution with probability of success

$$\mathbb{P}(y|X = x) = \frac{\exp(x^T\boldsymbol{\beta}^*)}{1 + \exp(x^T\boldsymbol{\beta}^*)}. \tag{3.17}$$

Example 1: Let $n = 50, 100$, $p = 100, 1000$ and $\boldsymbol{\beta}^* = (3, 3, 3, 3, 3, -7.5, 2, 0, \ldots, 0)^T$. X_i's all follow the standard normal distribution with equal correlation 0.5 except that

X_7 is independent with all other X_i ($i \neq 7$). We will condition all X_1, X_1 to X_2, X_1 to X_3 and X_1 to X_4, respectively. And we report the TPR on the non-conditioning set for each conditioning set. For example, if the conditioning set is X_1 to X_2, we then report the TPR for X_3 to X_7, X_4 to X_7 and X_5 to X_7. Bayesian Information Criteria (Schwarz et al., 1978) is used to select the best model. And note that, when we calculate the number of non-zero coefficient for GCAL, k, we are excluding the conditioning set. That is, when calculating GCAL on conditioning set X_1 to X_2, k does not count X_1 and X_2. 200 simulations are performed. *Example 1* is mainly focusing on comparing the performance of GCAL with adaptive Lasso.

Table 3.1: *Example 1.* TPR for important variables in the non-conditioning set.

$n = 50, p = 100, \rho = 0.5$					
TPR for Variables	Adaptive Lasso	GCAL on Condition Set			
		X_1	X_1 to X_2	X_1 to X_3	X_1 to X_4
TPR for X_2 to X_7	0.010	0.020			
TPR for X_3 to X_7	0.020	0.035	0.040		
TPR for X_4 to X_7	0.040	0.070	0.055	0.055	
TPR for X_5 to X_7	0.070	0.110	0.125	0.115	0.135
$n = 100, p = 1000, \rho = 0.5$					
TPR for Variables	Adaptive Lasso	GCAL on Condition Set			
		X_1	X_1 to X_2	X_1 to X_3	X_1 to X_4
TPR for X_2 to X_7	0.005	0.010			
TPR for X_3 to X_7	0.010	0.015	0.015		
TPR for X_4 to X_7	0.020	0.025	0.035	0.040	
TPR for X_5 to X_7	0.060	0.060	0.085	0.125	0.160

Example 2: Let $n = 100$, $p = 300, 1000$ and $\boldsymbol{\beta^*} = (3, 3, 3, 3, 3, -7.5, 2, 0, \ldots, 0)^T$. X_i's all follow the standard normal distribution with equal correlation ρ ($\rho = 0, 0.2, 0.4, 0.6$) except that X_7 is independent with all other X_i ($i \neq 7$). The setting is almost the same as *Example 1* except that we only condition on X_1 to X_2.

Table 3.2: *Example 2.* MMMS and standard deviation.

		$n = 100, p = 300$		
Method	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$
CSIS	68.5 (84.4)	64.5 (84.0)	62.0 (65.0)	84.0 (66.0)
CS-SCAL-VS-G	21.0 (86.2)	22.0 (87.4)	18.5 (66.2)	20.0 (67.5)
		$n = 100, p = 1000$		
Method	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$
CSIS	216.5 (279.6)	189.0 (243.5)	196.5 (291.0)	299.0 (250.1)
CS-SCAL-VS-G	207.0 (290.2)	143.5 (263.0)	138.5 (311.7)	201.5 (275.5)

Example 1 is mainly to compare our method with adaptive Lasso, where we perform a simultaneous estimation. Variables with non-zero coefficient are kept and selected. We can observe at least two nice behaviors of GCAL under Table 3.1. Firstly, with some prior information (conditioning set), GCAL can better select the other remaining important variables compared with adaptive Lasso. Secondly, most of the time the TRP will go up, with the the conditioning set growing larger.

Example 2 mainly focuses on comparing our variable screening method with CSIS. We use the median and standard deviation of minimum model size as the comparison criteria. Under Table 3.2, we observe that CS-SCAL-VS-G has a much lower MMMS than CSIS with roughly the same standard deviation.

Example 3: This example has similar settings as in Fan and Song (2010) with $p = 5000$. We generate the predictors from

$$X_j = (\varepsilon_j + a_j\varepsilon)/\sqrt{1 + a_j^2}, \tag{3.18}$$

where ε and $\{\varepsilon_j\}_{j=1}^{p/3}$ follow iid standard normal distribution, $\{\varepsilon_j\}_{j=p/3+1}^{2p/3}$ follow iid double exponential distribution with location parameter 0 and scale parameter 1 and

$\{\varepsilon_j\}_{j=2p/3+1}^{p}$ are iid and follow a mixture normal distribution with two components $\mathrm{N}(-1,1)$, $\mathrm{N}(1,0.5)$ and equal mixture proportion. The predictors are standardized to have mean 0 and variance 1. We condition on X_1 to X_4.

Let $p = 5000$ and $s = 12$. The constants $a_1, \ldots, a_{100}$ are the same and chosen such that the correlation $\rho = \mathrm{corr}(X_i, X_j) = 0, 0.2, 0.4, 0.6$ and 0.8 among the first 100 variables and $a_{101} = \cdots = a_{5,000} = 0$. The true coefficient $\boldsymbol{\beta}^*$ is generated from an alternating sequence of 1 and 1.3.

Table 3.3: *Example 3.* MMMS and standard deviation.

	$n = 100, p = 5000$			
Method	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$
CSIS	185 (132)	31 (38)	29 (21)	32 (18)
CS-SCAL-VS-G	185 (132)	31 (38)	29 (21)	32 (18)

Table 3.3 shows the result for *Example 3*. Since we use the CSIS as our base ranking. Unfortunately, CS-SCAL-VS-G seems that CS-SCAL-VS-G does not improve the result of CSIS for this model.

Real Data

Leukemia data from high-density Affymetrix oligonucleotide arrays were previously analyzed in Golub et al. (1999), and are available at `http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`. There are in total 7129 genes and 72 patients from two classes: 47 in class ALL (acute lymphocytic leukemia) and 25 in class AML (acute mylogenous leukemia). Among these 72 patients, 38 (27 in class ALL and 11 in class AML) are set to be training samples and 34 (20 in class ALL and 14 in class AML) are set as test samples. We fit 3 different models, namely logistic regression with

Lasso penalty, SCAD penalty and our GCAL method. Based on the model results of Lasso and SCAD, we pick Gene-461 and Gene-6854 as our conditioning set.

Table 3.4: Classification errors of Leukemia dataset.

Method	Training Error	Testing Error	Selected Gene
Lasso	0/38	5/34	461, 804, 1834, 1882, 2354, 4535, 5039, 5772, 6218, 6378, 6854
SCAD	0/38	4/34	461, 1834, 1882, 2354, 4535, 5039, 5772, 6218, 6378, 6854
GCAL	0/38	4/34	461, 6854 + 571, 1035, 1929, 2214, 2890, 4560, 6989

As we can observe in Table 3.4, with a smaller testing error, GCAL can find very different genes when compared with Lasso and SCAD. This result may further suggest different important genes in classifying acute lymphocytic and acute mylogenous leukemia.

3.6 Discussion

In this chapter, we extend CAL to generalized linear models. We apply CAL on likelihood function with L_1 penalty. We develop the asymptotic and oracle properties for CAL under generalized linear models. Numerical studies and real data example show that GCAL and CS-SCAL-VS-G have better results for both variable selection and model estimation.

Chapter 4 The Conditionally Penalized Estimate and Its Oracle Properties

4.1 Introduction

One of the widely used high-dimensional variable selection techniques is through penalized least square or penalized likelihood estimation. Consider the linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector and $\mathbf{X}$ is an $n \times d$ design matrix, and $\boldsymbol{\varepsilon}$ is an n-dimensional noise vector. The penalized likelihood has the form

$$\frac{1}{n}\ell_n(\boldsymbol{\beta}) - \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{4.2}$$

where $\ell_n(\boldsymbol{\beta})$ is the log-likelihood function and $p_\lambda(\cdot)$ is a penalty function indexed by the regularization parameter $\lambda > 0$. We simultaneously select variables and estimate their associated regression coefficients by maximizing the penalized likelihood (4.2). When $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, maximizing the penalized likelihood is equivalent, up to an affine transformation of the log-likelihood, to minimizing the penalized least squares (PLS) problem

$$\frac{1}{2n}\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{4.3}$$

where $\|\cdot\|$ denotes the L^2-norm. A natural generalization of penalized L_0-regression is penalized L_q-regression, called bridge regression in Frank and Friedman (1993), in which $p_\lambda(|\theta|) = \lambda|\theta|^q$ for $0 < q \leq 2$. Special cases include best subset selection

($q = 0$) and ridge regression ($q = 2$). Penalized L_1-regression ($q = 1$) is called the Lasso by Tibshirani (1996). Fan and Li (2001) advocate penalty functions that give estimators with three properties:

1. *Sparsity*: The resulting estimator automatically sets small estimated coefficients to zero to accomplish variable selection and reduce model complexity.

2. *Unbiasedness*: The resulting estimator is nearly unbiased, especially when the true coefficient β_j is large, to reduce model bias.

3. *Continuity*: The resulting estimator is continuous in the data to reduce instability in model prediction (Breiman et al., 1996).

Fan and Li (2001) also show the corresponding conditions on $p(|\theta|)$ in order to satisfy the above three properties. A sufficient condition for unbiasedness is that $p'_\lambda(|\theta|) = 0$ for large $|\theta|$. A sufficient condition for the resulting estimator to be a thresholding rule is that the minimum of the $|\theta| + p'_\lambda(|\theta|) = 0$ is positive. A sufficient and necessary condition for continuity is that the minimum of the function is $|\theta| + p'_\lambda(|\theta|) = 0$ attained at 0. In other words, a penalty function satisfying the conditions of sparsity and continuity must be singular at the origin.

It is known that the convex L_q penalty with $q > 1$ does not satisfy the sparsity condition, whereas the convex L_1 penalty does not satisfy the unbiasedness condition, and the concave L_q penalty with $0 \leq q < 1$ does not satisfy the continuity condition. Zou (2006) propose the Adaptive Lasso which adds weights to β_j's, so that the minimizer will enjoy the oracle properties. Fan and Li (2001) introduce the smoothly clipped absolute deviation (SCAD), whose derivative is given by

$$p'_\lambda(\beta) = \lambda\Big\{I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda)\Big\} \quad \text{for some } a > 2, \tag{4.4}$$

where $a = 3.7$ is often used (suggested by a Bayesian argument).

4.2 The Conditionally Penalized Estimate and Its Oracle Properties in GLM

We adopt the same setting of GLM as in McCullagh and Nelder (1989). For a canonical parameter $\theta = \mathbf{x}^T\boldsymbol{\beta}^*$, generalized linear models consider that the generic density belongs to an exponential family

$$f(y|\mathbf{x}, \theta) = h(y)\exp(y\theta - \phi(\theta)). \tag{4.5}$$

Without loss of generality, let $\mathcal{C}$ be the index set of the first q conditional variables, that is $\mathcal{C} = \{1, 2, \ldots, q\}$. Let $\mathcal{D}$ be the index set of the remaining $d = p - q$ variables, that is $\mathcal{D} = \{q+1, q+2, \ldots, p\}$. And we will use the notation:

$$\boldsymbol{\beta}_{\mathcal{C}} = (\beta_1, \ldots, \beta_q)^T \in \mathbf{R}^q \qquad \text{and} \qquad \boldsymbol{\beta}_{\mathcal{D}} = (\beta_{q+1}, \ldots, \beta_p)^T \in \mathbf{R}^d.$$

The covariates have been standardized so that $E(X_j) = 0$ and $E(X_j^2) = 1$, for $j = 1, 2, \ldots, p$. Let $\mathcal{A} = \{j \in \mathcal{D} : \beta_j^* \neq 0\} = \{q+1, q+2, \ldots, q+s\}$, that is, the first s variables in $\mathcal{D}$ are active and they constitute the set $\mathcal{A}$. Let $\mathcal{A}^c = \{j \in \mathcal{D} : j \notin \mathcal{A}\}$. Suppose that $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the response vector and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, $i = 1, \ldots, n$, is the ith sample. Let $\mathbf{X}$ be the design matrix. We also have the following assumption about the Fisher information matrix,

$$\mathbf{I}(\boldsymbol{\beta}^*) = \begin{bmatrix} \mathbf{I}_{\mathcal{C}\mathcal{C}} & \mathbf{I}_{\mathcal{C}\mathcal{D}} \\ \mathbf{I}_{\mathcal{D}\mathcal{C}} & \mathbf{I}_{\mathcal{D}\mathcal{D}} \end{bmatrix}, \tag{4.6}$$

$$\mathbf{I}_{\mathcal{C}\mathcal{D}} = \begin{bmatrix} \mathbf{I}_{\mathcal{C}\mathcal{D}1} & \mathbf{I}_{\mathcal{C}\mathcal{D}2} \end{bmatrix} \qquad \text{and} \qquad \mathbf{I}_{\mathcal{D}\mathcal{D}} = \begin{bmatrix} \mathbf{I}_{ss} & \mathbf{I}_{s(d-s)} \\ \mathbf{I}_{(d-s)s} & \mathbf{I}_{(d-s)(d-s)} \end{bmatrix}. \tag{4.7}$$

where $\mathbf{I}_{\mathcal{C}\mathcal{C}}$ is a $q \times q$ matrix, $\mathbf{I}_{\mathcal{C}\mathcal{D}}$ is a $q \times d$ matrix, $\mathbf{I}_{\mathcal{D}\mathcal{D}}$ is a $d \times d$ matrix. $\mathbf{I}_{ss}$, $\mathbf{I}_{s(d-s)}$ and $\mathbf{I}_{(d-s)(d-s)}$ are $s \times s$, $s \times (d-s)$ and $(d-s) \times (d-s)$ matrices, respectively. $\mathbf{I}_{\mathcal{C}\mathcal{D}1}$ is a $q \times s$ matrix and $\mathbf{I}_{\mathcal{C}\mathcal{D}2}$ is a $q \times (d-s)$ matrix.

With the knowledge of some prior information, we propose a conditional estimate for any penalty function that satisfy certain conditions under the generalized linear model situations.

Definition 5. *Suppose that $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$ is a root-n-consistent estimator of $\boldsymbol{\beta}_{\mathcal{C}}^{*}$. We assume that $\hat{\boldsymbol{\beta}}(\text{mle})$ is the maximum likelihood estimates in the GLM. The weight vector is constructed as $\hat{\boldsymbol{\omega}} = 1/|\hat{\boldsymbol{\beta}}(\text{mle})|^{\gamma}$ for some $\gamma > 0$. The conditionally penalized estimate (CPE) $\hat{\boldsymbol{\beta}}^{(n)}(\text{gcal})$ is given by*

$$\hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}(\text{cpe}) = \underset{\boldsymbol{\beta}_{\mathcal{D}}}{\text{argmin}} \sum_{i=1}^{n} \Big(-y_i \big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}} \big) + \phi \big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}} \big) \Big) + n \sum_{j \in \mathcal{D}} p_{\lambda}(|\beta_j|), \tag{4.8}$$

where $p_{\lambda}(\cdot)$ is the penalty function.

We prove that the above conditional estimates enjoy the oracle properties for any penalty functions that satisfy certain conditions. The theorem is given below.

Theorem 5. *Suppose that $\lambda_n \to 0$, $\max\{p''_{\lambda_n}(|\beta_j^*|) : \beta^* \neq 0\} \to 0$, $\max\{\sqrt{n} p'_{\lambda_n}(|\beta_j^*|) : \beta^* \neq 0\} \to 0$, $\sqrt{n} p'_{\lambda_n}(0) \to \infty$ and $\sqrt{n} p''_{\lambda_n}(0)$ is finite. Let $\mathcal{A}_n = \{j \in \mathcal{D} : \hat{\boldsymbol{\beta}}_j^{(n)} \neq 0\}$, that is, the estimated active set in $\mathcal{D}$. Let $\mathcal{A} = \{j \in \mathcal{D} : \beta_j^* \neq 0\}$, that is, the true active set in $\mathcal{D}$. Under the regularity conditions in the Appendix, the conditionally penalized estimate (CPE, Definition 5) must satisfy the following:*

1. *Consistency in variable selection:* $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$;

2. *Asymptotic normality:* $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(n)} - \boldsymbol{\beta}_{\mathcal{A}}^{*}) \to_d \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{ss}^{-1})$.

Remark 1. For SCAD penalty, if $\beta^* > 0$, $p''_{\lambda_n}(\beta^*) \to 0$ and $\sqrt{n}p'_{\lambda_n}(\beta^*) \to 0$. Also note that $\sqrt{n}p'_{\lambda_n}(0) = \sqrt{n}\lambda_n$, thus $\sqrt{n}p'_{\lambda_n}(0) \to \infty$ is equivalent to $\sqrt{n}\lambda_n \to \infty$. Combining $p''_{\lambda_n}(0) = 0$, it's clear that SCAD penalty satisfies the assumptions of Theorem 5.

Remark 2. For Adaptive Lasso penalty, if $\beta^* > 0$, $p'_{\lambda_n}(\beta^*) = \lambda_n \hat{w}_j/n \to 0$ due to the assumptions that $\lambda_n/\sqrt{n} \to 0$ and $\hat{w}_j \to_p (\beta_j^*)^{-\gamma}$ in Zou (2006). Also note that $p''(\beta) = 0$ for $\forall\beta$. Since $\lambda_n n^{(\gamma-1)/2} \to \infty$ and $\hat{w}_j n^{-\gamma/2} = O_p(1)$, $\sqrt{n}p'_{\lambda_n}(0) = \lambda_n \hat{w}_j/\sqrt{n} \to \infty$. Thus, Adaptive Lasso penalty also satisfies the assumption in Theorem 5.

Similar with GCAL, we can simultaneously estimate $\boldsymbol{\beta}_{\mathcal{C}}$ and $\boldsymbol{\beta}_{\mathcal{D}}$ in CPE if we do not use the previously estimated $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$ as the values of $\boldsymbol{\beta}_{\mathcal{C}}$. The definition can be found below.

Definition 6. *Suppose that $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$ is a root-n-consistent estimator of $\boldsymbol{\beta}_{\mathcal{C}}^{*}$. We assume that $\hat{\boldsymbol{\beta}}(\mathrm{mle})$ is the maximum likelihood estimates in the GLM. The weight vector is constructed as $\hat{\boldsymbol{\omega}} = 1/|\hat{\boldsymbol{\beta}}(\mathrm{mle})|^{\gamma}$ for some $\gamma > 0$. The GCAL estimates $\hat{\boldsymbol{\beta}}^{(n)}(\mathrm{gcal})$ are given by*

$$\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}, \hat{\boldsymbol{\beta}}^{(n)}(\mathrm{cpe}) = \underset{\boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\beta}_{\mathcal{D}}}{\operatorname{argmin}} \sum_{i=1}^{n} \Big(-y_i \big(\mathbf{x}_{i\mathcal{C}}^T \boldsymbol{\beta}_{\mathcal{C}} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}} \big) + \phi\big(\mathbf{x}_{i\mathcal{C}}^T \boldsymbol{\beta}_{\mathcal{C}} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}} \big) \Big) + \lambda_n \sum_{j\in\mathcal{D}} \hat{w}_j |\beta_j|, \tag{4.9}$$

where λ_n varies with n.

We also show that CPE under Definition 6 enjoys the oracle properties.

Theorem 6. *Suppose that* $\lambda_n/\sqrt{n} \to 0$ *and* $\lambda_n n^{(\gamma-1)/2} \to \infty$. *Let* $\mathcal{A}_n = \{j \in \mathcal{D} : \hat{\boldsymbol{\beta}}_j^{(n)}(\text{cpe}) \neq 0\}$, *that is, the estimated active set. The GCAL estimates under Definition 6 must satisfy the following:*

1. *Consistency in variable selection:* $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$;
2. *Asymptotic normality:* $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(n)}(\text{cpe}) - \boldsymbol{\beta}_{\mathcal{A}}^*) \to_d \text{N}(\mathbf{0}, \mathbf{I}_{ss}^{-1})$.

The proof of Theorem 6 is also deferred to the Appendix.

4.3 The Sufficient Conditionally Penalized Estimate

The sufficient conditionally penalized estimate variable screening (SCPE-VS) algorithm is very similar to the ones in linear and generalized linear models.

Sufficient Conditionally Penalized Estimate Variable Screening Algorithm (SCPE-VS)

0. First calculate two rankings–marginal distance correlation rankings between X_i and Y and in-between distance correlation rankings between X_i and X_j $(i \neq j)$ conditioning on Y. Combine two rankings by taking out the highest s_0 variables from the in-between correlation rankings and putting them on top of the marginal rankings. Update the design matrix $\mathbf{X}$ by reordering the columns based on the combined rankings. Without loss of generality, we will always use the ordered design matrix hereafter.

1. Separate the ordered predictors into to several sets sequentially, such that each set contains $\lfloor \delta n \rfloor$ $(0 < \delta < 1)$ variables except the last set. The last set has whatever variables left (less than $\lfloor \delta n \rfloor$). Let $\mathbf{X}_1, \ldots, \mathbf{X}_k$ be the separated k

sub-design matrices with $p_1, \ldots, p_k$ ($p = \sum_{i=i}^{k} p_i$) number of predictors, respectively, and $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_k$ be the estimated coefficients corresponding to sub-design matrices $\mathbf{X}_1, \ldots, \mathbf{X}_k$, respectively.

2. Get initial estimates through a sequence of linear regressions. Let $\hat{\boldsymbol{\beta}}_1^{(0)}, \ldots, \hat{\boldsymbol{\beta}}_k^{(0)}$ all be zeros. For m^{th} iteration, set $\hat{\boldsymbol{\beta}}_1^{(m)}, \ldots, \hat{\boldsymbol{\beta}}_k^{(m)}$ equal to $\hat{\boldsymbol{\beta}}_1^{(m-1)}, \ldots, \hat{\boldsymbol{\beta}}_k^{(m-1)}$, respectively. Compute a CPE based target function (4.2) using $\mathbf{y} - \sum_{i \neq 1} \mathbf{X}_i \hat{\boldsymbol{\beta}}_i^{(m)}$ as the response variable on predictor $\mathbf{X}_1$. Update $\hat{\boldsymbol{\beta}}_1^{(m)}$ using the estimated coefficients. Next, compute a CPE based target function (4.2) using $\mathbf{y} - \sum_{i \neq 2} \mathbf{X}_i \hat{\boldsymbol{\beta}}_i^{(m)}$ as the response on predictor $\mathbf{X}_2$. Update $\hat{\boldsymbol{\beta}}_2^{(m)}$ using the estimated coefficients. Keep doing this until m_0^{th} iteration, when $\hat{\boldsymbol{\beta}}^{(m_0)}$ and $\hat{\boldsymbol{\beta}}^{(m_0-1)}$ are close enough by some criterion, e.g. ℓ_2-norm. Here, $\hat{\boldsymbol{\beta}}^{(m_0)}$ indicates the estimated coefficient vector of all ordered predictors.

3. Set $\hat{\boldsymbol{\beta}}^{(0)}(\text{scpe})$ equal to $\hat{\boldsymbol{\beta}}^{(m_0)}$. For j^{th} iteration, set $\hat{\boldsymbol{\beta}}_1^{(j)}(\text{scpe}), \ldots, \hat{\boldsymbol{\beta}}_k^{(j)}(\text{scpe})$ equal to $\hat{\boldsymbol{\beta}}_1^{(j-1)}(\text{scpe}), \ldots, \hat{\boldsymbol{\beta}}_k^{(j-1)}(\text{scpe})$, respectively. For each fixed λ_n, compute a CPE based target function (4.2) using $\mathbf{y} - \sum_{i \neq 1} \mathbf{X}_i \hat{\boldsymbol{\beta}}_i^{(j)}(\text{scpe})$ as the response on predictor $\mathbf{X}_1$ with weights $1/|\hat{\boldsymbol{\beta}}_1^{(m_0)}|^{\gamma}$. Update $\hat{\boldsymbol{\beta}}_1^{(j)}(\text{scpe})$ using the estimated coefficients. Compute a CPE based target function (4.2) using $\mathbf{y} - \sum_{i \neq 2} \mathbf{X}_i \hat{\boldsymbol{\beta}}_i^{(j)}(\text{scpe})$ as the response on predictor $\mathbf{X}_2$ with weights $1/|\hat{\boldsymbol{\beta}}_2^{(m_0)}|^{\gamma}$. Update $\hat{\boldsymbol{\beta}}_2^{(j)}(\text{scpe})$ using the estimated coefficients. Repeat until $\mathbf{X}_k$ is fitted. Use RIC to forcibly select less than n variables combining all the pieces. Remove those variables with zero estimated coefficients in $\mathbf{X}_1$ from $\mathbf{X}_1$. Combine $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$ to form the newly updated $\mathbf{X}$ which has fewer columns.

4. Repeat Step 0 to Step 3 until there's no zero estimated coefficients in current $\mathbf{X}_1$ or current $\mathbf{X}_1$ has less than $\lfloor \delta n \rfloor$ columns. Perform an Adaptive Lasso fit

with original $\mathbf{y}$ on current $\mathbf{X}_1$ with corresponding weights in $1/|\hat{\boldsymbol{\beta}}^{(m_0)}|^{\gamma}$. Let the $\mathbf{X}_{\text{Deleted}}$ be the design matrix only containing the deleted variable column. Order the columns of $\mathbf{X}_{\text{Deleted}}$ as described in Step 0. Update the design matrix $\mathbf{X}$ by putting the ordered columns in $\mathbf{X}_{\text{Deleted}}$ after the remaining (non-deleted) variable columns in $\mathbf{X}$. Return the column number (rank) of the current $\mathbf{X}$ as the screening ranking of all the variables.

Conditioning Set SCPE-VS Algorithm (CS-SCPE-VS)

The Conditioning Set SCPE Variable Selection Algorithm (CS-SCPE-VS) is almost the same as SCPE-VS except that we incorporate a pre-known set of variables—conditioning set $\mathbf{X}_{\mathcal{C}}$, in the model. Then the remaining variables have a corresponding design matrix $\mathbf{X}_{\mathcal{D}}$. Perform the same Steps 0-4 on $\mathbf{X}_{\mathcal{D}}$ as in SCPE-VS except that before Step 2 and Step 3, first fit a generalized linear model using $\mathbf{y} - \mathbf{X}_{\mathcal{D}}\hat{\boldsymbol{\beta}}_{\mathcal{D}}$ as the response on predictor $\mathbf{X}_{\mathcal{C}}$ and update $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(m)}$ and $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(j)}$(scpe), correspondingly.

4.4 Numerical Studies

In our numerical studies, we use R package **ncvreg** for SCAD penalty. CPE and CS-SCPE-VS are written in C and R codes based on the source code of **ncvreg** package. We write our own CSIS code in R for all the simulations. All the codes are available upon request.

We use several different metrics to evaluate the performance of CPE and CS-SCPE-VS. True positive rate (TPR) is the proportion of successfully selecting all important variables in the model. We use TPR as the evaluation metric in *Example 1*. Minimum model size is the largest variable rank among all the important variables.

We report the median of minimum model size (MMMS) in *Example 2* and *Example 3*.

Simulations

In this section, simulation data are given by iid copies of (X^T, Y), where the conditional distribution of Y given $X = x$ is a binomial distribution with probability of success

$$\mathbb{P}(y|X = x) = \frac{\exp(x^T\boldsymbol{\beta^*})}{1 + \exp(x^T\boldsymbol{\beta^*})}. \tag{4.10}$$

Example 1: Let $n = 50, 100$, $p = 100, 1000$ and $\boldsymbol{\beta^*} = (3, 3, 3, 3, 3, -7.5, 2, 0, \ldots, 0)^T$. X_i's all follow the standard normal distribution with equal correlation 0.5 except that X_7 is independent with all other X_i ($i \neq 7$). We will condition all X_1, X_1 to X_2, X_1 to X_3 and X_1 to X_4, respectively. And we report the TPR on the non-conditioning set for each conditioning set. For example, if the conditioning set is X_1 to X_2, we then report the TPR for X_3 to X_7, X_4 to X_7 and X_5 to X_7. Bayesian Information Criteria (Schwarz et al., 1978) is used to select the best model. And note that, when we calculate the number of non-zero coefficient for GCAL, k, we are excluding the conditioning set. That is, when calculating GCAL on conditioning set X_1 to X_2, k does not count X_1 and X_2. 200 simulations are performed. *Example 1* is mainly focusing on comparing the performance of GCAL with adaptive Lasso.

Example 1 is mainly to compare conditional SCAD versus original SCAD, where we perform a simultaneous estimation. Variables with non-zero coefficient are kept and selected. Based on Table 4.1, we can observe at least two nice behaviors of CPE with SCAD penalty. Firstly, with some prior information (conditioning set), CPE-SCAD can better select the other remaining important variables compared with original SCAD. Secondly, TRP will go up, with the the conditioning set growing

Table 4.1: *Example 1.* TPR for important variables in the non-conditioning set.

$n = 50, p = 100, \rho = 0.5$					
TPR for Variables	SCAD	CPE-SCAD with Condition Set			
		X_1	X_1 to X_2	X_1 to X_3	X_1 to X_4
TPR for X_2 to X_7	0.180	0.285			
TPR for X_3 to X_7	0.240	0.355	0.390		
TPR for X_4 to X_7	0.305	0.435	0.450	0.520	
TPR for X_5 to X_7	0.420	0.590	0.595	0.635	0.630
$n = 100, p = 1000, \rho = 0.5$					
TPR for Variables	SCAD	CPE-SCAD with Condition Set			
		X_1	X_1 to X_2	X_1 to X_3	X_1 to X_4
TPR for X_2 to X_7	0.210	0.335			
TPR for X_3 to X_7	0.280	0.380	0.440		
TPR for X_4 to X_7	0.355	0.450	0.510	0.545	
TPR for X_5 to X_7	0.465	0.620	0.650	0.650	0.650

larger.

4.5 Discussion

In this chapter, we introduce the Conditionally Penalized Estimate (CPE) for generalized linear models. It works with different penalty functions such as SCAD. We demonstrate the asymptotic properties of CPE. We propose Sufficient Conditionally Penalized Estimate Variable Screening (SCPE-VS) and Conditioning Set Sufficient Conditionally Penalized Estimate Variable Screening (CS-SCPE-VS) algorithms based on CPE. Simulations and real data examples are evaluated to show the good performance of CPE, SCPE-VS and CS-SCPE-VS. Overall, we can establish an entire theory and methods for our approach: loss function, penalty term with a conditional set. We summarize all cases we discuss in Table 4.2.

Conditional penalized approach is a novel statistical method. It appears that our approach can be extended to survival analysis and longitudinal data etc. These discussions are under future consideration. We hope that our research will open a new discussion and bring more specific and fine statistical methods especially in Big Data.

Table 4.2: Summary for different loss function and regularization with a conditional set.

Loss Function Type	Regularization	Conditional Set	Chapter
Linear Model (Least Square)	L_1	Yes	Chapter 2
	SCAD	Yes	Chapter 4
Generalized Linear Model (Likelihood)	L_1	Yes	Chapter 3
	SCAD	Yes	Chapter 4
Other Loss Function (such as Survival Analysis)	L_1	Yes	Future Study
	SCAD	Yes	Future Study

Appendices

A Supplementary Materials for Chapter 2

Proof of Theorem 1

We first prove the asymptotic normality part. Let $\mathbf{u}_{\mathcal{C}} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} - \boldsymbol{\beta}_{\mathcal{C}}^*)$, $\mathbf{u}_{\mathcal{D}} = \sqrt{n}(\boldsymbol{\beta}_{\mathcal{D}} - \boldsymbol{\beta}_{\mathcal{D}}^*)$ and

$$\Psi_n(\mathbf{u}_{\mathcal{D}}) = \left\| \mathbf{y} - \sum_{j\in\mathcal{C}} \mathbf{x}_j \left(\beta_j^* + \frac{u_j}{\sqrt{n}}\right) - \sum_{j\in\mathcal{D}} \mathbf{x}_j \left(\beta_j^* + \frac{u_j}{\sqrt{n}}\right) \right\|^2 + \lambda_n \sum_{j\in\mathcal{D}} \hat{w}_j \left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right|.$$

Let $\hat{\mathbf{u}}_{\mathcal{D}} = \operatorname{argmin} \Psi_n(\mathbf{u}_{\mathcal{D}})$; then $\hat{\boldsymbol{\beta}}^{CAL} = \boldsymbol{\beta}_{\mathcal{D}}^* + \frac{\hat{\mathbf{u}}_{\mathcal{D}}}{\sqrt{n}}$. Note that

$$\begin{aligned}
\Psi_n(\mathbf{u}_{\mathcal{D}}) &= \left\| \boldsymbol{\varepsilon} - \sum_{j\in\mathcal{C}} \mathbf{x}_j \frac{u_j}{\sqrt{n}} - \sum_{j\in\mathcal{D}} \mathbf{x}_j \frac{u_j}{\sqrt{n}} \right\|^2 + \lambda_n \sum_{j\in\mathcal{D}} \hat{w}_j \left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| \\
&= \left(\boldsymbol{\varepsilon} - \frac{1}{\sqrt{n}} \mathbf{X}_{\mathcal{C}} \mathbf{u}_{\mathcal{C}} - \frac{1}{\sqrt{n}} \mathbf{X}_{\mathcal{D}} \mathbf{u}_{\mathcal{D}} \right)^T \left(\boldsymbol{\varepsilon} - \frac{1}{\sqrt{n}} \mathbf{X}_{\mathcal{C}} \mathbf{u}_{\mathcal{C}} - \frac{1}{\sqrt{n}} \mathbf{X}_{\mathcal{D}} \mathbf{u}_{\mathcal{D}} \right) + \lambda_n \sum_{j\in\mathcal{D}} \hat{w}_j \left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| \\
&= \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} + \mathbf{u}_{\mathcal{C}}^T \frac{\mathbf{X}_{\mathcal{C}}^T \mathbf{X}_{\mathcal{C}}}{n} \mathbf{u}_{\mathcal{C}} + \mathbf{u}_{\mathcal{D}}^T \frac{\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}}}{n} \mathbf{u}_{\mathcal{D}} - 2\frac{\boldsymbol{\varepsilon}^T \mathbf{X}_{\mathcal{C}}}{\sqrt{n}} \mathbf{u}_{\mathcal{C}} - 2\frac{\boldsymbol{\varepsilon}^T \mathbf{X}_{\mathcal{D}}}{\sqrt{n}} \mathbf{u}_{\mathcal{D}} + 2\mathbf{u}_{\mathcal{C}}^T \frac{\mathbf{X}_{\mathcal{C}}^T \mathbf{X}_{\mathcal{D}}}{n} \mathbf{u}_{\mathcal{D}} \\
&\quad + \lambda_n \sum_{j\in\mathcal{D}} \hat{w}_j \left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right|
\end{aligned}$$

and

$$\Psi_n(\mathbf{0}) = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} + \mathbf{u}_{\mathcal{C}}^T \frac{\mathbf{X}_{\mathcal{C}}^T \mathbf{X}_{\mathcal{C}}}{n} \mathbf{u}_{\mathcal{C}} - 2\frac{\boldsymbol{\varepsilon}^T \mathbf{X}_{\mathcal{C}}}{\sqrt{n}} \mathbf{u}_{\mathcal{C}} + \lambda_n \sum_{j\in\mathcal{D}} \hat{w}_j |\beta_j^*|.$$

Let

$$V^{(n)}(\mathbf{u}_{\mathcal{D}}) = \Psi_n(\mathbf{u}_{\mathcal{D}}) - \Psi_n(\mathbf{0}) = \mathbf{u}_{\mathcal{D}}^T \frac{\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}}}{n} \mathbf{u}_{\mathcal{D}} - 2\frac{\boldsymbol{\varepsilon}^T \mathbf{X}_{\mathcal{D}}}{\sqrt{n}} \mathbf{u}_{\mathcal{D}} + 2\mathbf{u}_{\mathcal{C}}^T \frac{\mathbf{X}_{\mathcal{C}}^T \mathbf{X}_{\mathcal{D}}}{n} \mathbf{u}_{\mathcal{D}} + \lambda_n \sum_{j\in\mathcal{D}} \hat{w}_j \left[\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| - |\beta_j^*| \right].$$

Consider the second and third term in $V^{(n)}(\mathbf{u}_{\mathcal{D}})$. From matrix inverse, we know that

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}} & \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} \\ \mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}} & \mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{X}^{11} & \mathbf{X}^{12} \\ \mathbf{X}^{21} & \mathbf{X}^{22} \end{bmatrix},$$

where

$$\mathbf{X}^{11} = (\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1} + (\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}\mathbf{X}_{22\cdot1}^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}(\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}, \tag{11}$$

$$\mathbf{X}^{12} = -\mathbf{X}_{11\cdot2}^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}})^{-1}, \tag{12}$$

$$\mathbf{X}_{11\cdot2} = \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}} - \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}})^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}, \tag{13}$$

$$\mathbf{X}_{22\cdot1} = \mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}} - \mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}(\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}. \tag{14}$$

Observe that

$$\begin{aligned} \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}})^{-1}\mathbf{X}_{22\cdot1} &= \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} - \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}})^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}(\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} \\ &= [\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}} - \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}})^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}](\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} \\ &= \mathbf{X}_{11\cdot2}(\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}, \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbf{X}^{11}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}} + \mathbf{X}^{12}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}} &= \mathbf{I} + (\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}\mathbf{X}_{22\cdot1}^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}} - \mathbf{X}_{11\cdot2}^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}})^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}} \\ &= \mathbf{I} + [(\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}\mathbf{X}_{22\cdot1}^{-1} - \mathbf{X}_{11\cdot2}^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}})^{-1}]\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}} \\ &= \mathbf{I}. \end{aligned}$$

Also observe that

$$\begin{aligned}\mathbf{X}_{11\cdot 2}(\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} &= \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} - \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}})^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}(\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} \\ &= \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}})^{-1}[\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}} - \mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}(\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}] \\ &= \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}})^{-1}\mathbf{X}_{22\cdot 1}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbf{X}^{11}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} + \mathbf{X}^{12}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}} &= (\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} + (\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}\mathbf{X}_{22\cdot 1}^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}(\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} \\ &\quad - \mathbf{X}_{11\cdot 2}^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} \\ &= \mathbf{X}_{11\cdot 2}^{-1}[\mathbf{X}_{11\cdot 2}(\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1} \\ &\quad + \mathbf{X}_{11\cdot 2}(\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}\mathbf{X}_{22\cdot 1}^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}(\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}]\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} - \mathbf{X}_{11\cdot 2}^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} \\ &= \mathbf{X}_{11\cdot 2}^{-1}[\mathbf{I} - \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}})^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}(\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1} \\ &\qquad + \mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}(\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}})^{-1}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}(\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}})^{-1}]\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} - \mathbf{X}_{11\cdot 2}^{-1}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} \\ &= \mathbf{0}.\end{aligned}$$

Finally,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathcal{C}} &= \begin{bmatrix}\mathbf{X}^{11} & \mathbf{X}^{12}\end{bmatrix}\mathbf{X}^T Y \\ &= \begin{bmatrix}\mathbf{X}^{11} & \mathbf{X}^{12}\end{bmatrix}\begin{bmatrix}\mathbf{X}_{\mathcal{C}}^T \\ \mathbf{X}_{\mathcal{D}}^T\end{bmatrix}(\mathbf{X}_{\mathcal{C}}\boldsymbol{\beta}_{\mathcal{C}}^* + \mathbf{X}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^* + \boldsymbol{\varepsilon}) \\ &= (\mathbf{X}^{11}\mathbf{X}_{\mathcal{C}}^T + \mathbf{X}^{12}\mathbf{X}_{\mathcal{D}}^T)\boldsymbol{\varepsilon} + (\mathbf{X}^{11}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}} + \mathbf{X}^{12}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}})\boldsymbol{\beta}_{\mathcal{C}}^* + (\mathbf{X}^{11}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}} + \mathbf{X}^{12}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}})\boldsymbol{\beta}_{\mathcal{D}}^* \\ &= (\mathbf{X}^{11}\mathbf{X}_{\mathcal{C}}^T + \mathbf{X}^{12}\mathbf{X}_{\mathcal{D}}^T)\boldsymbol{\varepsilon} + \boldsymbol{\beta}_{\mathcal{C}}^*.\end{aligned}$$

Then the third term becomes

$$2\mathbf{u}_{\mathcal{C}}^T\frac{\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}}{n}\mathbf{u}_{\mathcal{D}} = 2\mathbf{u}_{\mathcal{D}}^T\frac{\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}}{n}\mathbf{u}_{\mathcal{C}} = 2\mathbf{u}_{\mathcal{D}}^T\frac{\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}}{n}\sqrt{n}(\mathbf{X}^{11}\mathbf{X}_{\mathcal{C}}^T + \mathbf{X}^{12}\mathbf{X}_{\mathcal{D}}^T)\boldsymbol{\varepsilon}.$$

The second term is

$$-2\frac{\boldsymbol{\varepsilon}^T\mathbf{X}_{\mathcal{D}}}{\sqrt{n}}\mathbf{u}_{\mathcal{D}} = -2\mathbf{u}_{\mathcal{D}}^T\frac{\mathbf{X}_{\mathcal{D}}^T}{\sqrt{n}}\boldsymbol{\varepsilon}.$$

Combining the second and third terms,

$$\begin{aligned}-2\frac{\boldsymbol{\varepsilon}^T\mathbf{X}_{\mathcal{D}}}{\sqrt{n}}\mathbf{u}_{\mathcal{D}} + 2\mathbf{u}_{\mathcal{C}}^T\frac{\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}}{n}\mathbf{u}_{\mathcal{D}} &= -2\mathbf{u}_{\mathcal{D}}^T\left[\frac{\mathbf{X}_{\mathcal{D}}^T}{\sqrt{n}} - \frac{\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}}{\sqrt{n}}(\mathbf{X}^{11}\mathbf{X}_{\mathcal{C}}^T + \mathbf{X}^{12}\mathbf{X}_{\mathcal{D}}^T)\right]\boldsymbol{\varepsilon} \\ &\to -2\mathbf{u}_{\mathcal{D}}^T\mathbf{W}.\end{aligned}$$

where $\mathbf{W} = N(\mathbf{0}, \sigma^2\Sigma_{*\mathcal{D}})$. Note that

$$\begin{aligned}\frac{\mathbf{X}_{11\cdot 2}}{n} &\to \Sigma_{\mathcal{C}\mathcal{C}} - \Sigma_{\mathcal{C}\mathcal{D}}\Sigma_{\mathcal{D}\mathcal{D}}^{-1}\Sigma_{\mathcal{D}\mathcal{C}} = \Sigma_{\mathcal{C}|\mathcal{D}}, \\ \frac{\mathbf{X}_{22\cdot 1}}{n} &\to \Sigma_{\mathcal{D}\mathcal{D}} - \Sigma_{\mathcal{D}\mathcal{C}}\Sigma_{\mathcal{C}\mathcal{C}}^{-1}\Sigma_{\mathcal{C}\mathcal{D}} = \Sigma_{\mathcal{D}|\mathcal{C}}, \\ n\mathbf{X}^{11} &\to \Sigma_{\mathcal{C}\mathcal{C}}^{-1} + \Sigma_{\mathcal{C}\mathcal{C}}^{-1}\Sigma_{\mathcal{C}\mathcal{D}}\Sigma_{\mathcal{D}|\mathcal{C}}^{-1}\Sigma_{\mathcal{D}\mathcal{C}}\Sigma_{\mathcal{C}\mathcal{C}}^{-1} \equiv \Sigma_{11}, \\ n\mathbf{X}^{12} &\to -\Sigma_{\mathcal{C}|\mathcal{D}}^{-1}\Sigma_{\mathcal{C}\mathcal{D}}\Sigma_{\mathcal{D}\mathcal{D}}^{-1} \equiv \Sigma_{12}.\end{aligned}$$

Thus,

$$\begin{aligned}
&\left[\frac{\mathbf{X}_{\mathcal{D}}^T}{\sqrt{n}} - \frac{\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}}{\sqrt{n}}(\mathbf{X}^{11}\mathbf{X}_{\mathcal{C}}^T + \mathbf{X}^{12}\mathbf{X}_{\mathcal{D}}^T)\right]\left[\frac{\mathbf{X}_{\mathcal{D}}^T}{\sqrt{n}} - \frac{\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}}{\sqrt{n}}(\mathbf{X}^{11}\mathbf{X}_{\mathcal{C}}^T + \mathbf{X}^{12}\mathbf{X}_{\mathcal{D}}^T)\right]^T \\
&\quad = \left[\frac{\mathbf{X}_{\mathcal{D}}^T}{\sqrt{n}} - \frac{\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}}{\sqrt{n}}(\mathbf{X}^{11}\mathbf{X}_{\mathcal{C}}^T + \mathbf{X}^{12}\mathbf{X}_{\mathcal{D}}^T)\right]\left[\frac{\mathbf{X}_{\mathcal{D}}}{\sqrt{n}} - (\mathbf{X}_{\mathcal{C}}\mathbf{X}^{11} + \mathbf{X}_{\mathcal{D}}\mathbf{X}^{12})\frac{\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}}{\sqrt{n}}\right] \\
&\quad = \frac{\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}}{n} - \left(\frac{\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}}{\sqrt{n}}\mathbf{X}^{11} + \frac{\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}}{\sqrt{n}}\mathbf{X}^{12}\right)\frac{\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}}{\sqrt{n}} - \frac{\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}}{\sqrt{n}}\left(\mathbf{X}^{11}\frac{\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}}{\sqrt{n}} + \mathbf{X}^{12}\frac{\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}}{\sqrt{n}}\right) \\
&\qquad + \frac{\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}}{\sqrt{n}}\left(\mathbf{X}^{11}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{C}}\mathbf{X}^{11} + \mathbf{X}^{11}\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}\mathbf{X}^{12} + \mathbf{X}^{12}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{C}}\mathbf{X}^{11} + \mathbf{X}^{12}\mathbf{X}_{\mathcal{D}}^T\mathbf{X}_{\mathcal{D}}\mathbf{X}^{12}\right)\frac{\mathbf{X}_{\mathcal{C}}^T\mathbf{X}_{\mathcal{D}}}{\sqrt{n}} \\
&\quad \to \Sigma_{\mathcal{D}\mathcal{D}} - (\Sigma_{\mathcal{D}\mathcal{C}}\Sigma_{11} + \Sigma_{\mathcal{D}\mathcal{D}}\Sigma_{12})\Sigma_{\mathcal{C}\mathcal{D}} - \Sigma_{\mathcal{D}\mathcal{C}}(\Sigma_{11}\Sigma_{\mathcal{C}\mathcal{D}} + \Sigma_{12}\Sigma_{\mathcal{D}\mathcal{D}}) \\
&\qquad + \Sigma_{\mathcal{D}\mathcal{C}}(\Sigma_{11}\Sigma_{\mathcal{C}\mathcal{C}}\Sigma_{11} + \Sigma_{11}\Sigma_{\mathcal{C}\mathcal{D}}\Sigma_{12} + \Sigma_{12}\Sigma_{\mathcal{D}\mathcal{C}}\Sigma_{11} + \Sigma_{12}\Sigma_{\mathcal{D}\mathcal{D}}\Sigma_{12})\Sigma_{\mathcal{C}\mathcal{D}} \\
&\quad = \Sigma_{\mathcal{D}\mathcal{D}} - 2\Sigma_{\mathcal{D}\mathcal{C}}\Sigma_{11}\Sigma_{\mathcal{C}\mathcal{D}} - 2\Sigma_{\mathcal{D}\mathcal{D}}\Sigma_{12}\Sigma_{\mathcal{D}\mathcal{D}} \\
&\qquad + \Sigma_{\mathcal{D}\mathcal{C}}(\Sigma_{11}\Sigma_{\mathcal{C}\mathcal{C}}\Sigma_{11} + \Sigma_{11}\Sigma_{\mathcal{C}\mathcal{D}}\Sigma_{12} + \Sigma_{12}\Sigma_{\mathcal{D}\mathcal{C}}\Sigma_{11} + \Sigma_{12}\Sigma_{\mathcal{D}\mathcal{D}}\Sigma_{12})\Sigma_{\mathcal{C}\mathcal{D}} \\
&\quad \equiv \Sigma_{*\mathcal{D}}.
\end{aligned}$$

Now consider the limiting behavior of the fourth term in $V^{(n)}(\mathbf{u}_{\mathcal{D}})$. If $\beta_j^* \neq 0$, then $\hat{w}_j \to_p |\beta_j^*|^{-\gamma}$ and $\sqrt{n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}| - |\beta_j^*|) \to u_j sign(\beta_j^*)$. By Slutsky's theorem, we have $\frac{\lambda_n}{n}\hat{w}_j\sqrt{n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}| - |\beta_j^*|) \to_p 0$. If β_0^*, then $\sqrt{n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}| - |\beta_j^*|) = |u_j|$ and $\frac{\lambda_n}{n}\hat{w}_j = \frac{\lambda_n}{n}n^{\gamma/2}|\sqrt{n}\hat{\beta}_j|^{-\gamma}$, where $\sqrt{n}\hat{\beta}_j = O_p(1)$. Thus, again, by Slutsky's theorem, we see that $V^{(n)}(\mathbf{u}_{\mathcal{D}}) \to_d V(\mathbf{u}_{\mathcal{D}})$ for every $\mathbf{u}_{\mathcal{D}}$, where

$$V(\mathbf{u}_{\mathcal{D}}) = \begin{cases} \mathbf{u}_{\mathcal{A}}^T\Sigma_{ss}\mathbf{u}_{\mathcal{A}} - 2\mathbf{u}_{\mathcal{A}}^T\mathbf{W}_{\mathcal{A}} & \text{if } u_j = 0 \text{ for } j \notin \mathcal{A} \\ \infty & \text{otherwise,} \end{cases}$$

where $\mathbf{W}_{\mathcal{A}} = N(\mathbf{0}, \sigma^2\Sigma_{*\mathcal{A}})$ and $\Sigma_{*\mathcal{A}}$ is the upper left $s \times s$ corner of $\Sigma_{*\mathcal{D}}$. $V^{(n)}(\mathbf{u}_{\mathcal{D}})$ is convex, and the unique minimum of $V(\mathbf{u}_{\mathcal{D}})$ is $(\Sigma_{ss}^{-1}\mathbf{W}_{\mathcal{A}}, \mathbf{0})^T$. Following the epi-convergence results of Geyer (1994) and Knight and Fu (2000), we have

$$\hat{\mathbf{u}}_{\mathcal{A}} \to_d \Sigma_{ss}^{-1}\mathbf{W}_{\mathcal{A}} \qquad and \qquad \hat{\mathbf{u}}_{\mathcal{A}^c} \to_d \mathbf{0}. \tag{15}$$

Then we finish proving the asymptotic normality part.

Now we consider the consistency part. $\forall j \in \mathcal{A}$, the asymptotic normality result indicates that $\hat{\beta}_j^{\mathrm{CAL}} \to_p \beta_j^*$; thus $P(j \in \mathcal{A}_n) \to 1$. Then it suffices to show that $\forall j' \notin \mathcal{A}$, $P(j' \in \mathcal{A}_n) \to 0$. Consider the event $j' \in \mathcal{A}_n$. By the KKT optimalirt conditions, we know that $2\mathbf{x}_{j'}^T(\mathbf{y}-\mathbf{X}_{\mathcal{C}}\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}-\mathbf{X}_{\mathcal{D}}\hat{\boldsymbol{\beta}}^{CAL}) = \lambda_n \hat{w}_j$. Note that $\lambda_n\hat{w}_j/\sqrt{n} = \frac{\lambda_n}{\sqrt{n}} n^{\gamma/2} \frac{1}{\sqrt{n}\hat{\beta}_{j'}} \to_p \infty$, whereas

$$2\frac{\mathbf{x}_{j'}^T(\mathbf{y}-\mathbf{X}_{\mathcal{C}}\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}-\mathbf{X}_{\mathcal{D}}\hat{\boldsymbol{\beta}}^{\mathrm{CAL}})}{\sqrt{n}} = 2\frac{\mathbf{x}_{j'}^T\boldsymbol{\varepsilon}}{\sqrt{n}} + 2\frac{\mathbf{x}_{j'}^T\mathbf{X}_{\mathcal{C}}}{\sqrt{n}}\sqrt{n}(\boldsymbol{\beta}_{\mathcal{C}}^* - \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}) + 2\frac{\mathbf{x}_{j'}^T\mathbf{X}_{\mathcal{D}}}{\sqrt{n}}\sqrt{n}(\boldsymbol{\beta}_{\mathcal{D}}^* - \hat{\boldsymbol{\beta}}^{CAL}).$$

And we know that $2\mathbf{x}_{j'}^T\mathbf{X}_{\mathcal{C}}\sqrt{n}(\boldsymbol{\beta}_{\mathcal{C}}^* - \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)})/\sqrt{n} + 2\mathbf{x}_{j'}^T\mathbf{X}_{\mathcal{D}}\sqrt{n}(\boldsymbol{\beta}_{\mathcal{D}}^* - \hat{\boldsymbol{\beta}}^{CAL})/\sqrt{n} \to_d$ some normal distribution and $2\mathbf{x}_{j'}^T\boldsymbol{\varepsilon}/\sqrt{n} \to_d \mathrm{N}(\mathbf{0}, 4\|\mathbf{x}_{j'}^T\|^2\sigma^2)$. Thus

$$P(j' \in \mathcal{A}_n) \leqslant P(2\mathbf{x}_{j'}^T(\mathbf{y}-\mathbf{X}_{\mathcal{C}}\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}-\mathbf{X}_{\mathcal{D}}\hat{\boldsymbol{\beta}}^{\mathrm{CAL}}) = \lambda_n\hat{w}_j) \to 0.$$

Proof of Theorem 2

Since $\mathbf{X}_{\mathcal{C}}$ part is not penalized, it's easily shown that it's a MLE and has the asymptotic behavior. Then the proof is the same as in Theorem 1.

B Supplementary Materials for Chapter 3

Proof of Theorem 3

Before we prove the Theorem 3, similarly with Zou (2006), we assume the following regularity conditions:

1. The Fisher information matrix is finite and positive definite,

$$\mathbf{I}(\boldsymbol{\beta}^*) = E[\phi''(\mathbf{x}^T\boldsymbol{\beta}^*)\mathbf{x}\mathbf{x}^T]. \tag{16}$$

2. There is a sufficiently large enough open set $\mathcal{O}$ that contains $\boldsymbol{\beta}^*$ such that $\forall \boldsymbol{\beta} \in \mathcal{O}$,

$$|\phi'''(\mathbf{x}^T\boldsymbol{\beta})| \leq M(\mathbf{x}) < \infty \tag{17}$$

and

$$E[M(\mathbf{x})|x_j x_k x_l|] < \infty \tag{18}$$

for all $q+1 \leq j,k,l \leq p$, that is, $j,k,l \in \mathcal{D}$.

We start the proof by the asymptotic normality part. Let $\mathbf{u}_\mathcal{C} = \sqrt{n}(\hat{\boldsymbol{\beta}}_\mathcal{C}^{(n)} - \boldsymbol{\beta}_\mathcal{C}^*)$ and $\mathbf{u}_\mathcal{D} = \sqrt{n}(\boldsymbol{\beta}_\mathcal{D} - \boldsymbol{\beta}_\mathcal{D}^*)$, then $\hat{\boldsymbol{\beta}}_\mathcal{C}^{(n)} = \boldsymbol{\beta}_\mathcal{C}^* + \frac{\mathbf{u}_\mathcal{C}}{\sqrt{n}}$ and $\boldsymbol{\beta}_\mathcal{D} = \boldsymbol{\beta}_\mathcal{D}^* + \frac{\mathbf{u}_\mathcal{D}}{\sqrt{n}}$. Define

$$\begin{aligned}\Gamma_n(\mathbf{u}_\mathcal{D}) = \sum_{i=1}^n \Bigg\{ &- y_i\Big[\mathbf{x}_{i\mathcal{C}}^T(\boldsymbol{\beta}_\mathcal{C}^* + \frac{\mathbf{u}_\mathcal{C}}{\sqrt{n}}) + \mathbf{x}_{i\mathcal{D}}^T(\boldsymbol{\beta}_\mathcal{D}^* + \frac{\mathbf{u}_\mathcal{D}}{\sqrt{n}})\Big] \\ &+ \phi\Big[\mathbf{x}_{i\mathcal{C}}^T(\boldsymbol{\beta}_\mathcal{C}^* + \frac{\mathbf{u}_\mathcal{C}}{\sqrt{n}}) + \mathbf{x}_{i\mathcal{D}}^T(\boldsymbol{\beta}_\mathcal{D}^* + \frac{\mathbf{u}_\mathcal{D}}{\sqrt{n}})\Big]\Bigg\} + \lambda_n \sum_{i\in\mathcal{D}} \hat{w}_j\Big|\beta_j^* + \frac{u_j}{\sqrt{n}}\Big|.\end{aligned}$$

Let $\hat{\mathbf{u}}_\mathcal{D} = \operatorname{argmin}_{\mathbf{u}_\mathcal{D}} \Gamma_n(\mathbf{u}_\mathcal{D})$, then $\hat{\mathbf{u}}_\mathcal{D}^{(n)} = \sqrt{n}(\boldsymbol{\beta}_\mathcal{D}^{*(n)} - \boldsymbol{\beta}_\mathcal{D}^*)$. Let

$$\begin{aligned}H^{(n)}(\mathbf{u}_\mathcal{D}) &= \Gamma_n(\mathbf{u}_\mathcal{D}) - \Gamma_n(\mathbf{0}) \\ &= \sum_{i=1}^n \Bigg\{ - y_i\mathbf{x}_{i\mathcal{D}}^T\frac{\mathbf{u}_\mathcal{D}}{\sqrt{n}} + \phi\Big[\mathbf{x}_{i\mathcal{C}}^T(\boldsymbol{\beta}_\mathcal{C}^* + \frac{\mathbf{u}_\mathcal{C}}{\sqrt{n}}) + \mathbf{x}_{i\mathcal{D}}^T(\boldsymbol{\beta}_\mathcal{D}^* + \frac{\mathbf{u}_\mathcal{D}}{\sqrt{n}})\Big] \\ &\quad - \phi\Big[\mathbf{x}_{i\mathcal{C}}^T(\boldsymbol{\beta}_\mathcal{C}^* + \frac{\mathbf{u}_\mathcal{C}}{\sqrt{n}}) + \mathbf{x}_{i\mathcal{D}}^T\boldsymbol{\beta}_\mathcal{D}^*\Big]\Bigg\} + \lambda_n \sum_{i\in\mathcal{D}} \hat{w}_j\left[\Big|\beta_j^* + \frac{u_j}{\sqrt{n}}\Big| - \Big|\beta_j^*\Big|\right].\end{aligned}$$

Then using the Taylor expansion, we have

$$H^{(n)}(\mathbf{u}_{\mathcal{D}}) \equiv A_1^{(n)} + A_2^{(n)} + A_3^{(n)} + A_4^{(n)},$$

with

$$\begin{aligned}
A_1^{(n)} &= -\sum_{i=1}^{n} \left\{ y_i - \phi'\Big[\mathbf{x}_{i\mathcal{C}}^T(\boldsymbol{\beta}_{\mathcal{C}}^* + \frac{\mathbf{u}_{\mathcal{C}}}{\sqrt{n}}) + \mathbf{x}_{i\mathcal{D}}^T\boldsymbol{\beta}_{\mathcal{D}}^*\Big] \right\} \frac{\mathbf{x}_{i\mathcal{D}}^T\mathbf{u}_{\mathcal{D}}}{\sqrt{n}}, \\
A_2^{(n)} &= \sum_{i=1}^{n} \frac{1}{2}\phi''\Big[\mathbf{x}_{i\mathcal{C}}^T(\boldsymbol{\beta}_{\mathcal{C}}^* + \frac{\mathbf{u}_{\mathcal{C}}}{\sqrt{n}}) + \mathbf{x}_{i\mathcal{D}}^T\boldsymbol{\beta}_{\mathcal{D}}^*\Big] \mathbf{u}_{\mathcal{D}}^T \frac{\mathbf{x}_{i\mathcal{D}}\mathbf{x}_{i\mathcal{D}}^T}{n}\mathbf{u}_{\mathcal{D}}, \\
A_3^{(n)} &= \frac{\lambda_n}{\sqrt{n}} \sum_{i\in\mathcal{D}} \hat{w}_j\sqrt{n}\left[\Big|\beta_j^* + \frac{u_j}{\sqrt{n}}\Big| - \Big|\beta_j^*\Big| \right],
\end{aligned}$$

and

$$A_4^{(n)} = n^{-3/2}\sum_{i=1}^{n} \frac{1}{6}\phi'''\Big[\mathbf{x}_{i\mathcal{C}}^T(\boldsymbol{\beta}_{\mathcal{C}}^* + \frac{\mathbf{u}_{\mathcal{C}}}{\sqrt{n}}) + \mathbf{x}_{i\mathcal{D}}^T\tilde{\boldsymbol{\beta}}_{\mathcal{D}}^*\Big](\mathbf{x}_{i\mathcal{D}}^T\mathbf{u}_{\mathcal{D}})^3,$$

where $\tilde{\boldsymbol{\beta}}_{\mathcal{D}}^*$ is between $\boldsymbol{\beta}_{\mathcal{D}}^*$ and $\boldsymbol{\beta}_{\mathcal{D}}^* + \frac{\mathbf{u}_{\mathcal{D}}}{\sqrt{n}}$. We now analyze the asymptotic behavior of each term. By the well known properties of the exponential family,

$$E[y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)|\mathbf{x}_i, \boldsymbol{\beta}^*] = 0, \tag{19}$$

and

$$Var[y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)|\mathbf{x}_i, \boldsymbol{\beta}^*] = E\big\{[y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)]^2|\mathbf{x}_i, \boldsymbol{\beta}^*\big\} = \phi''(\mathbf{x}_i^T\boldsymbol{\beta}^*), \tag{20}$$

then we have

$$E\Big\{[y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)](\mathbf{x}_{i\mathcal{D}}^T\mathbf{u}_{\mathcal{D}})\Big\} = E\Big\{E(y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)|\mathbf{x}_i, \boldsymbol{\beta}^*)(\mathbf{x}_{i\mathcal{D}}^T\mathbf{u}_{\mathcal{D}})\Big\} = 0, \tag{21}$$

and

$$
\begin{aligned}
Var\Big\{[Y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)](\mathbf{x}_{i\mathcal{D}}^T\mathbf{u}_{\mathcal{D}})\Big\} &= E\Big\{E\big\{[Y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)]^2|\mathbf{x}_i, \boldsymbol{\beta}^*\big\}^2(\mathbf{x}_{i\mathcal{D}}^T\mathbf{u}_{\mathcal{D}})^2\Big\} \\
&= \mathbf{u}_{\mathcal{D}}^T E\big[\phi''(\mathbf{x}_i^T\boldsymbol{\beta}^*)\mathbf{x}_{i\mathcal{D}}\mathbf{x}_{i\mathcal{D}}^T\big]\mathbf{u}_{\mathcal{D}} \\
&= \mathbf{u}_{\mathcal{D}}^T\mathbf{I}_{\mathcal{D}\mathcal{D}}\mathbf{u}_{\mathcal{D}}.
\end{aligned} \tag{22}
$$

Also note that, since $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$ is a root-n consistent estimate of $\boldsymbol{\beta}_{\mathcal{C}}^*$, $\frac{\mathbf{u}_{\mathcal{C}}}{\sqrt{n}} = (\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} - \boldsymbol{\beta}_{\mathcal{C}}^*) \xrightarrow{p} 0$. Therefore,

$$
\phi'\Big[\mathbf{x}_{i\mathcal{C}}^T(\boldsymbol{\beta}_{\mathcal{C}}^* + \frac{\mathbf{u}_{\mathcal{C}}}{\sqrt{n}}) + \mathbf{x}_{i\mathcal{D}}^T\boldsymbol{\beta}_{\mathcal{D}}^*\Big] \xrightarrow{p} \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*), \tag{23}
$$

and

$$
\phi''\Big[\mathbf{x}_{i\mathcal{C}}^T(\boldsymbol{\beta}_{\mathcal{C}}^* + \frac{\mathbf{u}_{\mathcal{C}}}{\sqrt{n}}) + \mathbf{x}_{i\mathcal{D}}^T\boldsymbol{\beta}_{\mathcal{D}}^*\Big] \xrightarrow{p} \phi''(\mathbf{x}_i^T\boldsymbol{\beta}^*). \tag{24}
$$

By the central limit theorem and the Slutsky's theorem, we have

$$
A_1^{(n)} \xrightarrow{d} -\mathbf{u}_{\mathcal{D}}^T\mathrm{N}(\mathbf{0}, \mathbf{I}_{\mathcal{D}\mathcal{D}}). \tag{25}
$$

For the second term $A_2^{(n)}$, we observe that

$$
\sum_{i=1}^{n} \phi''\Big[\mathbf{x}_{i\mathcal{C}}^T(\boldsymbol{\beta}_{\mathcal{C}}^* + \frac{\mathbf{u}_{\mathcal{C}}}{\sqrt{n}}) + \mathbf{x}_{i\mathcal{D}}^T\boldsymbol{\beta}_{\mathcal{D}}^*\Big]\frac{\mathbf{x}_{i\mathcal{D}}\mathbf{x}_{i\mathcal{D}}^T}{n} \xrightarrow{p} \mathbf{I}_{\mathcal{D}\mathcal{D}}. \tag{26}
$$

Thus,

$$
A_2^{(n)} \xrightarrow{p} \frac{1}{2}\mathbf{u}_{\mathcal{D}}^T\mathbf{I}_{\mathcal{D}\mathcal{D}}\mathbf{u}_{\mathcal{D}}. \tag{27}
$$

Since the limiting behavior of the third term $A_3^{(n)}$ has been discussed in the proof of

Theorem 1, here we just list the results as follows:

$$\frac{\lambda_n}{\sqrt{n}}\hat{w}_j\sqrt{n}\left[\left|\beta_j^* + \frac{u_j}{\sqrt{n}}\right| - \left|\beta_j^*\right|\right] \xrightarrow{p} \begin{cases} 0 & \text{if } \beta_j^* \neq 0 \text{ for } j \in \mathcal{D}, \\ 0 & \text{if } \beta_j^* = 0 \text{ and } u_j = 0 \text{ for } j \in \mathcal{D}, \\ \infty & \text{if } \beta_j^* = 0 \text{ and } u_j \neq 0 \text{ for } j \in \mathcal{D}. \end{cases} \tag{28}$$

For the fourth term $A_4^{(n)}$, we observe that

$$\phi'''\left[\mathbf{x}_{i\mathcal{C}}^T(\boldsymbol{\beta}_{\mathcal{C}}^* + \frac{\mathbf{u}_{\mathcal{C}}}{\sqrt{n}}) + \mathbf{x}_{i\mathcal{D}}^T\tilde{\boldsymbol{\beta}}_{\mathcal{D}}^*\right] = \phi'''(\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}^*), \tag{29}$$

where $\tilde{\boldsymbol{\beta}}^*$ is between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}$. By the regularity condition 2, the fourth term $A_4^{(n)}$ can be bounded as

$$6\sqrt{n}A_4^{(n)} \leq \sum_{i=1}^{n} \frac{1}{n} M(\mathbf{x})|\mathbf{x}_{i\mathcal{D}}^T\mathbf{u}_{\mathcal{D}}|^3 \xrightarrow{p} E[M(\mathbf{x})|\mathbf{x}_{\mathcal{D}}^T\mathbf{u}_{\mathcal{D}}|^3] < \infty. \tag{30}$$

Therefore, by Slutsky's theorem, we see that $H^{(n)}(\mathbf{u}_{\mathcal{D}}) \xrightarrow{d} H(\mathbf{u}_{\mathcal{D}})$ for every $\mathbf{u}_{\mathcal{D}}$, where

$$H(\mathbf{u}_{\mathcal{D}}) = \begin{cases} \frac{1}{2}\mathbf{u}_{\mathcal{A}}^T\mathbf{I}_{ss}\mathbf{u}_{\mathcal{A}} - \mathbf{u}_{\mathcal{A}}^T\mathbf{W}_{\mathcal{A}} & \text{if } u_j = 0 \ \forall j \in \mathcal{A}^c, \\ \infty & \text{otherwise,} \end{cases} \tag{31}$$

where $\mathbf{W} = \mathrm{N}(\mathbf{0}, \mathbf{I}_{\mathcal{D}\mathcal{D}})$, $H^{(n)}$ is convex and the unique minimum of H is $(\mathbf{I}_{ss}^{-1}W_{\mathcal{A}}, 0)^T$. Then we have

$$\hat{\mathbf{u}}_{\mathcal{A}} \xrightarrow{d} \mathbf{I}_{ss}^{-1}\mathbf{W}_{\mathcal{A}} \qquad \text{and} \qquad \hat{\mathbf{u}}_{\mathcal{A}^c} \xrightarrow{d} \mathbf{0}. \tag{32}$$

Since $\mathbf{W}_{\mathcal{A}} = \mathrm{N}(\mathbf{0}, \mathbf{I}_{ss})$, the asymptotic normality part is proven.

Now we show the consistency part. For $\forall j \in \mathcal{A}$, the asymptotic normality indicates that $\hat{\beta}_j^{(n)}(\text{gcal}) \xrightarrow{p} \beta_j^*$. Hence, it suffices to show that for $\forall j' \notin \mathcal{A}$, $P(j' \in$

$\mathcal{A}_n) \to 0$. Consider the event $j' \in \mathcal{A}_n$. By the KKT optimality conditions, we must have

$$\sum_{i=1}^{n} x_{ij'} \Big[y_i - \phi' \big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}(\text{gcal}) \big) \Big] = \lambda_n \hat{w}_{j'}.$$

Thus, we have

$$P(j' \in \mathcal{A}_n) \leq P\Big(\sum_{i=1}^{n} x_{ij'} \Big[y_i - \phi' \big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}(\text{gcal}) \big) \Big] = \lambda_n \hat{w}_{j'} \Big).$$

By using Taylor expansion, we can have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ij'} \Big[y_i - \phi' \big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}(\text{gcal}) \big) \Big] = B_0^{(n)} + B_1^{(n)} + B_2^{(n)}.$$

with

$$\begin{aligned} B_0^{(n)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ij'} \Big[y_i - \phi' \big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}}^{*} \big) \Big], \\ B_1^{(n)} &= \left[\frac{1}{n} \sum_{i=1}^{n} x_{ij'} \phi'' \big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}}^{*} \big) \mathbf{x}_{i\mathcal{D}}^T \right] \sqrt{n} (\boldsymbol{\beta}_{\mathcal{D}}^{*} - \hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}(\text{gcal})), \end{aligned}$$

and

$$B_2^{(n)} = -\frac{1}{\sqrt{n}} \left[\frac{1}{n} \sum_{i=1}^{n} x_{ij'} \phi''' \big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \tilde{\boldsymbol{\beta}}_{\mathcal{D}} \big) \right] \Big(\mathbf{x}_{i\mathcal{D}}^T \sqrt{n} (\hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}(\text{gcal}) - \boldsymbol{\beta}_{\mathcal{D}}^{*}) \Big)^2,$$

where $\tilde{\boldsymbol{\beta}}_{\mathcal{D}}$ is between $\hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}(\text{gcal})$ and $\boldsymbol{\beta}_{\mathcal{D}}^{*}$. By the previous proof, we know that

$$B_0^{(n)} \xrightarrow{d} \text{N}(0, \text{I}_{j'j'}).$$

Similarly by previous proof, we have

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij'}\phi''\big(\mathbf{x}_{i\mathcal{C}}^T\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T\boldsymbol{\beta}_{\mathcal{D}}^{*}\big)\mathbf{x}_{i\mathcal{D}}^T \xrightarrow{p} \mathbf{I}_{j'},$$

where $\mathbf{I}_{j'}$ is the j'th row of $\mathbf{I}_{\mathcal{D}\mathcal{D}}$. Thus, combining (53), it implies that

$$B_1^{(n)} \xrightarrow{d} \text{ some random variable.}$$

From the regularity condition 2 and (53), we observe that

$$B_2^{(n)} = O_p\Big(\frac{1}{\sqrt{n}}\Big).$$

By the assumptions of the theorem, we also have

$$\frac{\lambda_n \hat{w}_{j'}}{\sqrt{n}} = \frac{\lambda_n}{\sqrt{n}} n^{\gamma/2} \frac{1}{|\sqrt{n}\hat{\beta}_{j'}(\text{gcal})|^{\gamma}} \xrightarrow{p} \infty.$$

Therefore,

$$P(j' \in \mathcal{A}_n) \to 0.$$

The proof is finished.

Proof of Theorem 2

Since $\mathbf{X}_{\mathcal{C}}$ part is not penalized, it's easily shown that it's a MLE and has the asymptotic behavior. Then the proof is the same as in Theorem 1.

C Supplementary Materials for Chapter 4

Proof of Theorem 5

The proof is constructed based on the environment setup in Zou (2006). Similarly with Zou (2006), we assume the following regularity conditions:

1. The Fisher information matrix is finite and positive definite,

$$\mathbf{I}(\boldsymbol{\beta}^*) = E[\phi''(\mathbf{x}^T\boldsymbol{\beta}^*)\mathbf{x}\mathbf{x}^T]. \tag{33}$$

2. There is a sufficiently large enough open set $\mathcal{O}$ that contains $\boldsymbol{\beta}^*$ such that $\forall \boldsymbol{\beta} \in \mathcal{O}$,

$$|\phi'''(\mathbf{x}^T\boldsymbol{\beta})| \leq M(\mathbf{x}) < \infty \tag{34}$$

and

$$E[M(\mathbf{x})|x_j x_k x_l|] < \infty \tag{35}$$

for all $q+1 \leq j,k,l \leq p$, that is, $j,k,l \in \mathcal{D}$.

We start the proof by the asymptotic normality part. Let $\mathbf{u}_{\mathcal{D}} = \sqrt{n}(\boldsymbol{\beta}_{\mathcal{D}} - \boldsymbol{\beta}^*_{\mathcal{D}})$, then $\boldsymbol{\beta}_{\mathcal{D}} = \boldsymbol{\beta}^*_{\mathcal{D}} + \frac{\mathbf{u}_{\mathcal{D}}}{\sqrt{n}}$. Define

$$\Gamma_n(\mathbf{u}_{\mathcal{D}}) = \sum_{i=1}^{n} \Big\{ -y_i\Big[\mathbf{x}_{i\mathcal{C}}^T\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T(\boldsymbol{\beta}^*_{\mathcal{D}} + \frac{\mathbf{u}_{\mathcal{D}}}{\sqrt{n}})\Big] + \phi\Big[\mathbf{x}_{i\mathcal{C}}^T\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T(\boldsymbol{\beta}^*_{\mathcal{D}} + \frac{\mathbf{u}_{\mathcal{D}}}{\sqrt{n}})\Big] \Big\}$$
$$+ n\sum_{j\in\mathcal{D}} p_{\lambda_n}\big(\Big|\boldsymbol{\beta}^*_{\mathcal{D}} + \frac{\mathbf{u}_{\mathcal{D}}}{\sqrt{n}}\Big|\big).$$

Let $\hat{\mathbf{u}}_{\mathcal{D}} = \text{argmin}_{\mathbf{u}_{\mathcal{D}}}\, \Gamma_n(\mathbf{u}_{\mathcal{D}})$, then $\hat{\mathbf{u}}^{(n)}_{\mathcal{D}} = \sqrt{n}(\hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{D}} - \boldsymbol{\beta}^*_{\mathcal{D}})$. Let

$$
\begin{aligned}
H^{(n)}(\mathbf{u}_{\mathcal{D}}) &= \Gamma_n(\mathbf{u}_{\mathcal{D}}) - \Gamma_n(\mathbf{0}) \\
&= \sum_{i=1}^{n} \Big\{ - y_i \mathbf{x}_{i\mathcal{D}}^T \frac{\mathbf{u}_{\mathcal{D}}}{\sqrt{n}} + \phi\Big[\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T (\boldsymbol{\beta}_{\mathcal{D}}^* + \frac{\mathbf{u}_{\mathcal{D}}}{\sqrt{n}})\Big] - \phi\Big[\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}}^*\Big] \Big\} \\
&\qquad + \sum_{j \in \mathcal{D}} n\Big[p_{\lambda_n}\big(\Big|\boldsymbol{\beta}_{\mathcal{D}}^* + \frac{\mathbf{u}_{\mathcal{D}}}{\sqrt{n}}\Big|\big) - p_{\lambda_n}\big(\Big|\boldsymbol{\beta}_{\mathcal{D}}^*\Big|\big)\Big].
\end{aligned}
$$

Then using the Taylor expansion, we have

$$
H^{(n)}(\mathbf{u}_{\mathcal{D}}) \equiv A_1^{(n)} + A_2^{(n)} + A_3^{(n)} + A_4^{(n)},
$$

with

$$
\begin{aligned}
A_1^{(n)} &= -\sum_{i=1}^{n} \Big\{ y_i - \phi'\Big[\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}}^*\Big] \Big\} \frac{\mathbf{x}_{i\mathcal{D}}^T \mathbf{u}_{\mathcal{D}}}{\sqrt{n}}, \\
A_2^{(n)} &= \sum_{i=1}^{n} \frac{1}{2} \phi''\Big[\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}}^*\Big] \mathbf{u}_{\mathcal{D}}^T \frac{\mathbf{x}_{i\mathcal{D}} \mathbf{x}_{i\mathcal{D}}^T}{n} \mathbf{u}_{\mathcal{D}}, \\
A_3^{(n)} &= \sum_{j \in \mathcal{D}} n\Big[p_{\lambda_n}\big(\Big|\boldsymbol{\beta}_{\mathcal{D}}^* + \frac{\mathbf{u}_{\mathcal{D}}}{\sqrt{n}}\Big|\big) - p_{\lambda_n}\big(\Big|\boldsymbol{\beta}_{\mathcal{D}}^*\Big|\big)\Big],
\end{aligned}
$$

and

$$
A_4^{(n)} = n^{-3/2} \sum_{i=1}^{n} \frac{1}{6} \phi'''\Big[\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \tilde{\boldsymbol{\beta}}_{\mathcal{D}}^*\Big] (\mathbf{x}_{i\mathcal{D}}^T \mathbf{u}_{\mathcal{D}})^3,
$$

where $\tilde{\boldsymbol{\beta}}_{\mathcal{D}}^*$ is between $\boldsymbol{\beta}_{\mathcal{D}}^*$ and $\boldsymbol{\beta}_{\mathcal{D}}^* + \frac{\mathbf{u}_{\mathcal{D}}}{\sqrt{n}}$. We now analyze the asymptotic behavior of each term. By the well known properties of the exponential family,

$$
E[y_i - \phi'(\mathbf{x}_i^T \boldsymbol{\beta}^*) | \mathbf{x}_i, \boldsymbol{\beta}^*] = 0, \tag{36}
$$

and

$$Var[y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)|\mathbf{x}_i, \boldsymbol{\beta}^*] = E\big\{[y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)]^2|\mathbf{x}_i, \boldsymbol{\beta}^*\big\} = \phi''(\mathbf{x}_i^T\boldsymbol{\beta}^*), \tag{37}$$

then we have

$$E\Big\{[y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)](\mathbf{x}_{i\mathcal{D}}^T\mathbf{u}_{\mathcal{D}})\Big\} = E\Big\{E(y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)|\mathbf{x}_i, \boldsymbol{\beta}^*)(\mathbf{x}_{i\mathcal{D}}^T\mathbf{u}_{\mathcal{D}})\Big\} = 0, \tag{38}$$

and

$$\begin{aligned} Var\Big\{[Y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)](\mathbf{x}_{i\mathcal{D}}^T\mathbf{u}_{\mathcal{D}})\Big\} &= E\Big\{E\big\{[Y_i - \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*)]^2|\mathbf{x}_i, \boldsymbol{\beta}^*\big\}^2(\mathbf{x}_{i\mathcal{D}}^T\mathbf{u}_{\mathcal{D}})^2\Big\} \\ &= \mathbf{u}_{\mathcal{D}}^T E\big[\phi''(\mathbf{x}_i^T\boldsymbol{\beta}^*)\mathbf{x}_{i\mathcal{D}}\mathbf{x}_{i\mathcal{D}}^T\big]\mathbf{u}_{\mathcal{D}} \\ &= \mathbf{u}_{\mathcal{D}}^T\mathbf{I}_{\mathcal{D}\mathcal{D}}\mathbf{u}_{\mathcal{D}}. \end{aligned} \tag{39}$$

Also note that, since $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}$ is a root-n consistent estimate of $\boldsymbol{\beta}_{\mathcal{C}}^*$, $\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} \xrightarrow{p} \boldsymbol{\beta}_{\mathcal{C}}^*$. Therefore,

$$\phi'\Big[\mathbf{x}_{i\mathcal{C}}^T\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T\boldsymbol{\beta}_{\mathcal{D}}^*\Big] \xrightarrow{p} \phi'(\mathbf{x}_i^T\boldsymbol{\beta}^*), \tag{40}$$

and

$$\phi''\Big[\mathbf{x}_{i\mathcal{C}}^T\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T\boldsymbol{\beta}_{\mathcal{D}}^*\Big] \xrightarrow{p} \phi''(\mathbf{x}_i^T\boldsymbol{\beta}^*). \tag{41}$$

By the central limit theorem and the Slutsky's theorem, we have

$$A_1^{(n)} \xrightarrow{d} -\mathbf{u}_{\mathcal{D}}^T\mathrm{N}(\mathbf{0}, \mathbf{I}_{\mathcal{D}\mathcal{D}}). \tag{42}$$

For the second term $A_2^{(n)}$, we observe that

$$\sum_{i=1}^{n} \phi''\Big[\mathbf{x}_{i\mathcal{C}}^T(\boldsymbol{\beta}_{\mathcal{C}}^* + \frac{\mathbf{u}_{\mathcal{C}}}{\sqrt{n}}) + \mathbf{x}_{i\mathcal{D}}^T\boldsymbol{\beta}_{\mathcal{D}}^*\Big]\frac{\mathbf{x}_{i\mathcal{D}}\mathbf{x}_{i\mathcal{D}}^T}{n} \xrightarrow{p} \mathbf{I}_{\mathcal{D}\mathcal{D}}. \tag{43}$$

Thus,

$$A_2^{(n)} \xrightarrow{p} \frac{1}{2}\mathbf{u}_{\mathcal{D}}^T \mathbf{I}_{\mathcal{D}\mathcal{D}} \mathbf{u}_{\mathcal{D}}. \tag{44}$$

Now consider the limiting behavior of the third term $A_3^{(n)}$. If $\beta_j^* \neq 0$, using Taylor expansion, we can have

$$n\Big[p_{\lambda_n}\big(\Big|\beta_j^* + \frac{u_j}{\sqrt{n}}\Big|\big) - p_{\lambda_n}\big(|\beta_j^*|\big)\Big] = n\Big[\mathrm{sgn}(\beta_j^*)p'_{\lambda_n}(|\beta_j^*|)\frac{u_j}{\sqrt{n}} + \frac{1}{2}p''_{\lambda_n}(|\beta_j^*|)\frac{u_j^2}{n}\big(1+o(1)\big)\Big] \tag{45}$$

$$= \mathrm{sgn}(\beta_j^*)\sqrt{n}p'_{\lambda_n}(|\beta_j^*|)u_j + \frac{1}{2}p''_{\lambda_n}(|\beta_j^*|)u_j^2\big(1+o(1)\big). \tag{46}$$

Since $\max\{p''_{\lambda_n}(|\beta_j^*|) : \beta_j^* \neq 0\} \to 0$ as $n \to \infty$ and $\sqrt{n}p'_{\lambda_n}(|\beta_j^*|) \to 0$ for $j \in \{k \in \mathcal{D} : \beta_k^* \neq 0\}$ as $n \to \infty$, then $n\Big[p_{\lambda_n}\big(\Big|\beta_j^* + \frac{u_j}{\sqrt{n}}\Big|\big) - p_{\lambda_n}\big(|\beta_j^*|\big)\Big] \xrightarrow{p} 0$. If $\beta_j^* = 0$ and $u_j = 0$, clearly

$$n\Big[p_{\lambda_n}\big(\Big|\beta_j^* + \frac{u_j}{\sqrt{n}}\Big|\big) - p_{\lambda_n}\big(|\beta_j^*|\big)\Big] = 0.$$

If $\beta_j^* = 0$ and $u_j \neq 0$,

$$n\Big[p_{\lambda_n}\big(\Big|\beta_j^* + \frac{u_j}{\sqrt{n}}\Big|\big) - p_{\lambda_n}\big(|\beta_j^*|\big)\Big] = np_{\lambda_n}\big(\Big|\frac{u_j}{\sqrt{n}}\Big|\big).$$

By Taylor expansion,

$$np_{\lambda_n}\big(\Big|\frac{u_j}{\sqrt{n}}\Big|\big) = n\Big[\mathrm{sgn}(u_j)p'_{\lambda_n}(0)\frac{u_j}{\sqrt{n}} + \frac{1}{2}p''_{\lambda_n}(0)\frac{u_j^2}{n}\big(1+o(1)\big)\Big] \tag{47}$$

$$= \mathrm{sgn}(\beta_j^*)\sqrt{n}p'_{\lambda_n}(0)u_j + \frac{1}{2}p''_{\lambda_n}(0)u_j^2\big(1+o(1)\big). \tag{48}$$

Since $\sqrt{n}p'_{\lambda_n}(0) \to \infty$,

$$n\Big[p_{\lambda_n}\big(\Big|\beta_j^* + \frac{u_j}{\sqrt{n}}\Big|\big) - p_{\lambda_n}\big(|\beta_j^*|\big)\Big] \xrightarrow{p} \infty.$$

We conclude the limiting behavior of the third term $A_3^{(n)}$ as follows:

$$n\Big[p_{\lambda_n}\big(\Big|\beta_j^* + \frac{u_j}{\sqrt{n}}\Big|\big) - p_{\lambda_n}\big(|\beta_j^*|\big)\Big] \xrightarrow{p} \begin{cases} 0 & \text{if } \beta_j^* \neq 0 \text{ for } j \in \mathcal{D}, \\ 0 & \text{if } \beta_j^* = 0 \text{ and } u_j = 0 \text{ for } j \in \mathcal{D}, \\ \infty & \text{if } \beta_j^* = 0 \text{ and } u_j \neq 0 \text{ for } j \in \mathcal{D}. \end{cases} \tag{49}$$

For the fourth term $A_4^{(n)}$, we observe that

$$\phi'''\Big[\mathbf{x}_{i\mathcal{C}}^T\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T\tilde{\boldsymbol{\beta}}_{\mathcal{D}}^*\Big] = \phi'''(\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}^*), \tag{50}$$

where $\tilde{\boldsymbol{\beta}}^*$ is between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^* + \frac{\mathbf{u}}{\sqrt{n}}$. By the regularity condition 2, the fourth term $A_4^{(n)}$ can be bounded as

$$6\sqrt{n}A_4^{(n)} \leq \sum_{i=1}^{n} \frac{1}{n} M(\mathbf{x})|\mathbf{x}_{i\mathcal{D}}^T\mathbf{u}_{\mathcal{D}}|^3 \xrightarrow{p} E[M(\mathbf{x})|\mathbf{x}_{\mathcal{D}}^T\mathbf{u}_{\mathcal{D}}|^3] < \infty. \tag{51}$$

Therefore, by Slutsky's theorem, we see that $H^{(n)}(\mathbf{u}_{\mathcal{D}}) \xrightarrow{d} H(\mathbf{u}_{\mathcal{D}})$ for every $\mathbf{u}_{\mathcal{D}}$, where

$$H(\mathbf{u}_{\mathcal{D}}) = \begin{cases} \frac{1}{2}\mathbf{u}_{\mathcal{A}}^T\mathbf{I}_{ss}\mathbf{u}_{\mathcal{A}} - \mathbf{u}_{\mathcal{A}}^T\mathbf{W}_{\mathcal{A}} & \text{if } u_j = 0 \ \forall j \in \mathcal{A}^c, \\ \infty & \text{otherwise,} \end{cases} \tag{52}$$

where $\mathbf{W} = \mathrm{N}(\mathbf{0}, \mathbf{I}_{\mathcal{D}\mathcal{D}})$, $H^{(n)}$ is convex and the unique minimum of H is $(\mathbf{I}_{ss}^{-1}\mathbf{W}_{\mathcal{A}}, 0)^T$. Then we have

$$\hat{\mathbf{u}}_{\mathcal{A}} \xrightarrow{d} \mathbf{I}_{ss}^{-1}\mathbf{W}_{\mathcal{A}} \qquad \text{and} \qquad \hat{\mathbf{u}}_{\mathcal{A}^c} \xrightarrow{d} \mathbf{0}. \tag{53}$$

Since $\mathbf{W}_{\mathcal{A}} = \mathrm{N}(\mathbf{0}, \mathbf{I}_{ss})$, the asymptotic normality part is proven.

Now we show the consistency part. For $\forall j \in \mathcal{A}$, the asymptotic normality indicates that $\hat{\beta}_j^{(n)} \xrightarrow{p} \beta_j^*$. Hence, it suffices to show that for $\forall j' \notin \mathcal{A}$, $P(j' \in \mathcal{A}_n) \to 0$. Consider the event $j' \in \mathcal{A}_n$. By the KKT optimality conditions, we must have

$$\sum_{i=1}^{n} x_{ij'} \Big[y_i - \phi'\big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}\big) \Big] = n p'_{\lambda_n}(\hat{\beta}_{j'}^{(n)}).$$

Thus, we have

$$P(j' \in \mathcal{A}_n) \leq P\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ij'} \Big[y_i - \phi'\big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}\big) \Big] = \sqrt{n} p'_{\lambda_n}(\hat{\beta}_{j'}^{(n)}) \Big).$$

By using Taylor expansion, we can have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ij'} \Big[y_i - \phi'\big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}\big) \Big] = B_0^{(n)} + B_1^{(n)} + B_2^{(n)}$$

with

$$\begin{aligned}
B_0^{(n)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ij'} \Big[y_i - \phi'\big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}}^{*}\big) \Big], \\
B_1^{(n)} &= \left[\frac{1}{n} \sum_{i=1}^{n} x_{ij'} \phi''\big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \boldsymbol{\beta}_{\mathcal{D}}^{*}\big) \mathbf{x}_{i\mathcal{D}}^T \right] \sqrt{n} (\boldsymbol{\beta}_{\mathcal{D}}^{*} - \hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}),
\end{aligned}$$

and

$$B_2^{(n)} = -\frac{1}{\sqrt{n}} \left[\frac{1}{n} \sum_{i=1}^{n} x_{ij'} \phi'''\big(\mathbf{x}_{i\mathcal{C}}^T \hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T \tilde{\boldsymbol{\beta}}_{\mathcal{D}}\big) \right] \Big(\mathbf{x}_{i\mathcal{D}}^T \sqrt{n} (\hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)} - \boldsymbol{\beta}_{\mathcal{D}}^{*}) \Big)^2,$$

where $\tilde{\boldsymbol{\beta}}_{\mathcal{D}}$ is between $\hat{\boldsymbol{\beta}}_{\mathcal{D}}^{(n)}$ and $\boldsymbol{\beta}_{\mathcal{D}}^{*}$. By the previous proof, we know that

$$B_0^{(n)} \xrightarrow{d} \mathrm{N}(0, \mathrm{I}_{j'j'}).$$

Similarly by previous proof, we have

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij'}\phi''\big(\mathbf{x}_{i\mathcal{C}}^T\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)} + \mathbf{x}_{i\mathcal{D}}^T\boldsymbol{\beta}_{\mathcal{D}}^{*}\big)\mathbf{x}_{i\mathcal{D}}^T \xrightarrow{p} \mathbf{I}_{j'},$$

where $\mathbf{I}_{j'}$ is the j'th row of $\mathbf{I}_{\mathcal{D}\mathcal{D}}$. Thus, combining (53), it implies that

$$B_1^{(n)} \xrightarrow{d} \text{ some random variable.}$$

From the regularity condition 2 and (53), we observe that

$$B_2^{(n)} = O_p\Big(\frac{1}{\sqrt{n}}\Big).$$

By Taylor expansion again, we observe that

$$\sqrt{n}p'_{\lambda_n}(\hat{\beta}_{j'}^{(n)}) = \sqrt{n}p'_{\lambda_n}(0) + p''_{\lambda_n}(0)\sqrt{n}\hat{\beta}_{j'}^{(n)}\big(1 + o(1)\big).$$

By previous arguments and assumptions, $\sqrt{n}\hat{\beta}_{j'}^{(n)} \xrightarrow{d} 0$ and $\sqrt{n}p'_{\lambda_n}(0) \to \infty$. Thus,

$$\sqrt{n}p'_{\lambda_n}(\hat{\beta}_{j'}^{(n)}) \xrightarrow{p} \infty.$$

Therefore,

$$P(j' \in \mathcal{A}_n) \to 0.$$

The proof is finished.

Now we consider the consistency part. $\forall j \in \mathcal{A}$, the asymptotic normality result indicates that $\hat{\beta}_j^{\text{CAL}} \to_p \beta_j^*$; thus $P(j \in \mathcal{A}_n) \to 1$. Then it suffices to show that $\forall j' \notin \mathcal{A}$, $P(j' \in \mathcal{A}_n) \to 0$. Consider the event $j' \in \mathcal{A}_n$. By the KKT optimalirt conditions, we know that $2\mathbf{x}_{j'}^T(\mathbf{y}-\mathbf{X}_{\mathcal{C}}\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}-\mathbf{X}_{\mathcal{D}}\hat{\boldsymbol{\beta}}^{\text{CAL}}) = \lambda_n\hat{w}_j$. Note that $\lambda_n\hat{w}_j/\sqrt{n} = \frac{\lambda_n}{\sqrt{n}}n^{\gamma/2}\frac{1}{\sqrt{n}\hat{\beta}_{j'}} \to_p \infty$, whereas

$$2\frac{\mathbf{x}_{j'}^T(\mathbf{y}-\mathbf{X}_{\mathcal{C}}\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}-\mathbf{X}_{\mathcal{D}}\hat{\boldsymbol{\beta}}^{\text{CAL}})}{\sqrt{n}} = 2\frac{\mathbf{x}_{j'}^T\boldsymbol{\varepsilon}}{\sqrt{n}}+2\frac{\mathbf{x}_{j'}^T\mathbf{X}_{\mathcal{C}}}{\sqrt{n}}\sqrt{n}(\boldsymbol{\beta}_{\mathcal{C}}^*-\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)})+2\frac{\mathbf{x}_{j'}^T\mathbf{X}_{\mathcal{D}}}{\sqrt{n}}\sqrt{n}(\boldsymbol{\beta}_{\mathcal{D}}^*-\hat{\boldsymbol{\beta}}^{CAL}).$$

And we know that $2\mathbf{x}_{j'}^T\mathbf{X}_{\mathcal{C}}\sqrt{n}(\boldsymbol{\beta}_{\mathcal{C}}^*-\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)})/\sqrt{n}+2\mathbf{x}_{j'}^T\mathbf{X}_{\mathcal{D}}\sqrt{n}(\boldsymbol{\beta}_{\mathcal{D}}^*-\hat{\boldsymbol{\beta}}^{CAL})/\sqrt{n} \to_d$ some normal distribution and $2\mathbf{x}_{j'}^T\boldsymbol{\varepsilon}/\sqrt{n} \to_d \text{N}(\mathbf{0}, 4\|\mathbf{x}_{j'}^T\|^2\sigma^2)$. Thus

$$P(j' \in \mathcal{A}_n) \leqslant P(2\mathbf{x}_{j'}^T(\mathbf{y}-\mathbf{X}_{\mathcal{C}}\hat{\boldsymbol{\beta}}_{\mathcal{C}}^{(n)}-\mathbf{X}_{\mathcal{D}}\hat{\boldsymbol{\beta}}^{\text{CAL}}) = \lambda_n\hat{w}_j) \to 0.$$

Proof of Theorem 2

Since $\mathbf{X}_{\mathcal{C}}$ part is not penalized, it's easily shown that it's a MLE and has the asymptotic behavior. Then the proof is the same as in Theorem 1.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Symposium on Information Theory*, pages 267–281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Barut, E., Fan, J., and Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association*, 111(515):1266–1277.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232.

Breiman, L. et al. (1996). Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6):2350–2383.

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.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):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. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484.

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10(Sep):2013–2038.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., Simon, N., and Tibshirani, R. (2016). Lasso and elastic-net regularized generalized linear models. r-package version 2.0-5. 2016.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

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

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Mallows, C. L. (1973). Some comments on c p. *Technometrics*, 15(4):661–675.

McCullagh, P. and Nelder, J. A. (1989). Generalized linear models, no. 37 in monograph on statistics and applied probability.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics*, pages 1012–1030.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Shi, P. and Tsai, C.-L. (2002). Regression model selection—a residual likelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):237–252.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494.

Xie, J. and Yin, X. (2018). The conditional adaptive lasso and its sufficient variable selection algorithm.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

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 (Statistical Methodology)*, 67(2):301–320.

Vita

Jin Xie

Education

- **M.S. in Statistics** University of Kentucky, Lexington, KY, 2012-2014
- **B.S. in Statistics** Nanjing University, Nanjing, China, 2008-2012

Experience

- **Research Assistant** University of Kentucky, 2015-2017
- **Teaching Assistant** University of Kentucky, 2013-2015