ANALYSIS OF BIG DATA BY SPLIT-AND-CONQUER AND PENALIZED REGRESSIONS: NEW METHODS AND THEORIES

BY XUEYING CHEN

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Statistics and Biostatistics

Written under the direction of

Minge Xie and Cun-Hui Zhang

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2013

ABSTRACT OF THE DISSERTATION

Analysis of Big Data by Split-and-Conquer and Penalized Regressions: New Methods and Theories

by Xueying Chen

Dissertation Director: Minge Xie and Cun-Hui Zhang

This dissertation develops methodologies for analysis of big data and its related theoretical properties. Recent years, tremendous progress has been made in analysis of big data, especially techniques via penalization and shrinkages. However, there are still many challenging problems to be solved. This dissertation focuses on two settings where (i) the data is too large to fit into a single computer or too expensive to perform a computationally intensive data analysis; or (ii) there are unknown group structures of highly correlated variables. In this dissertation, we first propose a *Split-and-Conquer* approach to analyze extraordinarily large data. Then, under linear regression settings with highly correlated variables, we investigate model selection properties of OSCAR (octagonal shrinkage and clustering algorithm for regression) estimators (Bondell & Reich, 2008) and propose a more general method *Group OSCAR* which incorporates both prior knowledge of group structures and correlation patterns among explanatory variables.

We first propose a split-and-conquer approach and illustrate it using a computationally intensive penalized regression method. We show that the combined result is asymptotically equivalent to the corresponding analysis result of using the entire data all together. In addition, we demonstrate that the approach has an inherent advantage of being more resistant to false model selections. Furthermore, when a computational intensive algorithm is used, we show that the split-and-conquer approach can substantially reduce computing time and computer memory requirement.

Detecting meaningful 'groups' of highly correlated variables has been studied a lot. OSCAR estimators provide a feasible way to perform variable selection and clustering simultaneously. However, no theoretical results are provided for OSCAR estimators. In this dissertation, we provide a set of mild conditions under which OSCAR estimators are able to select the true model and keep the order of the coefficients by their magnitudes when the correlations are high.

In the last part of this dissertation, we propose a new method. This method not only takes use of known group structures but also incorporates the correlation patterns leading to the underlying unknown group structure. It extends most of the model selections methods in the literature, and has a general grouping effect.

Acknowledgements

I would like to express my gratitude to my advisor Professor Minge Xie for his constant support and advice throughout these years. I thank him for providing me lots of opportunities and guiding me throughout the research projects. His motivation and confidence in me always encouraged me to overcome challenges in the dissertation. Without his patience and continuous support, this dissertation would be impossible.

My thanks also go to Professor Cun-Hui Zhang who also provided me guidance and had in depth discussions in my research. His deep understanding in penalized regression models offered me is invaluable for me to complete my research.

I am also grateful to Professor Lee Dicker and Professor Xiaodong Lin for their precious time and efforts to serve on my thesis committee. I would thank Dr. Jerry Cheng who has kindly cleaned and gave me access to the dataset used in the research projects.

Last but definitely not the least, I would like thank all my friends in Department of Statistics, especially Wentao Li, Guang Yang and Wenqian Qiao who have been supportive and helped me a lot in every aspect. I deeply appreciate their friendship that made this experience enjoyable.

Dedication

To my family

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Consider a generalized linear model:

$$E(y_i) = g(\boldsymbol{x}_i'\boldsymbol{\beta}), i = 1, \ldots, n$$

where y_i is a response variable and $\boldsymbol{x}_i$ is a $p \times 1$ explanatory vector, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and g is a link function. Both the sample size n and the number of parameters p can be potentially very large. We assume that, given $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, the conditional distribution of $\boldsymbol{y} = (y_1, \ldots, y_n)'$ follows the canonical exponential distribution:

$$f(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_0(y_i; \theta_i) = \prod_{i=1}^{n} \left\{ c(y_i) \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi} \right] \right\}, \tag{1.1}$$

where $\theta_i = \boldsymbol{x}_i'\boldsymbol{\beta}, i = 1, \ldots, n$. The log-likelihood function $\log f(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta})$ is then given by

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}) = [\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} - \mathbf{1}'\boldsymbol{b}(\boldsymbol{X}\boldsymbol{\beta})]/n, \tag{1.2}$$

where $\boldsymbol{b}(\boldsymbol{\theta}) = (b(\theta_1), \ldots, b(\theta_n))'$ for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$. In the case when p is large (or grows with n) and $\boldsymbol{\beta}$ is sparse (i.e., many elements of $\boldsymbol{\beta}$ are zero), a penalized likelihood estimator is often used, which is defined as, in a general form,

$$\hat{\boldsymbol{\beta}}^{(a)} = \text{argmax}_{\boldsymbol{\beta}} \left\{ \ell(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X})/n - \rho(\boldsymbol{\beta}; \lambda_a) \right\}. \tag{1.3}$$

Here, $\boldsymbol{y}$ is a $n \times 1$ response vector, $\boldsymbol{X}$ is a $n \times p$ matrix; ρ is the penalty function with tuning parameter λ_a. The superscript a refers to the result obtained by analyzing *all* data simultaneously. Depending on the choice of penalty function $\rho(\boldsymbol{\beta}; \lambda_a)$, we have bridge regression (Frank & Friedman, 1993), LASSO estimator (Tibshirani, 1996; Chen et al., 2001), LARS algorithm (Efron et al., 2004), SCAD estimator (Fan & Li, 2001) and MCP estimators (Zhang, 2010), among others. In this dissertation, we focus on a setting used in the review article of Fan & Lv (2011) which covers most commonly used penalty functions currently used in practice. Under the setting, Fan & Lv (2011) show that the penalized estimators under the generalized linear models (1.3) have good asymptotic properties, such as model selection consistency and asymptotic normality etc., under some regularity conditions.

Although penalized regression has been a successful method to perform variable selection when p is large, computational intensive algorithms restrain its application, especially when massive data is available. To solve computing problems, we propose a split-and-conquer approach for the situation that n is extraordinarily large, too large to perform the aforementioned penalized regression using a single computer or available computing resources to us. In this case, without touching the existing penalized regression methods, we split the whole dataset into K subsets of smaller sample sizes. Each subset is then analyzed separately, provided that such an analysis can be performed on the smaller subsets. A set of K results are obtained. Subsequently, the K results are combined to obtain a final result. We prove that, under some mild conditions and with a suitable choice of K, our combined estimator using the split-and-conquer approach is asymptotically equivalent to the penalized estimator obtained from analyzing entire data all together. The combined estimator can keep the sparsity property

and is model selection consistent as long as the penalized estimators from the imposed penalty function are sparse and model selection consistent. When asymptotic normality is attainable, the combined estimator does not lose any efficiency through the split-and-conquer process, in the sense that it has the same asymptotic variance as the penalized estimator using entire data all together. In other words, although the combined estimator may not be exactly the same as the one using the entire data all together, it is as asymptotically efficient and asymptotically equivalent as the penalized estimator analyzing the entire data all together. In addition, the split-and-conquer approach involves random splitting in the first step. Taking advantage of this procedure, we further provides an upper bound for the number of false selected variables and lower bound for the truly selected variables. In fact, split-and-conquer is a very general approach that can be applied for any statistical analysis method. As long as a computational intensive algorithm with computing expenses at the order of $O(n^a)$, $a > 1$, is used, we show using a simple calculation, as well as demonstrate using numerical examples, that the split-and-conquer approach can reduce computing time and computer memory requirement. The details of the proposed method are presented in Chapter 2.

In Chapter 3, we restrain models to be simple linear regression model

$$\boldsymbol{y} = \sum_{i=1}^{p} \boldsymbol{x}_i \beta_i + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} = (y_1, ..., y_n)^T$ is the response vector of n observations, $\boldsymbol{x}_1 = (x_{11}, ..., x_{1n})^T$ ... $\boldsymbol{x}_p = (x_{p1}, \ldots, x_{pn})^T$ are the vectors of p explanatory variables, $\beta_1, \ldots, \beta_p$ are the corresponding regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is the vector of independent random errors. We assume that the $\boldsymbol{x}$'s are standardized so that $\sum_{j=1}^{n} x_{ij} = 0$ and $\sum_{j=1}^{n} x_{ij}^2 = n$, $i = 1, \ldots, p$. The aforementioned methods and also many other papers published about penalized regression, however, are designed to select individual

variables with low correlations. They are not particularly effective to incorporate the correlation structures among variables. However, detecting meaningful 'groups' of predictors are important to prediction accuracy and model interpretation (Zou & Hastie, 2005; Bondell & Reich, 2008). Here, highly correlated variables should be grouped in the sense that they are jointly included or excluded and have similar or exactly the same coefficients. Compared with other estimators, OSCAR estimators are more desirable since they achieve an exact grouping property, that is, the absolute coefficients of two highly correlated variables are enforced to be exactly the same. While the grouping effect has been found to be desirable, a fundamental question of OSCAR estimators is whether they can select the true predictors that have nonzero coefficients. If yes, what conditions are needed for OSCAR estimators to be model selection consistent? We find that, unlike other penalized estimators which obtain grouping property by adding a quadratic or L_2 norm-like penalty on coefficients' differences, OSCAR penalty function can be rewritten as a L_1 norm penalty on coefficients plus a L_1 norm penalty on coefficients' differences. The nondifferentiation of L_1 norm function distinguishes OSCAR estimators from others and leads to the technique difficulties. In this dissertation, we consider a more restrictive definition of sign consistency which requires estimators to keep the magnitude order of the coefficients beside selecting the true model. We provide a set of mild conditions under which OSCAR estimators are sign consistent. The details are presented in Chapter 3.

In practice, group structures can be retrieved from different sources. On one hand, as mentioned above, the data itself contain group structures which are not available beforehand. For instance, identical or highly correlated variables may need to be grouped together because of their similarity. In other words, the underlying group information

is implied in the correlation patterns among explanatory variables. On the other hand, prior knowledge can provide information about group structures. For example, dummy variables created to represent different factors of one categorical variable are considered as one group or the experts who collect the data may suggest to split the explanatory variables into several groups. The existing approaches either only take correlation patterns into consideration (Zou & Hastie, 2005; Bondell & Reich, 2008), or consider the scenario that group structures are completely given (Yuan & Lin, 2006; Zhao & Yu, 2006). We propose Group OSCAR that is able to capture group features in the data from both sources for variable selection. The proposed penalty consists of two regularity functions. In particular, a representative variable, which is a weighted average of explanatory variable in a certain group, is created for each known group. The weights will be automatically determined by the algorithm according to corresponding coefficients' magnitudes. Then, if two representative variables are highly correlated, the corresponding two groups will be merged. In addition, model selection consistency property is explored. The proposed method is presented in Chapter 4.

The rest of this thesis is organized as follows. In Chapter 2, we develop a general split-and-conquer approach for extraordinarily large data analysis problem and demonstrate it using computationally intensive penalized regression. In Chapter 3, we explore model selection properties of OSCAR estimators. In Chapter 4, we propose Group OSCAR that is able to incorporate both prior group information and correlation structures among explanatory variables.

Chapter 2

A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data

2.1 Introduction

In this chapter, we consider the situation that the data size is extraordinarily large, too large to fit into a single computer or be analyzed with available computing resources. We propose a split-and-conquer approach to solve the problem and illustrate it using the aforementioned penalized regression methods. Specifically, we split the whole dataset into K subsets of smaller sample sizes. Each subset is then analyzed separately, provided that such an analysis can be performed on the smaller subsets. A set of K results are obtained. Subsequently, the K results are combined to obtain a final result. Our task is to investigate whether the combined overall result can be the same or as good as the result that is obtained from analyzing the entire dataset and, if conditions are needed, what they are. Although the split-and conquer approach can apply more generally, to facilitate our discussion, we focus on a penalized regression setting considered in the review article of Fan & Lv (2011), which covers most commonly used penalty functions currently used in penalized regression practice such as LASSO, SCAD, MCP and others. Under the setting, Fan & Lv (2011) show that the penalized estimators under the generalized linear models (1.3) have good asymptotic properties, such as model selection consistency and asymptotic normality, etc. We investigate in this paper

specifically whether the combined result from the proposed split-and-conquer method using the corresponding penalized regression still retains these desired properties and, if so, under which conditions.

The idea behind the proposed split-and-conquer approach is simple and straightforward. Its essence can be illustrated using a simple special case of the regular Gaussian linear regression where we have finite p and non-sparse $\boldsymbol{\beta}$. In particular, the ordinary least squares estimator using entire data all at once in this case is

$$\hat{\boldsymbol{\beta}}^{(a)} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

When we split the dataset into K pieces, the ordinary least squares estimator obtained from the k^{th} subset is $\hat{\boldsymbol{\beta}}_k = (\boldsymbol{X}_k'\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k'\boldsymbol{y}_k$, where $\boldsymbol{X}_k$ is the design matrix and $\boldsymbol{y}_k$ is the response vector for data in the k^{th} subset. These K least square estimators can be combined, using the inverse of $\hat{\boldsymbol{\beta}}_k$'s variance $S_k \overset{\mathrm{d}}{=} \boldsymbol{X}_k'\boldsymbol{X}_k$ as their combining weights, to form a new estimator

$$\hat{\boldsymbol{\beta}}^{(c)} = (\sum_{k=1}^{K}\boldsymbol{X}_k'\boldsymbol{X}_k)^{-1}\sum_{k=1}^{K}(\boldsymbol{X}_k'\boldsymbol{X}_k)\hat{\boldsymbol{\beta}}_k = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sum_{k=1}^{K}\boldsymbol{X}_k'\boldsymbol{y}_k = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

This combined new estimator $\hat{\boldsymbol{\beta}}^{(c)}$ is identical to $\hat{\boldsymbol{\beta}}^{(a)}$. Thus, we do not lose any information through the split-and-conquer approach in the case of the regular Gaussian linear regression. For penalized estimators and under generalized linear models, the results are not so straightforward. Our specific aim in this paper is to investigate whether we have any similar results to support the split-and-conquer approach under generalized linear models and for penalized estimators. We also investigate whether there are any special properties and benefits for more complex settings beyond this simple case involving a small fixed p and the least squares estimation.

We prove that, under some mild conditions and with a suitable choice of K, our

combined estimator using the split-and-conquer approach is asymptotically equivalent to the penalized estimator obtained from analyzing entire data all at once. Here, the number of splitting K should be relatively large so that each subset is small enough and can by analyzed using computing resources available to us. But it should not be too large either, because each subset should contain enough data to provide a meaningful estimator for the unknown regression parameter $\boldsymbol{\beta}$. The combined estimator can keep the sparsity property and is model selection consistent as long as the penalized estimators from the imposed penalty function have the properties of sparsity and model selection consistency. When asymptotic normality is attainable, the combined estimator does not lose any efficiency through the split-and-conquer process, in the sense that it has the same asymptotic variance as the penalized estimator using entire data all at once. In other words, although the combined estimator may not be exactly the same as the one using the entire data all at once, it is as asymptotically efficient and asymptotically equivalent as the penalized estimator analyzing the entire data all together. There is no price to pay under a regular Gaussian linear regression. But we show that under generalized linear models and with more complicated settings, it requires stronger conditions such as larger coefficients signals or slower growth rate of p to retain the aforementioned desired properties for the combined estimator from a split-and-conquer approach.

The split-and-conquer approach involves combining the results of subsets that are obtained from random splitting. Utilizing this procedure, improvements over the regular penalized estimators in model selection can be expected through a majority voting in the combining step. As a result, we are able to establish an upper bound for the expected number of falsely selected variables and a lower bound for the expected number of

truly selected variables. Several papers in the literature have found that averaging over independent observations can reduce the impact of random errors. For instance, Fan et al. (2010) propose refitted cross-validation to attenuate false correlations among the random errors and explanatory variables that they call spurious variables. Meinshausen & Buhlmann (2010) introduce stability selection which is a combination of subsampling and model selection algorithms. They get an exact error control bound because the data from subsampling are independent. Shah & Samworth (2012) propose a variant of stability selection with improved error control property. Similarly, the split-and-conquer approach provides resistance to selection errors caused by spurious correlations and keeps a large amount of variables that are in the true model at the same time. Usually, this control on the selected variables are not available for conventional penalized estimators.

Furthermore, when a computational intensive algorithm with computing expenses at the order of $O(n^a)$, $a > 1$, is used, we demonstrate that the split-and-conquer approach can reduce computing time and computer memory requirement. For instance, consider the example of linear regression with L_1 norm penalty function. The LARS (Efron et al., 2004) algorithm, which has been considered by some researchers as a fast and efficient algorithm to solve the LASSO problem, requires $O(n^3)$ computations when $p \geq n$. The computing time can be costly when both n and p are extraordinarily large. In this case, we show both mathematically and numerically the proposed split-and-conquer approach with LARS can save up to $(1 - 1/K^2)\%$ computing time, where K is the number of splitting. Indeed, our numerical studies in Section 4 provide several examples across different models and penalized methods in which the proposed split-and-conquer approach provides substantial savings in computing time while producing

comparable estimators. In some practice, the split-and-conquer approach provides a, if not the only, feasible way to carry out such analysis.

The split and conquer approach is intuitive and similar ideas have also been explored by the others. For instance, Mackey et al. (2011) propose a divide-and-conquer method for matrix factorization, in which the authors partition a large-scale matrix into submatrix, factor each submatrix and then combine submatrix estimates. A similar practice can also be found in the computer sciences community under the name of parallel and distributed computing (see, e.g., Andrews (2000)). However, the research there focuses mainly on computer sciences aspects, such as accessing to a shared memory, exchanging information between processors, etc. There is no systemic and theoretical study from statistical prospects, especially on the combination methodology and the statistical performance of the overall result from a statistical analysis.

The rest of this chapter is organized as follows. Section 2.2 proposes a split-and-conquer approach and a combined estimator under the generalized linear regression models. Section 2.3 studies theoretical properties of the combined estimator and also investigate issues related to error bound controls and computing time. Section 2.4 illustrates the results using simulations and real data from an application of cargo screening in the U.S. Port-of-Entries (POEs) practices. Section 2.5 provides further discussions. Section 2.6 gives mathematical proofs of theorems and lemmas.

2.2 Split-and-conquer for penalized regressions

Suppose $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters that lies in the parameter space Ω and the true parameter, denoted by $\boldsymbol{\beta}^0$, is sparse. Let us divide the entire dataset of size n into K subsets and the k^{th} subset has n_k observations: $(\boldsymbol{x}_{k,i}, y_{k,i})$, $i = 1, \ldots, n_k$. For the

k^{th} subset, the log-likelihood function is

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}_k, \boldsymbol{X}_k) = [\boldsymbol{y}_k' \boldsymbol{X}_k \boldsymbol{\beta} - \mathbf{1}' \boldsymbol{b}(\boldsymbol{X}_k \boldsymbol{\beta})]/n_k, \quad k = 1, \ldots, K,$$

where $\boldsymbol{y}_k = (y_{k,1}, \ldots, y_{k,n_k})'$ is a $n_k \times 1$ response vector and $\boldsymbol{X}_k = (\boldsymbol{x}_{k,1}', \ldots, \boldsymbol{x}_{k,n_k}')'$ is a $n_k \times p$ matrix. Corresponding to (1.3), the penalized estimator for the k^{th} subset is:

$$\hat{\boldsymbol{\beta}}_k = \text{argmax}_{\boldsymbol{\beta}} \left\{ \ell(\boldsymbol{\beta}; \boldsymbol{y}_k, \boldsymbol{X}_k)/n_k - \rho(\boldsymbol{\beta}; \lambda_k) \right\},$$

where $\rho(\boldsymbol{\beta}; \lambda_k)$ is the penalty function with tuning parameter λ_k. To simplify our discussion and following Fan & Lv (2011), we write $\rho(\boldsymbol{\beta}; \lambda_k) = \sum_{j=1}^{p} \rho(\beta_j; \lambda_k)$ and assume that the penalty function $\rho(\beta_j; \lambda_k)$ satisfy the following condition:

- **(PC)** *Assume $\rho(t; \lambda)$ is increasing and concave in $t \in [0, \infty)$, and has a continuous derivative $\rho'(t; \lambda)$ with $\rho'(0+; \lambda) > 0$. In addition, $\rho'(0+; \lambda)$ is increasing in $\lambda \in [0, \infty)$ and $\rho'(0+; \lambda)$ is independent of λ.*

The class of penalty functions satisfying Condition (PC) covers most commonly used penalty functions, including L_1 penalty, SCAD, MCP, among others.

From Fan & Lv (2011), the penalized estimator $\hat{\boldsymbol{\beta}}_k$ has the so-called sparsity property with many zero entries. Let us denote by $\hat{\mathcal{A}}_k = \{j : \hat{\beta}_{k,j} \neq 0\}$ the set of selected variables (non-zero elements) of $\hat{\boldsymbol{\beta}}_k$. Also, for any indices set S, denote by $\hat{\boldsymbol{\beta}}_{k,S}$ a $|S| \times 1$ vector that is formed by the elements of $\hat{\boldsymbol{\beta}}_k$ whose indices are in S. Thus, $\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}_k}$ is the sub-vector that contains only the non-zero elements of $\hat{\boldsymbol{\beta}}_k$. Note that, since each $\hat{\boldsymbol{\beta}}_k$ is estimated from a different subset of data, $\hat{\mathcal{A}}_k$ can be different from one to another and the K vectors $\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}_k}$, $k = 1, \ldots, K$, may have different lengths.

In order to obtain a combined estimator of $\boldsymbol{\beta}$ from $\hat{\boldsymbol{\beta}}_k$'s that retains good performance, we use a majority voting method. There are two considerations. First, the

combined estimator should be formed based on the estimators from the subsets $\hat{\boldsymbol{\beta}}_k$'s. A variable that is not selected in any of $\hat{\mathcal{A}}_k = \{j : \hat{\beta}_{k,j} \neq 0\}$ should also be excluded by the combined estimator. On the other hand, $\hat{\mathcal{A}}_k$ are subject to selection errors because only a portion of data is analyzed and the penalized likelihood estimator does not always guarantee the perfect selection. There is always some mismatch between the set $\hat{\mathcal{A}}_k$ from the analysis of the k^{th} subset and the true nonzero set, say $\mathcal{A} \overset{\mathrm{d}}{=} \{j : \beta_j^0 \neq 0\}$. We apply a majority voting method to handle these issues. In our majority voting method, we define

$$\hat{\mathcal{A}}^{(c)} \overset{\mathrm{d}}{=} \left\{ j : \sum_{k=1}^{K} \mathbf{I}(\hat{\beta}_{k,j} \neq 0) > w \right\}$$

as the set of selected variables of the combined estimator, where $w \in [0, K)$ is a prespecified threshold and $\mathbf{I}$ is the indicator function. We always have $\hat{\mathcal{A}}^{(c)} \subset \bigcup_{k=1}^{K} \hat{\mathcal{A}}_k$. When the numbers of elements in $\hat{\mathcal{A}}_k$, denoted by $|\hat{\mathcal{A}}_k|$, for $k = 1, \ldots, K$, are small and the sets $\hat{\mathcal{A}}_k$ have lots of common elements, the numbers of elements in $\hat{\mathcal{A}}^{(c)}$, denoted by $|\hat{\mathcal{A}}^{(c)}|$, can be much smaller than p. In one extreme casein which the threshold $w \geq K - 1$, the majority voting set $\hat{\mathcal{A}}^{(c)}$ contains only those variables that are selected by all penalized estimators from the subsets. In the other extreme case in which $w = 0$, $\hat{\mathcal{A}}^{(c)}$ contains those variables that are selected by at least one penalized estimator from the subsets.

The majority voting idea is closely connected with the novel developments by Meinshausen & Buhlmann (2010) and Shah & Samworth (2012) on stability selection. For example, we may view the quantity $\sum_{k=1}^{K} \mathbf{I}(\hat{\beta}_{k,j} \neq 0)/K$ as a variant version of $\hat{\Pi}_j^{\lambda}$, the probability of variable j to be selected with tuning parameter λ, used in Meinshausen & Buhlmann (2010). However, the goal of Meinshausen & Buhlmann (2010) and Shah & Samworth (2012) is to develop stable penalized estimators, which is different from ours. Although our development also cares about performance and stable estimation,

the main focus is to investigate whether we can analyze extreme large data by splitting the task and thus computational feasibility is the forefront of our development. Different from Meinshausen & Buhlmann (2010) and Shah & Samworth (2012) (which will be computationally infeasible for extraordinarily large data due to multiple rounds of calculation and subsampling), the proposed majority voting approach only requires one-round calculation and each subset has much less observations especially when K is large. These difference can help improve computing efficiency and time which in turn increase the feasibility of handling extraordinarily large data. Also, instead of using the same tuning parameter λ for all subsets, our $\hat{\boldsymbol{\beta}}_k$ from each subset is calculated from subset k with tuning parameter λ_k chosen by a criterion, e.g. AIC, BIC or cross-validation, independently within each subset. The K tuning parameters are often not the same. Finally, our development applies to the case when $K \longrightarrow \infty$. It subsumes the situation discussed in Meinshausen & Buhlmann (2010) or Shah & Samworth (2012) in which $K = 2$ or finite.

We introduce the following notations. For any $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, define

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = (\mu(\theta_1), \ldots, \mu(\theta_n))' \text{ and } \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \text{diag}(\sigma(\theta_1), \ldots, \sigma(\theta_n)),$$

where $\mu(\theta) = \partial b(\theta)/\partial\theta$ and $\sigma(\theta) = \partial^2 b(\theta)/\partial^2\theta$. We also define weight matrices

$$\hat{\boldsymbol{S}}_k \stackrel{\text{d}}{=} \boldsymbol{X}_k' \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k) \boldsymbol{X}_k, \tag{2.1}$$

where $\hat{\boldsymbol{\theta}}_k = \boldsymbol{X}_k \hat{\boldsymbol{\beta}}_k$. The weight matrix $\boldsymbol{S}_k$ comes from the second order condition of the penalized likelihood function. It is approximately the inverse of the covariance matrix $\boldsymbol{S}_k^0 = \boldsymbol{X}_k' \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0) \boldsymbol{X}_k$ with $\boldsymbol{\theta}_k^0 = \boldsymbol{X}_k \boldsymbol{\beta}^0$, where $\boldsymbol{\beta}^0$ is the true coefficients.

We propose to use the following combined estimator, which is the weighted average of $\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}^{(c)}}$, $k = 1, \ldots, K$:

$$\hat{\boldsymbol{\beta}}^{(c)} \stackrel{\mathrm{d}}{=} \boldsymbol{A}\left(\sum_{k=1}^{K} \boldsymbol{A}'\boldsymbol{S}_k\boldsymbol{A}\right)^{-1} \sum_{k=1}^{K} \boldsymbol{A}'\boldsymbol{S}_k\boldsymbol{A}\hat{\boldsymbol{\beta}}_{k,\hat{\mathcal{A}}^{(c)}}, \tag{2.2}$$

where $\boldsymbol{E} = \mathrm{diag}(\mathbf{I}_{\sum_{k=1}^{K}\mathbf{I}(\hat{\beta}_{k,j}\neq 0)>w})$ and $\boldsymbol{A} = \boldsymbol{E}_{\hat{\mathcal{A}}^{(c)}}$. Here, for any set S, $\boldsymbol{E}_S$ stands for an $p \times |S|$ submatrix of $\boldsymbol{E}$ formed by columns whose indices are in S.

2.3 Theoretical results

In this section, we investigate the asymptotic properties of the combined estimator $\hat{\boldsymbol{\beta}}^{(c)}$ defined in (2.2), and compare it with the penalized estimator $\hat{\boldsymbol{\beta}}^{(a)}$ that is obtained from analyzing the entire dataset all at once as defined in (1.3).

2.3.1 Sign consistency

We first show that the combined estimator is sign consistent, i.e. each component of the combined estimator has the same sign as that of its true value, given that the penalized estimator obtained from each subset is consistent.

Denote by $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_p^0)$ the true parameter. Also, denote by $\mathcal{A} = \{i : \beta_i^0 \neq 0\}$ the true nonzero set and $\mathcal{B} \stackrel{\mathrm{d}}{=} \mathcal{A}^c$ its complement or the set of noise variables. We write the minimal signal as $\beta_* = \min\{|\beta_j^0| : \beta_j^0 \neq 0\}$. For any indices set S, $\boldsymbol{X}_S$ stands for an $n \times |S|$ submatrix of $\boldsymbol{X}$ formed by columns with indices in S. Similarly, $\boldsymbol{X}_{k,S}$ stands for an $n_k \times |S|$ submatrix of $\boldsymbol{X}_k$ formed by columns with indices in S.

In order to obtain model selection consistency of the combined estimator, we require certain regularity conditions on the design matrix. Assumption **A1** basically requires that the norms of the design matrices are proportional to the sample sizes in the subsets and the entire data. These conditions are mild and often satisfied in practice. More

specifically, we assume

$$\textbf{A1}\quad \begin{aligned}
&\|[\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\beta}^0_{\mathcal{A}})\boldsymbol{X}_{k,\mathcal{A}}]^{-1}\|_\infty = O(b_{s,K}n_k^{-1}),\\
&\|[\boldsymbol{X}'_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{X}_{\mathcal{A}}\boldsymbol{\beta}^0_{\mathcal{A}})\boldsymbol{X}_{\mathcal{A}}]^{-1}\|_\infty = O(b_{s,K}n^{-1}),\\
&\|\boldsymbol{X}'_{k,\mathcal{A}^c}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}[\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}]^{-1}\|_\infty \le \min\{C\rho'(0+)/\rho'(\beta_*/2;\lambda_k), O(n_k^\alpha)\},\\
&\max_{\boldsymbol{\delta}\in\mathcal{N}_0}\max_{j=1,\ldots,p}\lambda_{\max}[\boldsymbol{X}'_{k,\mathcal{A}}\mathrm{diag}\{|\boldsymbol{x}_{k,j}|\circ|\boldsymbol{\mu}''(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})|\}\boldsymbol{X}_{k,\mathcal{A}}] = O(n_k),
\end{aligned}$$

where $\{b_{s,K}\}$ is a diverging sequence of positive numbers that depends on s and K; $C\in(0,1)$, $\alpha\in[0,1/2]$ and $\mathcal{N}_0=\{\boldsymbol{\delta}\in\Re^{s_n}:\|\boldsymbol{\delta}-\boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty\le\beta_*/2\}$. Here, the derivative is taken componentwise and $\circ$ is componentwise product.

Since the weight matrices $\hat{\boldsymbol{S}}_k=\boldsymbol{X}'_k\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_k$ use $\hat{\boldsymbol{\theta}}_k=\boldsymbol{X}_k\hat{\boldsymbol{\beta}}$ rather than the true $\boldsymbol{\theta}^0_k$, we further assume the following conditions to control the bias caused by an estimation of the weight matrix.

$$\textbf{A2}\quad \begin{aligned}
&b(\theta)\ \text{ has the third derivative }\ b'''(\theta),\\
&\|\boldsymbol{X}^T_{k,\mathcal{A}}D_k\boldsymbol{X}_{k,\mathcal{A}}\{\boldsymbol{X}^T_{k,\mathcal{A}}\boldsymbol{\Sigma}(\theta^0_k)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\|_\infty = o(1),\\
&\|I+\{\boldsymbol{X}^T_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)\boldsymbol{X}_{\mathcal{A}}\}^{-1}\textstyle\sum_{k=1}^K\boldsymbol{X}^T_{k,\mathcal{A}}D_k\boldsymbol{X}_{k,\mathcal{A}}\|_\infty = O(1),
\end{aligned}$$

where $D_k=\mathrm{diag}(d_{kj})$, $d_{kj}=b'''(\boldsymbol{x}_j^T\boldsymbol{\delta})\boldsymbol{x}_j^T(\boldsymbol{\delta}-\boldsymbol{\beta}^0_{\mathcal{A}})$ with $\boldsymbol{\delta}\in\mathcal{N}_0$ and $\boldsymbol{x}_j$ in subset k, and $D=\mathrm{diag}(D_k)$.

Let $v_{n,K}$ and $u_{n,K}$ be two diverging sequences depending on the total sample size n and the number of subsets K. Assume the following conditions

$$\textbf{A3}\quad \begin{aligned}
&b_{s,K}v_{n,K}/(n\beta_*)=o(1), && b_{s,K}\rho'(\beta_*/2;\lambda_k)/\beta_*=o(1),\\
&n_k\beta_*s=o(1), && n^\alpha s\beta_*^2/(K^\alpha\lambda_k)=o(1),\\
&u_{n,K}/(n\lambda_k)=o(1), && v_{n,K}n^{\alpha-1}/(K^\alpha\lambda_k)=o(1),\\
&\lambda_k\kappa_0=o(\tau_0),
\end{aligned}$$

where $\kappa_0=\max_{\boldsymbol{\delta}\in\mathcal{N}_0}\kappa(\rho;\boldsymbol{\delta})$ and $\tau_0=\min_{\boldsymbol{\delta}\in\mathcal{N}_0}\lambda_{\min}[n_k^{-1}\boldsymbol{X}^T_{k,\mathcal{A}}\Sigma(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{k,\mathcal{A}}]$.

In Condition **A3**, $v_{n,k}$ and $u_{n,k}$ are related to the error tolerance level. Specifically, the probability of obtaining the correct signs of nonzero variables will increase with $v_{n,k}$; the probability of excluding variables with zero coefficients will increase with $u_{n,k}$. Therefore, smaller error tolerance probability will lead to larger penalty level (larger λ_k) and thus requires larger signal strength β_*.

Compared with the conditions in Fan & Lv (2011), Condition **A3** explains the inherent connections between the minimal signal β_*, model size s, the number of parameters p and the sample size n rather than just specifies a set of possible choices. Consider the signal strength used in Fan & Lv (2011), $\beta_* = O(n^{-\gamma}\log n)$, $\gamma \in (0, 1/2]$. In this case, we get the following conditions which are consistent with the conditions in Fan & Lv (2011):

A3′

$$\begin{aligned}
&\beta_* = O(n^{-\gamma}\log n), \gamma \in (0,1/2]; && s = O(n^{\alpha_0}), \alpha_0 \in (0,1),\\
&v_{n,K} = \sqrt{Kn\log n}; && u_{n,k} = K^{1/2}n^{1/2-\alpha_1}(\log n)^{1/2},\\
&b_{s,K} = o(\min\{K^{-1/2}n^{1/2-\gamma}\sqrt{\log n}, s^{-1}n^{\gamma}/\log n\}); && v_{n,K}n^{\alpha-1}/(K^{\alpha}\lambda_k) = o(1);\\
&\rho'(\beta_*/2;\lambda_k) = o(b_{s,K}^{-1}n^{-\gamma}\log n); && \lambda_k \gg n^{-\alpha_1}(\log n)^2/K^{\alpha};\\
&K = o\{\min(n^{1-2\gamma}\log n, n^{1-\alpha_0})\}; && \lambda_k\kappa_0 = o(\tau_0),
\end{aligned}$$

where $\alpha_1 = \min(1/2, 2\gamma - \alpha_0) - \alpha$ $\tau_0 = \min_{\boldsymbol{\delta}\in\mathcal{N}_0}\lambda_{\min}[n_k^{-1}\boldsymbol{X}_{k,\mathcal{A}}^T\Sigma(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})\boldsymbol{X}_{k,\mathcal{A}}]$, and $\kappa_0 = \max_{\boldsymbol{\delta}\in\mathcal{N}_0}\kappa(\rho;\boldsymbol{\delta})$.

Theorem 2.1 *Suppose the regularity A1, A2 and A3 (A3') are satisfied. Assume that the dataset is divided into K subsets and $n_k = O(n/K)$, then with probability at least*

$$1 - 2Ks\exp\{-v_{n,K}^2/(nK)\} - 2K(p-s)\exp\{-u_{n,K}^2/(nK)\},$$

the combined estimator is sign consistent, i.e. $\operatorname{sgn}(\hat{\boldsymbol{\beta}}^{(c)}) = \operatorname{sgn}(\boldsymbol{\beta}^0)$. *More specifically, we have* $\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_\infty \le \beta_*/2$ *and* $\hat{\boldsymbol{\beta}}_{\mathcal{A}^c}^{(c)} = 0$.

Because of the split-and-conquer approach, the growth rate of $\log p$ is controlled by $n^{1-\alpha_1}/K$ rather than $n^{1-\alpha_1}$. In other words, if the split-and-conquer approach wants to detect the same signal strength as complete dataset analysis, it pays the price of allowing only the slower rate $p = o(e^{(n^{1-\alpha}/K)})$ rather than $p = o(e^{(n^{1-\alpha})})$. When $K = O(1)$, the combined estimator achieves the same convergency order as the complete data analysis although the probability of sign consistency is still smaller.

2.3.2 Oracle property

In this subsection, we show that, after we strengthen the regularity conditions, our combined estimator can also have such an oracle property with a better rate of model selection consistency and asymptotic normality.

First we show that the combined estimator can converge at the order of $O(\sqrt{s/n})$ under L_2 norm. Furthermore, we show that the combined estimator obtains asymptotic normality with the same variance as the penalized estimator using entire data all together. Therefore, we fully establish the asymptotic equivalence between the combined estimator and the penalized estimator using entire data all together.

Assume the following conditions on the design matrix

A4

$$\min_{\boldsymbol{\delta}\in\mathcal{N}_0} \lambda_{\min}(\boldsymbol{X}_{k,\mathcal{A}}^T \boldsymbol{\Sigma}(\tilde{\boldsymbol{\theta}}_k^{\boldsymbol{\delta}}) \boldsymbol{X}_{k,\mathcal{A}}) = cn_k,$$

$$tr(\boldsymbol{X}_{k,\mathcal{A}}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0) \boldsymbol{X}_{k,\mathcal{A}}) = O(sn_k),$$

$$\|\boldsymbol{X}_{k,\mathcal{A}^c}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0) \boldsymbol{X}_{k,\mathcal{A}}\|_{2,\infty} = O(n_k),$$

$$\max_{\boldsymbol{\delta}\in\mathcal{N}_0} \max_{j=1,\ldots,p} \lambda_{\max}[\boldsymbol{X}'_{k,\mathcal{A}} \mathrm{diag}\{|\boldsymbol{x}_{k,j}| \circ |\boldsymbol{\mu}''(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})|\}\boldsymbol{X}_{k,\mathcal{A}}] = O(n_k),$$

$$\lambda_{\max}((\boldsymbol{X}_{\mathcal{A}}^T \Sigma(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}})^{-1}) = O(n^{-1}),$$

where $\tilde{\boldsymbol{\theta}}_k^{\boldsymbol{\delta}} = \boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta}$, $\|A\|_{2,\infty} = \max_{\|\boldsymbol{v}\|_2=1} \|A\boldsymbol{v}\|_\infty$ and $\mathcal{N}_0 = \{\boldsymbol{\delta} \in \Re^{s_n} : \|\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_\infty \leq \beta_*/2\}$. Here, the derivative is taken componentwise and $\circ$ is componentwise product.

Condition **A4** is generally stronger than condition **A1** as it restricts the L_2 norm instead of L_∞ norm of the design matrix.

Let $u_{n,K}$ be a diverging sequence depending on the total sample size n and the number of subsets K. We impose the following conditions on the tuning parameters and signal strength

$$\textbf{A5}\quad \begin{array}{ll} \sqrt{s/n_k}/\beta_* = o(1), & \sqrt{s/n_k}/\lambda_k = o(1), \\ \lambda_k \kappa_0 = o(1), & \rho'(\beta_*/2; \lambda_k) = \min\{O(n_k^{-1/2}), o(s^{-1/2} n^{-1/2})\}, \\ u_{n,K}/(n\lambda_k) = o(1), & pK \exp\{-u_{n,k}^2/(nK)\} = o(1), \end{array}$$

where $\kappa_0 = \max_{\boldsymbol{\delta} \in \mathcal{N}_0} \kappa(\rho; \boldsymbol{\delta})$.

Condition **A5** controls the bias term introduced by the penalty function. Compared with Condition **A3**, Condition **A5** also requires that the probability of selecting true model goes to 1 by restricting $u_{n,k}$.

$$\textbf{A6}\quad \begin{array}{l} \max_{i=1,\ldots,n} E|y_i - b'(\theta_i^0)|^3 = O(1), \\ \sum_{i=1}^n (\boldsymbol{z}_i' \boldsymbol{B}^{-1} \boldsymbol{z}_i)^{3/2} \to 0 \ \text{ as } \ n \to \infty, \end{array}$$

where $\boldsymbol{B} = \boldsymbol{X}_{\mathcal{A}}' \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}$ and $\boldsymbol{X}_{\mathcal{A}} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)'$.

The equivalence results are stated in the following theorem.

Theorem 2.2 *Suppose the regularity conditions A4-A5 hold and $s = O(n_k^{1/3})$. Assume the dataset is divided into K subsets, $K \le O(s)$, and $n_k = O(n/K)$.*

(i) With probability approaching 1, $\hat{\boldsymbol{\beta}}_{\mathcal{B}}^{(c)} = 0$ as $n \to \infty$ and $\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2 = O(\sqrt{s/n})$.

(ii) Suppose assumption A6 holds and further assume $s = o(n_k^{1/3}/K^{1/3})$ and $\rho'(\beta_*/2; \lambda_k) = o(s_n^{-1/2} n_k^{-1/2} K^{-1/2})$. If $\boldsymbol{D}$ is a $q \times s$ matrix such that $\boldsymbol{D}\boldsymbol{D}' \to \boldsymbol{G}$, where $\boldsymbol{G}$ is a

$q \times q$ symmetric positive definite matrix, we have

$$\boldsymbol{D}[\boldsymbol{X}_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)\boldsymbol{X}_{\mathcal{A}}]^{1/2}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) \xrightarrow{\mathrm{D}} N(\mathbf{0}, \phi\boldsymbol{G}).$$

Fan & Lv (2011) show that $\boldsymbol{D}[\boldsymbol{X}_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)\boldsymbol{X}_{\mathcal{A}}]^{1/2}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(a)} - \boldsymbol{\beta}_{\mathcal{A}}^0) \xrightarrow{\mathrm{D}} N(\mathbf{0}, \phi\boldsymbol{G})$. From Theorem 2.2 (ii), the combined estimator $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)}$ has the same limiting normal distribution as $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(a)}$ under similar conditions. Thus, even with K going to infinity, the combined estimator $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)}$ is asymptotically as efficient as the penalized estimator $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(a)}$ which is obtained using the entire data all together. Together with the fact that both estimators are model selection consistent, the combined estimator is asymptotically equivalent to the penalized estimator analyzing the entire data all together. In theorem 2.2, we require β_* needs to be larger than $O(\sqrt{s/n_k})$ to ensure the sign consistency for the penalized estimator of each subset. Compared with the signal strength $O(\sqrt{s/n})$ required by the penalized estimator using the entire dataset, the combined estimator needs larger coefficients to entail L_2 norm consistency. In addition, s has to be at the order of $O(n_k^{1/3})$ which is smaller than $O(n^{1/3})$ as needed by analyzing the entire dataset.

2.3.3 Error control

Since the observations are independent and the splitting is random, the majority voting proposed in our approach enables us to find an upper bound of the expected number of falsely selected variables and a lower bound of the expected number of truly selected variables for the combined estimator. The bounds lead to improved performance of the combined estimator on model selection. Let $\bar{s}_k = E(|\hat{\mathcal{A}}_k|)$ be the average number of selected variables of the penalized estimator from the k^{th} subset. Theorem 3 below provides an upper bound of the expected number of falsely selected variables and a

lower bound of the expected number of truly selected variables, both of which depend on the choice of the threshold w in the proposed majority voting method. A similar result is also provided by Meinshausen & Buhlmann (2010) and Fan et al. (2009) which only considered the $K = 2$ situation.

Theorem 2.3 *Assume the distribution of* $\{\mathbf{1}_{j\in\hat{\mathcal{A}}_k} : j \in \mathcal{A}\}$ *and* $\{\mathbf{1}_{j\in\hat{\mathcal{A}}_k} : j \in \mathcal{B}\}$ *are exchangeable for all* $k = 1, \ldots, K$*. Also, assume the penalized estimators used are not worse than random guessing, i.e.* $E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|)/E(|\mathcal{B} \cap \hat{\mathcal{A}}_k|) \geq |\mathcal{A}|/|\mathcal{B}|$*, for the set of selected variables* $\hat{\mathcal{A}}_k$ *of any penalized estimator. If* $s^* = \sup_k \bar{s}_k$*,* $s_* = \inf_k \bar{s}_k$ *and* $w \geq s^*K/p - 1$*, then for the combined estimator* $\hat{\boldsymbol{\beta}}^{(c)}$*,*

(i) the expected number of false selected variables has an upper bound: $E(|\mathcal{B}\cap\hat{\mathcal{A}}^{(c)}|) \leq |\mathcal{B}|\{1 - F(w|K, s^*/p)\}$*,*

(ii) the expected number of truly selected variables has a lower bound: $E(|\mathcal{A}\cap\hat{\mathcal{A}}^{(c)}|) \geq |\mathcal{A}|\{1 - F(w|K, s_*/p)\}$*,*

where $F(\cdot|m, q)$ *is the cumulative distribution function of binomial distribution with* m *trials and success probability* q*.*

In an extreme case with $w = K - 1$, the combined estimator selects a variable when it is selected by all penalized estimators from the K subsets. Then, the upper bound for the expectation of selected noise variables is $(s^*)^K/p^{K-1}$. Usually, it is hard to get s^*. However, as long as s^* is bounded by $c^{1/K}p^{1-1/K}$, the average number of noise variables is bounded by c, where c is constant. In sparse models, s^* is usually small and so is c. Therefore, the combined estimator controls the model selection error in a foreseeable way. In another extreme case with $w = 0$, the combined estimator selects a variable when it is selected by at least a penalized estimator from the K subsets. In

this case, the lower bound for the expected number of truly selected variables is tight, achieving the true number of non-zero set $|\mathcal{A}|$. However, in this latter case, the upper bound for the expected number of false selected variables is very loose, up to $|\mathcal{B}|$ the number of variables in the entire noise set.

Indeed, there is a trade off between the upper and lower bounds in Theorem 2.3 for the choice of w. A larger w typically gives us a smaller upper bound of the expected number of false selected variable as well as a smaller lower bound of the expected number of truly selected variables. A smaller w typically gives us a larger upper bound of the expected number of false selected variable as well as a larger lower bound of the expected number of truly selected variables. We use $w = K/2$ in our numerical studies in Section 2.4. It appears to be able to provide a good balance between selecting nonzero coefficients in the true model and excluding noise variables, provided that s^* is smaller than half of p. Our numerical studies show that when $w = 2$, the combined estimators select very few noise variables while keep most variables in the true model. Our empirical experience seems to suggest that the best choice of w is in $[K/3, K/2]$, depending on whether higher sensitivity or higher specificity is more desirable.

2.3.4 Computing issues

In this subsection, we discuss potential computing savings through the split-and-conquer approach. We have the following simple proposition for a computational demanding procedure.

Lemma 2.1 *Assume a penalized linear regression problem with n observations and p variables where $n \leq p$ and LARS algorithm is applied to perform variable selection and coefficients estimation. Then, the LARS algorithm using the entire dataset needs*

$5n^3/3 + 23n/6 + 4n^2(p - 7/8) + 6np$ *computing steps for the best case and* $23n^3/3 + 71n/6 + 8n^2(p - 31/16) + 12np$ *computing steps for the worst case.*

Theorem 2.4 *Under the assumption in lemma 2.1, suppose the dataset is split into* K *subsets and the* k^{th} *subset has* n_k *observations. If the computing effort of the combination is ignorable, with* $n_k = O(n/K)$, $K \geq 3$ *and* $n \geq 4(4 + 3p)/\{1 + 8p(1 - 2/K) + 31/K - 7\}$, $p \geq 2$, *split-and-conquer approach always has less computing steps compared with the LARS algorithm using the entire dataset.*

Theorem 2.4 provides an intuitive interpretation on how much computing time can be saved for LARS algorithm. In fact, the split-and-conquer approach can results in a computing saving by the order of K^2 times in most cases because LARS algorithm have a computing order of $O(n^3)$ when $n \geq p$. In the numerical example of a Gaussian regression with L1 penalty in Section 2.4.1, the exact order of the computing saving K^2 is achieved using the LARS algorithm. Figure 2.1 below demonstrates how computing time changes for different n using LARS algorithm. Detailed simulation settings can be found in section 2.4.1. According to Figure 2.1, the computing time is decreased dramatically for the split-and-conquer approach compared with the computing time required for analyzing the entire dataset.

In fact, the split-and-conquer approach can achieve tremendous computing savings for any statistical procedure that requires $O(n^a)$ computing steps, $a > 1$. Suppose the dataset is split into K subsets with almost equal sample size $n_k = O(n/K)$ and the computing effort of the combination is ignorable. Then, the split-and-conquer approach only needs $K \times O((n/K)^a)$, that is $O(n^a/K^{a-1})$, steps. Thus, using the split-and-conquer approach results in a computing saving by the order of K^{a-1} times. A

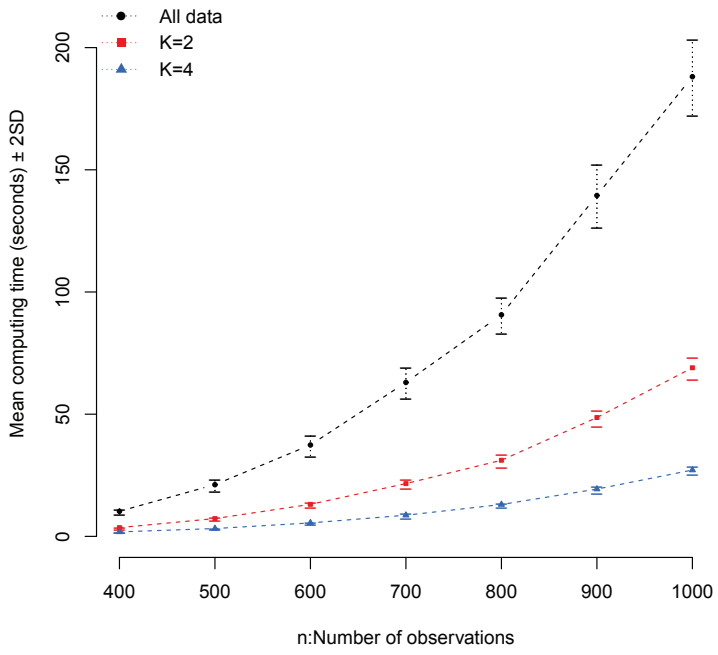

Figure 2.1: Computing time comparison for different K using LARS algorithm ($p = 2n$): Mean $\pm$ 2Standard Deviation (SD) over 100 replications

similar finding in a computational intensive robust multivariate scale estimation is also reported in Section 5.3 of Singh et al. (2005). Under more complex situations with generalized linear models and more complicated algorithms such as those studied in Section 2.4.2, the computing saving is less than what would be predicted by the simple calculation, although the computing time is still reduced in a great amount in all those examples. The complexity of an algorithm and its computing time are associated with its computing paths in search for a numerical solution of the optimization. Cross-validations used for selecting the tuning parameter in the penalized likelihood function add another degree of complexity to the problem. We speculate that the computing time for analysis of the K subset is different, sometimes substantially, from one to another in these more complex situations. This makes a prediction of computing savings a much harder task. Although we can not use the calculation to obtain computing savings in the more complex cases, it still provides an intuition that can help us understand why the split-and-conquer approach can reduce computing time.

2.4 Numerical studies

In this section, we provide numerical studies, using both simulation and real data, to illustrate the performance of the proposed split-and-conquer approach. We also compare the combined estimators with their corresponding penalized estimators obtained using the entire data all together, whenever the latter approach can be performed and does not reach the limits of our computer. The L_1 norm, SCAD and MCP, three of the most widely used penalty functions in the literature, are used in our illustration. We focus on two models, the Guassian linear regression model and the logistic model, with different choices of sample size n, number of parameters p and true model size s (the

number of nonzero regression parameters). All analyses are performed on a W35653 20GHz, 2G(RAM) workstation using R 2.13.1 under Windows 7.

2.4.1 Linear regression with L_1 norm penalty

We consider in this subsection a simple case with a linear regression and the L_1 norm penalty to demonstrate the properties of the combined estimator. In particular, the response variable $\boldsymbol{y}$ follows a Guassian linear model

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

where ε are IID $N(0,1)$ errors and the explanatory variables $\boldsymbol{X}$ are generated from a $N(0, \boldsymbol{I})$ distribution with $\boldsymbol{I}$ being identity matrix. In our simulation study, we generate p variables and the true model $\mathcal{A}^0$ contains $s = \lfloor\sqrt{p}\rfloor$ nonzero coefficients with values around $\sqrt{2K\log(p)/n}$. The total sample size is picked to be $n <= p$. To get the LASSO estimators using the L_1 norm penalty, the LARS algorithm (Efron et al., 2004) is applied and BIC criterion is used for selecting the tuning parameter. When $p \geq n$, the computing order of the LARS is $O(n^3)$ that is computationally intensive.

We repeat our simulation 100 times. For the final overall estimators, we record the mean of computing time and the number of selected nonzero coefficients. To demonstrate the error control property, we also calculate model selection sensitivity and model selection specificity. Here, model selection sensitivity is defined as the number of truly selected variables divided by the true model size, and model selection specificity is defined as the number of truly removed variables divided by the number of noise variables. The simulation results are shown in Table 2.1. In Table 2.1, $K = 1$ means the entire dataset is used to get the LASSO estimator; otherwise, the combined estimator proposed in this paper is used. To examine the performance of the combined estimator ,

Table 2.1: Comparison of the combined estimator and the complete estimator (with standard deviation in the parenthesis)

Simulation setting			Computing time (in second)	w	Model selection		
n	p	K			# selected variables	sensitivity (in %)	specificity (in %)
500	500	1	41.55 (5.37)	-	36.01 (5.87)	100 (0)	97.07 (1.23)
		2	5.77 (0.51)	1	66.56 (9.72)	100 (0)	90.68 (2.03)
		4	2.74 (0.46)	1	157.55 (16.05)	100 (0)	71.64 (3.36)
				2	37.49 (5.10)	98.68 (2.53)	96.70 (1.06)
		6	1.71 (0.22)	1	221.92 (14.25)	99.73 (1.08)	58.16 (2.93)
				2	63.96 (7.63)	96.73 (3.66)	91.07 (1.58)
				3	24.11 (2.61)	86.73 (7.33)	98.95 (0.42)
500	800	1	24.72 (1.77)	-	48.32 (6.60)	100 (0)	97.37 (0.86)
		2	8.10 (0.60)	1	102.75 (11.84)	99.93 (0.50)	90.31 (1.53)
		4	3.60 (0.38)	1	240.06 (14.99)	99.18 (1.89)	72.50 (1.94)
				2	50.96 (5.92)	92.29 (5.39)	96.75 (0.74)
		6	2.52 (0.27)	1	294.60 (11.50)	97.18 (2.93)	65.36 (1.50)
				2	69.31 (6.84)	83.50 (7.28)	94.05 (0.90)
				3	20.66 (3.34)	58.46 (9.71)	99.44 (0.27)
500	1000	1	28.06 (1.85)	-	59.09 (7.80)	100 (0)	97.20 (0.81)
		2	10.04 (0.60)	1	135.72 (16.58)	99.81 (1.32)	89.28 (1.71)
		4	4.48 (0.41)	1	284.18 (15.89)	97.03 (2.68)	73.85 (1.64)
				2	54.13 (5.80)	83.53 (6.86)	97.17 (0.56)
		6	2.92 (0.27)	1	325.83 (10.94)	93.19 (4.34)	69.42 (1.15)
				2	64.46 (5.84)	70.31 (7.58)	95.67 (0.63)
				3	16.60 (3.16)	41.88 (7.77)	99.67 (0.19)
1000	1000	1	393.10 (46.82)	-	47.86 (6.54)	100 (0)	98.36 (0.68)
		2	57.30 (2.87)	1	83.51 (12.31)	98.36 (0.68)	94.68 (1.27)
		4	20.21 (2.24)	1	217.77 (18.11)	100 (0)	80.81 (1.87)
				2	46.53 (4.72)	99.87 (0.62)	98.50 (0.49)
		6	12.66 (1.63)	1	381.51 (21.69)	99.94 (0.44)	63.89 (2.24)
				2	94.18 (8.31)	99.81 (0.75)	93.57 (0.86)
				3	37.51 (3.13)	97.59 (2.62)	99.35 (0.30)

we try different values of K and w, where $K = 2, 4, 6$ and $w = 1, \ldots, \lfloor K/2 \rfloor$.

According to Table 2.1, all estimators select some noise variables in addition to the true s nonzero variables. This is consistent with a known performance of the LASSO-type estimators that they usually intend to include more variables than desired in model selections. For the same fixed w, when K gets larger, the combined estimator gets worse because each subset has less data. But for the same K, the combined estimator with larger w shows the benefit of error control with high model selection specificity. Moreover, the computing time is decreasing when K is increasing. Since the computing

order for the LARS algorithm is $O(n^3)$ when $p \geq n$, Theorem 2.4 in Section 2.3.3 suggests that the split-and-conquer approach can save computing time by the order of K^2. This is exactly the order achieved in this simulation study, as indicated in column 4 in Table 2.1.

2.4.2 Generalized linear model with SCAD and MCP penalties

The SCAD and MCP estimators are two commonly used estimators that are obtained based on non-concave penalized likelihood functions. They often have a better performance than the LASSO estimators, in terms of selecting a tighter model and fewer noise variables. We consider in this subsection both the SCAD and MCP estimators under both the linear regression and logistic models.

For the linear regression case, the response variable $\boldsymbol{y}$ follows the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}+\varepsilon$, where ε are IID $N(0,1)$ errors. For the logistic regression case, the response variable $\boldsymbol{y}$ follows the Bernoulli distribution with the success probability $p(\boldsymbol{X}\boldsymbol{\beta}) = e^{\boldsymbol{X}\boldsymbol{\beta}}/(1+e^{\boldsymbol{X}\boldsymbol{\beta}})$. In our simulations, we consider two settings to generate the design matrix $\boldsymbol{X}$: one is for independent variables and the other is for correlated variables.

1. Independent variables: a set of p variables are generated from a $N(0, \boldsymbol{I})$ distribution, where $\boldsymbol{I}$ is identity matrix.

2. Correlated variables: a set of p variables are generated from a $N(0, \boldsymbol{\Sigma})$ distribution, where $\boldsymbol{\Sigma}(i,j) = 0.6^{|i-j|}$ is the covariance matrix.

We consider two settings of sample sizes: $n = 10000$ that is large but not too large and $n = 100000$ that is very large. In the linear regression, the number of parameters $p = 1000$ and in the logistic model, the number of parameters $p = 200$. In all cases, the

true model contains $s = 30$ nonzero coefficients. The true model size $s = 30$ is chosen to be relatively small compared with p and n. In order to get the SCAD and MCP estimators, the NCVREG algorithm (Breheny & Huang, 2011) is applied and a 10-fold cross-validation is used to select the tuning parameters.

The simulation is repeated 100 times. Similarly as in the example in Section 2.4.1, we record the computing time and the number of selected variables and calculate model selection sensitivity and specificity. In addition, the MSE (mean squared error) is calculated in the linear regression case and the misclassification rate with 0.5 as threshold is reported in the logistic regression case. The results are displayed in Table 2.2. In the table, $K = 1$ refers to the entire data is used all together with no splitting. For any $K > 1$, the proposed split-and-conquer approach is applied.

According to Table 2.2, the SCAD estimators performs similar to the MCP estimators. In either case, the combined estimator has good model selection results with high model selection sensitivity and specificity that are similar to those of the penalized estimator using entire data all together. Moreover, in the linear regression case, the combined estimator has a similar MSE to that of the penalized estimator using entire data all together. In the logistic regression case, the misclassification rate of the combined estimator is also close to that of the penalized estimator using entire data all together.

The computing time is reduced through the split-and-conquer procedure, although we cannot calculate the exact order of computing savings in these complicated settings. For both the SCAD and MCP penalties, the proposed split-and-conquer approach can reduce the computing time by almost 10 times on average in the linear regression setting. For the logistic model, the average saving is a little less. When the explanatory

variables are independent, the combined estimator needs about half of the time compared to directly performing the same analysis on the entire data all together. When the explanatory variables are correlated, the combined estimator by the proposed method can save up to 25% time compared to directly performing the same analysis on the entire data all together. When the sample size $n = 100000$, we are not able to perform either the SCAD or the MCP regression on the entire data all together due to computer memory limitations. However, the combined estimators can still be obtained using the split-and-conquer procedure.

We also compare the values of the combined estimators and the penalized estimators analyzing entire data all together in all the settings of Table 2.2 when both are available; see Figure 2.2. For the linear regression case, the boxplots of the $\boldsymbol{\beta}$ estimation in the true model $\mathcal{A} = \{j : \beta_j^0 \neq 0\}$ are plotted in the top panels. We can see that the estimation of the combined estimators has the similar mean and spread to those of the estimators using entire data all together. For the logistic regression, the boxplots of the $\boldsymbol{\beta}$ estimation in the true model are plotted in the bottom panels. In the logistic model case, the estimation of covariance matrix can influence the combined estimator. We use the maximum likelihood estimator based on only the selected variables in $\hat{\mathcal{A}}$ to get the weight matrix. Again, the combined estimators using the proposed split-and conquer approach perform similarly to the penalized estimators using entire data all together.

2.4.3 Numerical analysis on POEs manifest data

In this subsection, we study a set of manifest data collected at the US Port of Entries (POEs) to demonstrate an application of the split-and-conquer approach. To counter potential terrorists' threats, substantial efforts have been made in devising strategies

Table 2.2: Comparison of the combined estimator and the complete estimator (standard deviation in the parenthesis)

Part I: Linear regression								
Simulation setting					Model selection			
Design matrix	n	p	K	Computing time (in second)	# selected variables	sensitivity (in %)	specificity (in %)	MSE
SCAD: Linear regression								
Independent	10000	1000	1	815.27 (77.98)	34.58 (9.81)	100 (0)	99.53 (1.01)	1.00 (0.01)
			10	104.96 (9.55)	30 (0)	100 (0)	100 (0)	1.00 (0.01)
Correlated	10000	1000	1	755.4 (157.56)	34.00 (12.22)	96.00 (19.79)	99.46 (1.02)	0.96 (0.20)
			10	289.17 (61.03)	28.72 (6.13)	95.87 (19.78)	100 (0)	1.00 (0.01)
Independent	100000	1000	1	- (-)	- (-)	- (-)	- (-)	- (-)
			100	1136.70 (74.65)	30 (0)	100 (0)	100 (0)	1.00 (0.01)
Correlated	100000	1000	1	- (-)	- (-)	- (-)	- (-)	- (-)
			100	3074.53 (25.01)	30 (0)	100 (0)	100 (0)	1.06 (0.01)
MCP: Linear regression								
Independent	10000	1000	1	2243.45 (155.82)	34.58 (9.81)	100 (0)	99.79 (0.41)	1.00 (0.01)
			10	163.72 (12.95)	30 (0)	100 (0)	100 (0)	1.00 (0.01)
Correlated	10000	1000	1	1244.73 (80.86)	31.92 (5.69)	100 (0)	99.80 (0.59)	0.99 (0.01)
			10	442.14 (42.42)	29.98 (0.14)	99.93 (0.47)	100 (0)	1.01 (0.02)
Independent	100000	1000	1	- (-)	- (-)	- (-)	- (-)	- (-)
			100	1565.54 (132.38)	30 (0)	100 (0)	100 (0)	1.00 (0.01)
Correlated	100000	1000	1	- (-)	- (-)	- (-)	- (-)	- (-)
			100	4256.52 (215.60)	30 (0)	100 (0)	100 (0)	1.02 (0.01)

Part II: Logistic regression								
Simulation setting					Model selection			
Design matrix	n	p	K	Computing time (in second)	# selected variables	sensitivity (in %)	specificity (in %)	Misclassificaton rate (in %)
SCAD: Logistic regression								
Independent	10000	200	1	198.85 (5.88)	35.54 (5.71)	100 (0)	96.74 (3.36)	17.32 (0.40)
			5	116.49 (2.78)	31.70 (1.33)	100 (0)	99.00 (0.78)	17.40 (0.38)
Correlated	10000	200	1	463.61 (20.16)	38.18 (5.58)	99.33 (1.35)	95.02 (3.15)	9.90 (0.29)
			5	359.29 (7.94)	32.38 (2.42)	96.07 (2.75)	97.84 (1.27)	10.10 (0.26)
Independent	100000	200	1	- (-)	- (-)	- (-)	- (-)	- (-)
			20	1352.14 (76.2)	30 (0)	100 (0)	100 (0)	17.38 (0.12)
Correlated	100000	200	1	- (-)	- (-)	- (-)	- (-)	- (-)
			20	4014.48 (284.69)	29.97 (0.2)	99.87 (0.67)	100 (0)	9.96 (0.09)
MCP: Logistic regression								
Independent	10000	200	1	201.46 (6.74)	31.8 (2.77)	100 (0)	98.94 (1.63)	17.31 (0.34)
			5	118.85 (3.17)	30.24 (0.62)	99.87 (0.66)	99.84 (0.34)	17.38 (0.35)
Correlated	10000	200	1	582.182 (59.02)	35.48 (4.22)	98.73 (1.89)	96.55 (2.27)	9.84 (0.33)
			5	557.43 (22.7)	28.7 (1.63)	92.93 (3.85)	99.52 (0.60)	10.17 (0.32)
Independent	100000	200	1	- (-)	- (-)	- (-)	- (-)	- (-)
			20	1301.95 (63.27)	30 (0)	100 (0)	100 (0)	17.34 (0.13)
Correlated	100000	200	1	- (-)	- (-)	- (-)	- (-)	- (-)
			20	4485.9 (186.29)	29.58 (0.50)	98.60 (1.66)	100 (0)	10.00 (0.09)

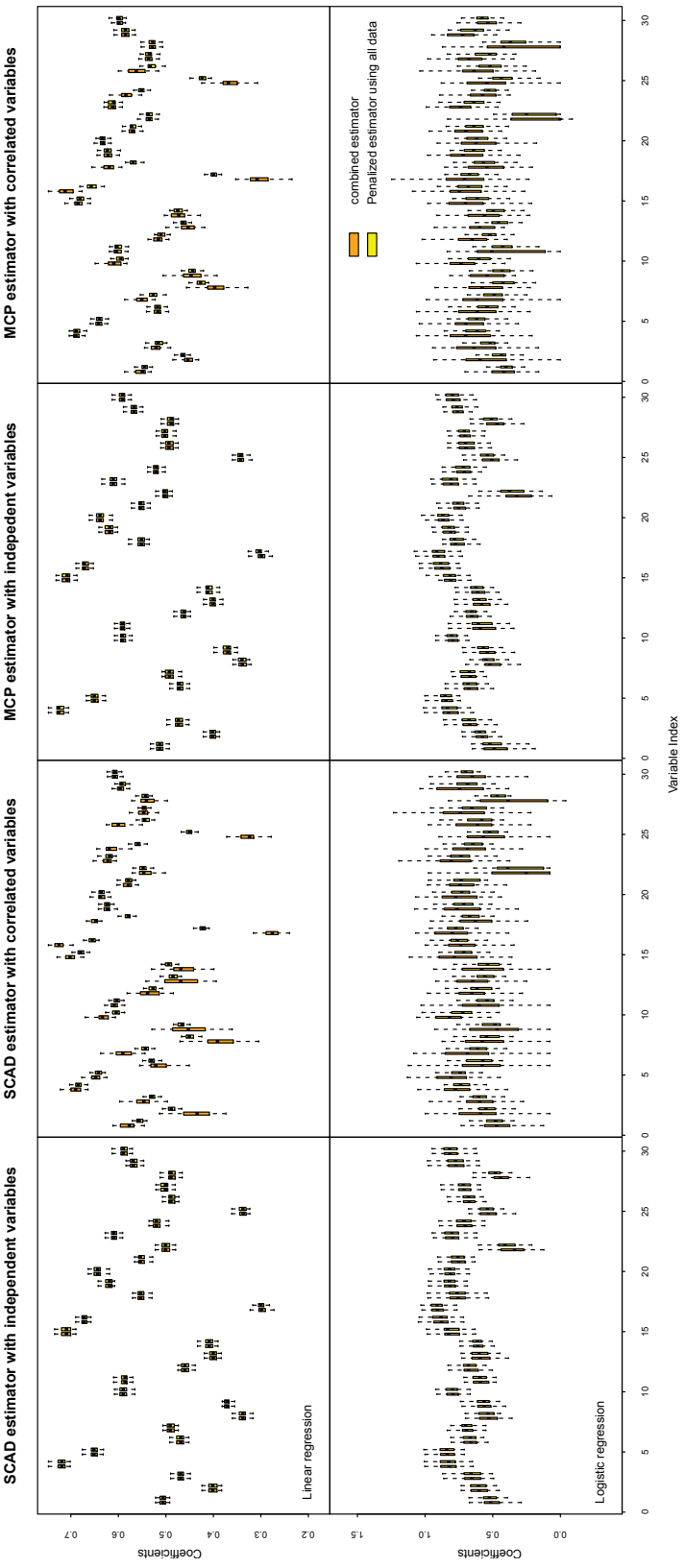

Figure 2.2: Comparison of parameter estimation for the combined estimator and the penalized estimator using all data. Box plots of estimation for variables in the true model. Orange: the combined estimator; Yellow: the estimator using all data. Top panels: Linear regression; bottom panels: Logistic regression

for inspecting containers coming through the US POEs every day to interdict illicit nuclear and chemical materials. Manifest data, compiled from the custom forms submitted by merchants or shipping companies, are collected by the US custom offices and the Department of Homeland Security (DHS). Analysis of the manifest data is a part of effort to build up layered defenses for the national security. In a nuclear detection project sponsored by the Command, Control, and Interoperability Center for Advanced Data Analysis (CCICADA), a Department of Homeland Security (DHS) Center of Excellence, we obtain a set of manifest data that contain all shipping records coming through the POEs across the US in February, 2009. The goal is to make quantitative evaluations of the manifest data and to develop an effective risk scoring approach that can be used to assess future shipments. In our project, a logistic regression model has been used to enhance the effectiveness of the real-time inspection system with binary response variable indicating high-risk shipments. Since not all information collected in the manifest data are relevant to risk scoring and there are also many redundant information, we need to determine the effects of different sources of information in the manifest data and penalized regression provides a way to evaluate the importance of these variables. Table 2.3 provides the definition and a description of some variables contained in the manifest data. Most of these variables are categorical and dummy variables for each categorical variable are created which results in $p = 213$ variables in total. There are also text fields that can potentially lead to a much larger p. To simply our discussion and without loss of our focus, we only illustrate the proposed split-and-conquer approach using this $p = 213$ variables and we do not consider any semantic analysis and text mining approaches in this paper.

Practical issues and challenges exist in carrying out this important task. Due to

Table 2.3: Manifest data: Dictionary of Variables

Variables	Number of Categories	Definition
$\boldsymbol{X}_1$	9	Vessel Country Code
$\boldsymbol{X}_2$	69	Voyage Number
$\boldsymbol{X}_3$	9	dp of Unlading
$\boldsymbol{X}_4$	14	Foreign Port Lading
$\boldsymbol{X}_5$	68	Foreign Port
$\boldsymbol{X}_6$	35	Inbond Entry Type
$\boldsymbol{X}_7$	17	Container Cotents

the enormous size of traffic and a large number of entry sites, it is impossible for us to analyze the whole data simultaneously on a single computer. For instance, there are 164721 shipments in one week from February 20, 2009 to February 26, 2009. A computer with 2 GB memory and 3.2GHz CPU fails to perform the SCAD penalized regression on the one-week data. Even if high-performance computer is available, it will takes a long time to carry out the task and this is very inefficient in practice, especially we may need to constantly update the models over the months and years. Nevertheless, we can solve this problem by applying the split-and-conquer approach with the assumption that the underlying regression model stays more or less the same over a short period of time of one week or one month.

Because of security concerns, the indicator of high-risk shipments are not accessible to us, but we have been told to use the rate 1% to 10% of cargo containers that need further inspections in the context of inspections of drugs and other illicit materials. To illustrate our approach, we turn to a simulation to generate the risk scores based on the given manifest data. In particular, potential influential characteristics are selected to generate the risk scores using logistic models. Then, we perform the SCAD penalized regression on everyday's data and combine the seven daily estimators together to obtain an overall combined estimator. Note that, due to the computing limitations of our

Table 2.4: Comparison of the combined weekly estimator and daily estimators (standard deviation in the parenthesis)

	Model selection			
	# of selected variables	Sensitivity (in %)	Specificity (in %)	Misclassificaton rate (in %)
Week (Combined)	21.06 (0.38)	95.25 (0.09)	99.95 (0.14)	3.97 (0.05)
Mon	32.66 (4.00)	92.53 (0.36)	94.2 (1.78)	3.99 (0.05)
Tues	29.18 (3.07)	95.4 (0.05)	96.14 (1.44)	3.98 (0.05)
Wed	9.22 (4.58)	23.13 (1.2	98.05 (1.18)	3.99 (0.05)
Thur	10.86 (4.6)	27.73 (1.08)	97.76 (1.28)	3.98 (0.05)
Fri	25.6 (2.09)	95.45 (0)	97.83 (0.98)	4.00 (0.05)
Sat	29.76 (3.47)	95 (0.14)	95.82 (1.61)	3.98 (0.05)
Sun	30.6 (3.31)	95.1 (0.12)	95.44 (1.57)	3.99 (0.05)

personal computer, we are not able to perform the SCAD analysis on the whole week of data all together.

The results from the split-and-conquer approach are displayed in Tables 2.4 and 2.5, in which we report the model selection sensitivity, model selection specificity, misclassification rate and the average estimates of the non-zero parameters from 100 replications, based on the split-and-conquer approach as well as the SCAD penalized regression using the data of a single day. The $s = 22$ non-zero parameters are from three categorical variables: Vessel Country Code, Foreign Port Landing and Container Contents. Clearly, the split-and-conquer approach succeeds in performing the penalized logistic regression analysis on the whole week manifest data. As we can see from Table 2.4, the split-and conquer approach has identified most influential variables in the manifest data. In particular, the combined estimator has both high model selection sensitivity and specificity. On a contrast, the daily estimators either select many more noise variables or exclude many influential variables. Also, the combined estimator is more stable than daily estimators because it has much smaller variances in the values of the average model size, model selection sensitivity and specificity. Although the combined

Table 2.5: Manifest data analysis through split-and-conquer approach

Categories	Week (Combined)	Daily estimation						
		Mon	Tues	Wed	Thur	Fri	Sat	Sun
Vessel country code								
PA	0.33(0.06)	0.2(0.17)	0.36(0.15)	0.07(0.14)	0.14(0.14)	0.46(0.07)	0.41(0.16)	0.4(0.14)
LR	1.78(0.07)	1.7(0.22)	1.75(0.19)	0.8(0.39)	1.64(0.16)	1.78(0.16)	1.75(0.17)	1.73(0.13)
DE	0.26(0.06)	0.22(0.17)	0.39(0.16)	0.01(0.06)	0.02(0.11)	0.47(0.11)	0.32(0.19)	0.31(0.2)
Foreign port lading								
570	1.54(0.05)	1.59(0.15)	1.56(0.13)	0.92(0.35)	1.36(0.33)	1.53(0.08)	1.58(0.17)	1.53(0.12)
582	0.9(0.07)	1(0.23)	1.1(0.14)	0.26(0.21)	0.36(0.23)	0.84(0.17)	0.92(0.26)	0.63(0.25)
580	1.13(0.06)	1.39(0.17)	0.85(0.23)	0.03(0.09)	0.45(0.29)	1.33(0.1)	0.72(0.23)	1.27(0.14)
Container contents								
Material	1.31(0.1)	1.98(0.24)	2.03(0.18)	0.12(0.27)	0.1(0.22)	2.06(0.17)	2(0.23)	1.97(0.24)
Animals	0.05(0.11)	0.27(0.21)	0.74(0.28)	0(0)	0(0)	0.63(0.21)	0.47(0.24)	0.46(0.25)
Entertainment	1.04(0.15)	1.55(0.36)	1.75(0.32)	0.03(0.12)	0.03(0.14)	1.85(0.23)	1.48(0.31)	1.56(0.33)
Industry	0.76(0.1)	1.39(0.25)	1.5(0.19)	0.03(0.22)	0.01(0.1)	1.55(0.18)	1.43(0.2)	1.44(0.18)
Cloth	0.65(0.08)	1.31(0.17)	1.37(0.12)	0.03(0.19)	0.02(0.13)	1.4(0.1)	1.32(0.17)	1.3(0.15)
Electro	0.44(0.13)	1.02(0.37)	1.09(0.28)	0.01(0.12)	0.01(0.12)	1.38(0.26)	0.91(0.26)	1.02(0.28)
Food	0.7(0.08)	1.41(0.14)	1.4(0.15)	0.02(0.17)	0.05(0.19)	1.46(0.11)	1.36(0.14)	1.34(0.12)
Furniture	1.34(0.11)	2.01(0.25)	2.09(0.22)	0.08(0.24)	0.12(0.23)	2.14(0.18)	2.01(0.26)	1.95(0.22)
Hardware	0.24(0.07)	0.88(0.18)	0.94(0.14)	0.01(0.1)	0(0.03)	0.97(0.1)	0.87(0.17)	0.9(0.15)
Health	0.53(0.09)	1.18(0.15)	1.23(0.13)	0.02(0.14)	0.01(0.12)	1.25(0.1)	1.19(0.15)	1.18(0.13)
Home	1.18(0.1)	1.91(0.24)	1.91(0.19)	0.09(0.26)	0.03(0.16)	1.95(0.15)	1.87(0.2)	1.83(0.2)
Motor	0.28(0.14)	0.89(0.3)	1.01(0.32)	0.03(0.25)	0.01(0.1)	1.19(0.29)	1.18(0.37)	1(0.33)
Media	0.98(0.11)	1.69(0.23)	1.75(0.26)	0.03(0.14)	0.02(0.13)	1.79(0.2)	1.47(0.29)	1.46(0.28)
Office	-0.17(0.13)	0.24(0.25)	0.55(0.26)	0.01(0.06)	0(0)	0.55(0.25)	0.4(0.25)	0.54(0.29)
Sporting	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Mature	0.45(0.08)	1.15(0.13)	1.17(0.13)	0.02(0.15)	0.01(0.1)	1.23(0.1)	1.14(0.14)	1.14(0.11)

estimator has a slightly smaller misclassification rate, all estimators have more or less the similar misclassification rates, which are on average slightly less than 4%.

In terms of estimation, as in Table 2.5, the combined estimator also has smaller variance than the penalized estimators that only use daily data. For the categories Animals and Office in Container Contents, some of the daily estimators fails to select them and they are not significant in the combined estimators. Also, the Sporting variable is left out in the model by all the estimators. But all other 19 variables are found by the combined estimator. The same performance can not be achieved by any of the penalized estimators using only daily data. By incorporating one-week information, the split-and-conquer approach provides more reliable results with better performance than any of the daily analysis.

2.5 Discussions

We propose in this chapter a split-and-conquer methodology for analysis of extraordinarily large data that is too large to be analyzed by the existing statistical methods. The split-and-conquer approach contains two operational steps. Firstly, the entire dataset is randomly split into non-overlapped small subsets, and each subset is analyzed separately using desired statistical procedures. Then, the results from all subsets are combined together and provide a final overall statistical inference that contains information from the entire dataset. We demonstrate the split-and-conquer approach for penalized regression models that are widely used in the analysis high-dimensional data.

The split-and-conquer approach provides an applicable way to analyze extraordinarily large datasets using available procedures. The approach is very general and can have many applications. As the entire dataset is split into smaller pieces, each subset requires a smaller storage space and computer memory when we perform our statistical analysis. Moreover, we have shown that the split-and-conquer approach needs less computing time when the desired statistical method is computationally intensive. Even in the case in which the desired statistical method is efficient, a reduced computing time can be expected operationally because we now can analyze different subsets at the same time using different computers. This computing improvement is very useful in many practical applications.

One important step in the split-and-conquer approach is the combination. We have demonstrated in our settings that the combined results obtained from the subsets do not cause any bias or efficiency loss, asymptotically. The specific combination method to be

used depends on the desired statistical procedure. As illustrated by penalized regressions in this paper, the properly weighted and linearly combined estimator is asymptotically equivalent to the one from analyzing the entire data all together although the combined estimator requires slightly stronger conditions. According to Singh et al. (2005), Xie et al. (2011) and Liu (2012), equivalent combined statistics or asymptotic efficiency are achievable for many other models. The proposed split-and-conquer approach can be easily extended to other problem settings as well as problems beyond point estimations including those using hypothesis testings and confidence intervals.

2.6 Appendix

For proving convenience, we state the two lemmas from Fan & Lv (2011).

Lemma 2.2 *Fan & Lv (2011)* $(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}, 0)$ *is a strict local maximizer if*

$$\boldsymbol{X}_{k,\mathcal{A}}^T \boldsymbol{y}_k - \boldsymbol{X}_{k,\mathcal{A}}^T \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_k) - n_k \lambda_k \bar{\rho}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}) = 0, \tag{2.3}$$

$$(n_k \lambda_k)^{-1} \|\boldsymbol{X}_{k,\mathcal{A}^c}^T [\boldsymbol{y}_k - \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_k)]\|_\infty < \rho'(0+), \tag{2.4}$$

$$\lambda_{\min}[\boldsymbol{X}_{k,\mathcal{A}}^T \Sigma(\hat{\boldsymbol{\theta}}) \boldsymbol{X}_{\mathcal{A}}] > n_k \lambda_k \kappa(\rho; \hat{\boldsymbol{\beta}}_{\mathcal{A}}). \tag{2.5}$$

Lemma 2.3 *Fan & Lv (2011) Let* $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ *be the* n*-dimensional independent random response vector and* $\boldsymbol{a} \in R^n$*. Then*

a) If $Y_1, \ldots, Y_n$ *are bounded in* $[c, d]$ *for some* $c, d \in R$*, then for any* $\epsilon \in (0, \infty)$

$$\mathrm{P}(|\boldsymbol{a}^T \boldsymbol{Y} - \boldsymbol{a}^T \boldsymbol{\mu}(\boldsymbol{\theta}^0)| > \epsilon) \leq 2 \exp[-2\epsilon^2 / (\|\boldsymbol{a}\|_2^2 (d-c)^2)].$$

b) If $Y_1, \ldots, Y_n$ *are unbounded and there exist some* $M, v_0 \in (0, \infty)$ *such that*

$$\max_{i=1,\ldots,n} E\{\exp[(Y_i - b'(\theta_i^0))/M] - 1 - |Y_i - b'(\theta_i^0)|/M\} M^2 \leq v_0/2$$

with $\boldsymbol{\theta}^0 = (\theta_i^0, \ldots, \theta_n^0)$*, then for any* $\epsilon \in (0, \infty)$

$$\mathrm{P}(|\boldsymbol{a}^T\boldsymbol{Y} - \boldsymbol{a}^T\boldsymbol{\mu}(\boldsymbol{\theta}^0)| > \epsilon) \leq 2\exp[-\epsilon^2/(2\|\boldsymbol{a}\|_2^2 v_0 + \|\boldsymbol{a}\|_\infty M\epsilon)].$$

Define $c_1 = 2/(d-c)^2$ for bounded responses and $c_1 = 1/(2v_0 + 2M)$ for unbounded responses for the following proofs.

Proof of Theorem 2.1:

According to the definition of the combined estimator, $\hat{\boldsymbol{\beta}}_{\mathcal{A}^c}^{(c)}$=0 if $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}^c} = 0$ for $k = 1, \ldots, K$. So, we will prove the theorem in two steps. First, we show $\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_\infty \leq \beta_*/2$. Then, we prove $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}^c} = 0$ for $k = 1, \ldots, K$.

Define $\boldsymbol{\xi}_k = \boldsymbol{X}_k^T\boldsymbol{y}_k - \boldsymbol{X}_k^T\boldsymbol{\mu}(\boldsymbol{\theta}_k^0)$, $\boldsymbol{\gamma}_k(\boldsymbol{\delta}) = (\gamma_{k1}, \ldots, \gamma_{ks}) = \boldsymbol{X}_k^T\boldsymbol{\mu}(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})$ and $\boldsymbol{\eta}_k(\boldsymbol{\delta}) = n_k\lambda_k\bar{\rho}(\boldsymbol{\delta})$. Consider events $E_{1k} = \{\|\boldsymbol{\xi}_{k,\mathcal{A}}\|_\infty \leq c_1^{-1/2} v_{n,K}/K\}$, and $E_{2k} = \{\|\boldsymbol{\xi}_{k,\mathcal{A}^c}\|_\infty \leq c_1^{-1/2} u_{n,K}/K\}$.

First show that there exists a solution of (2.3) in $\mathcal{N}_0 = \{\boldsymbol{\delta} : \|\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_\infty \leq \beta_*/2\}$, $k = 1, \ldots, K$. Equation (2.3) can be rewritten as

$$\begin{aligned} & \boldsymbol{\gamma}_{k,\mathcal{A}}(\boldsymbol{\delta}) - \boldsymbol{\gamma}_{k,\mathcal{A}}(\boldsymbol{\beta}_{\mathcal{A}}^0) - (\boldsymbol{\xi}_{k,\mathcal{A}} - \boldsymbol{\eta}_k(\boldsymbol{\delta})) \\ = \;& \boldsymbol{X}_{k,\mathcal{A}}^T\Sigma(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}(\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0) + \boldsymbol{r}_k - (\boldsymbol{\xi}_{k,\mathcal{A}} - \boldsymbol{\eta}_k(\boldsymbol{\delta})) = 0, \end{aligned}$$

where $\boldsymbol{r}_k = (r_{k1}, \ldots, r_{ks})$ and $r_{kj} = (\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0)^T\nabla^2\gamma_{kj}(\boldsymbol{\delta}_j)(\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0)$ with $\boldsymbol{\delta}_j$ being a s-dimensional vector on the segment between $\boldsymbol{\delta}$ and $\boldsymbol{\beta}_{\mathcal{A}}^0$.

It is equivalent to

$$\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0 = \{\boldsymbol{X}_{k,\mathcal{A}}^T\Sigma(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}(\boldsymbol{\xi}_{k,\mathcal{A}} - \boldsymbol{\eta}_k(\boldsymbol{\delta}) - \boldsymbol{r}_k). \tag{2.6}$$

In $\mathcal{N}_0$, $\|\boldsymbol{\eta}_k(\boldsymbol{\delta})\|_\infty \le n_k\lambda_k\rho'(\beta_*/2)$. By condition A1 and under event E_{1k}, we have

$$\begin{aligned} & \|\{\boldsymbol{X}_{k,\mathcal{A}}^T\Sigma(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}(\boldsymbol{\xi}_{k,\mathcal{A}} - \boldsymbol{\eta}_k(\boldsymbol{\delta}))\|_\infty \\ \le & \|\{\boldsymbol{X}_{k,\mathcal{A}}^T\Sigma(\boldsymbol{\theta}_k^0)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}\|_\infty\|\boldsymbol{\xi}_{k,\mathcal{A}} - \boldsymbol{\eta}_k(\boldsymbol{\delta})\|_\infty \\ = & O(c_1^{-1/2}b_{s,K}v_{n,K}/n + b_{s,K}\rho'(\beta_*/2;\lambda_k)) \end{aligned}$$

In addition, because of condition A1,

$$\|\boldsymbol{r}_k\|_\infty \le \max_{\boldsymbol{\delta}\in\mathcal{N}_0}\max_{j=1,\ldots,s}\lambda_{\max}[\boldsymbol{X}'_{k,\mathcal{A}}\mathrm{diag}\{|\boldsymbol{x}_{k,j}|\circ|\boldsymbol{\mu}''(\boldsymbol{X}_{k,\mathcal{A}}\boldsymbol{\delta})|\}\boldsymbol{X}_{k,\mathcal{A}}]\|\boldsymbol{\delta}-\boldsymbol{\beta}_{\mathcal{A}}^0\|_2^2 = O(n_k\beta_*^2 s).$$

By condition A3, we have

$$\|\boldsymbol{\delta} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_\infty = o(\beta_*).$$

Thus, there exists a solution $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}$ of (2.6) in $\mathcal{N}_0$. In addition,

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0 = (\sum_{k=1}^K \hat{\boldsymbol{S}}_k)^{-1}[\sum_{k=1}^K \hat{\boldsymbol{S}}_k\boldsymbol{S}_{k,0}^{-1}\{\boldsymbol{\xi}_{k,\mathcal{A}} - \boldsymbol{\eta}_k(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}) + \boldsymbol{r}_k\}]$$

By Taylor expansion,

$$\begin{aligned} b''(\boldsymbol{x}_j T\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}) &= b''(\boldsymbol{x}_j^T\boldsymbol{\beta}_{\mathcal{A}}^0) + b'''(\boldsymbol{x}_j^T\boldsymbol{\delta})\boldsymbol{x}_j^T(\boldsymbol{\delta}-\boldsymbol{\beta}_{\mathcal{A}}^0) \\ &\overset{\mathrm{d}}{=} b''(\boldsymbol{x}_j^T\boldsymbol{\beta}_{\mathcal{A}}^0) + d_{kj}, \end{aligned}$$

where $\boldsymbol{\delta}$ is a s-dimensional vector on the segment between $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}$ and $\boldsymbol{\beta}_{\mathcal{A}}^0$.

Written in matrix format

$$\hat{\boldsymbol{S}}_k = \boldsymbol{X}_{k,\mathcal{A}}\Sigma(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}} = \boldsymbol{X}_{k,\mathcal{A}}\{\Sigma(\boldsymbol{\theta}_k^0) + D_k\}\boldsymbol{X}_{k,\mathcal{A}},$$

where $D_k = \mathrm{diag}(d_{kj})$. We have

$$\begin{aligned} & \hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0 \\ = & \{\sum_{k=1}^K \boldsymbol{X}_{k,\mathcal{A}}^T\{\Sigma(\boldsymbol{\theta}_k^0) + D_k\}\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}[\sum_{k=1}^K\{I + \boldsymbol{X}_{k,\mathcal{A}}^T D_k\boldsymbol{X}_{k,\mathcal{A}}(\boldsymbol{S}_k^0)^{-1}\}(\boldsymbol{\xi}_{k,\mathcal{A}} - \boldsymbol{\eta}_k(\boldsymbol{\delta}) - \boldsymbol{r}_k)] \end{aligned}$$

Therefore, $\|\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty = O(b_{s,K}v_{n,K}/n + b_{s,K}\lambda_k\rho'(\beta_*/2) + b_{s,K}\beta_*^2 s) = o(\beta_*)$.

We still need to verify (2.4), i.e. show $\|\boldsymbol{z}_k\|_\infty < \rho'(0+)$, where $\boldsymbol{z}_k = (n_k\lambda_k)^{-1}\{\boldsymbol{\xi}_{k,\mathcal{A}^c} - [\boldsymbol{\gamma}_{k,\mathcal{A}^c}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}) - \boldsymbol{\gamma}_{k,\mathcal{A}^c}(\boldsymbol{\beta}^0_{\mathcal{A}})]\}$.

By Taylor expansion,

$$\begin{aligned}
& \boldsymbol{\gamma}_{k,\mathcal{A}^c}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}) - \boldsymbol{\gamma}_{k,\mathcal{A}^c}(\boldsymbol{\beta}^0_{\mathcal{A}}) \\
= \;& \boldsymbol{X}^T_{k,\mathcal{A}^c}\Sigma(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) + \boldsymbol{w}_k \\
= \;& \boldsymbol{X}^T_{k,\mathcal{A}^c}\Sigma(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}\{\boldsymbol{X}^T_{\mathcal{A}}\Sigma(\boldsymbol{\theta}^0)\boldsymbol{X}_{\mathcal{A}}\}^{-1}(\boldsymbol{\xi}_{k,\mathcal{A}} - \boldsymbol{\eta}_k(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}) - \boldsymbol{r}_k) + \boldsymbol{w}_k,
\end{aligned}$$

where $\boldsymbol{w}_k = (w_{k,s+1}, \ldots, w_{kp})$ and $w_{kj} = (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}})^T\nabla^2\gamma_{kj}(\boldsymbol{\delta}_j)(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}})$.

Similar to $\boldsymbol{r}_k$, $\|\boldsymbol{w}_k\|_\infty = O(n_k s\beta_*^2)$. Then, by condition A1, A3 and under event E_{2k},

$$\begin{aligned}
\|\boldsymbol{z}_k\|_\infty &= (n_k\lambda_k)^{-1}\|\boldsymbol{\xi}_{k,\mathcal{A}^c}\|_\infty + (\lambda_k n_k)^{-1}O(n^\alpha v_{n,K}/K + n^\alpha n_k s\beta_*^2) + C\rho'(0+) \\
&= o(1) + C\rho'(0+).
\end{aligned}$$

The

$$\begin{aligned}
& \mathrm{P}\{\cap_{k=1}^K(E_{1k}\cap E_{2k})\} \\
\geq\; & 1 - \sum_{k=1}^K \mathrm{P}(E^c_{1k}) - \sum_{k=1}^K \mathrm{P}(E^c_{2k}) \\
\geq\; & 1 - \sum_{k=1}^K\sum_{j=1}^s \mathrm{P}(|\xi_{kj}| > c_1^{-1/2}v_{n,K}/K) - \sum_{k=1}^K\sum_{j=s+1}^p \mathrm{P}(|\xi_{kj}| > c_1^{-1/2}u_{n,K}/K) \\
\geq\; & 1 - 2Ks\exp\{-v^2_{n,K}/(nK)\} - 2K(p-s)\exp\{-u^2_{n,K}/(nK)\}.
\end{aligned}$$

□

Proof of Theorem 2.2:

We first prove part (i). Similar to theorem 2.1, if $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}^c} = 0$, $k = 1, \ldots, K$, then $\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}^c} = 0$.

The first step is to show that $\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_2 = O_p(\sqrt{s/n_k})$. For each k, define event $F_k = \{\bar{Q}_k(\boldsymbol{\beta}^0_{\mathcal{A}}) > \max_{\boldsymbol{\delta}\in\partial\mathcal{N}_\tau} \bar{Q}_k(\boldsymbol{\delta})\}$, where $\bar{Q}_k(\boldsymbol{\delta}) = \ell(\boldsymbol{\delta}; \boldsymbol{y}_k, \boldsymbol{X}_{k,\mathcal{A}}) - \rho(\boldsymbol{\delta}; \lambda_k)$ is the penalized likelihood constrained on the subspace that $\{\boldsymbol{\beta} : \boldsymbol{\beta}_{\mathcal{A}^c} = 0\}$ and $\mathcal{N}_\tau = \{\boldsymbol{\delta} : \|\boldsymbol{\delta} - \boldsymbol{\beta}^0_{\mathcal{A}}\| \le \sqrt{s/n_k}\tau\}$.

Since $\sqrt{s/n_k}/\beta_* = o(1)$, $\beta_*/2 > \sqrt{s/n_k}\tau$ when n_k is large enough. Thus $\boldsymbol{\delta} \in \mathcal{N}_\tau$ will have $\text{sgn}(\boldsymbol{\delta}) = \text{sgn}(\boldsymbol{\beta}^0_{\mathcal{A}})$. By Taylor expansion, we have

$$\bar{Q}_k(\boldsymbol{\delta}) - \bar{Q}_k(\boldsymbol{\beta}^0_{\mathcal{A}}) = (\boldsymbol{\delta} - \boldsymbol{\beta}^0_{\mathcal{A}})^T \boldsymbol{v}_k - (\boldsymbol{\delta} - \boldsymbol{\beta}^0_{\mathcal{A}})^T V_k (\boldsymbol{\delta} - \boldsymbol{\beta}^0_{\mathcal{A}}),$$

where $\boldsymbol{v}_k = n_k^{-1} \boldsymbol{X}^T_{k,\mathcal{A}}[\boldsymbol{y}_k - \mu(\boldsymbol{\theta}^0_k)] - \bar{\rho}(\boldsymbol{\beta}^0_{\mathcal{A}}; \lambda_k)$ and $V_k = n_k^{-1} \boldsymbol{X}^T_{k,\mathcal{A}} \Sigma(\boldsymbol{\theta}^*_k) \boldsymbol{X}_{k,\mathcal{A}} + \text{diag}(\rho''(\boldsymbol{\beta}^*_k; \lambda_k))$ with $\boldsymbol{\theta}^*_k = \boldsymbol{X}_{k,\mathcal{A}} \boldsymbol{\beta}^*_k$ and $\boldsymbol{\beta}^*_k$ being a vector on the segment joining $\boldsymbol{\delta}$ and $\boldsymbol{\beta}^0_{\mathcal{A}}$.

By condition A4, we have $\lambda_{\min}(V_k) \ge c - \lambda_k \kappa_0 \ge c/2$. Therefore,

$$\max_{\boldsymbol{\delta}\in\partial\mathcal{N}_\tau} \bar{Q}_k(\boldsymbol{\delta}) - \bar{Q}_k(\boldsymbol{\beta}^0_{\mathcal{A}}) = \sqrt{s/n_k}\tau(\|\boldsymbol{v}_k\|_2 - c\sqrt{s/n_k}\tau/4),$$

and by condition A5, $\|\bar{\rho}(\boldsymbol{\beta}^0_{\mathcal{A}}; \lambda_k)\|_2 \le \sqrt{s}\rho'(\beta_*/2; \lambda_k) = O(n_k^{-1})$. Thus,

$$\text{P}(F_k) \ge \text{P}(\|\boldsymbol{v}_k\|_2^2 < c^2 s\tau^2/(16 n_k)) \ge 1 - 16 n_k E\|\boldsymbol{v}_k\|_2^2/(c^2 s\tau^2) = 1 - O(\tau^{-2}),$$

since $E\|\boldsymbol{v}_k\|_2^2 \le n_k^{-2}\phi tr(\boldsymbol{X}^T_{k,\mathcal{A}} \Sigma(\boldsymbol{\theta}^0_k) \boldsymbol{X}_{k,\mathcal{A}}) + \|\bar{\rho}(\boldsymbol{\beta}^0_{\mathcal{A}}; \lambda_k)\|_2^2 \le n_k^{-2}\phi tr(\boldsymbol{X}^T_{k,\mathcal{A}} \Sigma(\boldsymbol{\theta}^0_k) \boldsymbol{X}_{k,\mathcal{A}}) + s\rho'(\beta_*/2; \lambda_k) = O(s n_k^{-1})$.

Then, constrained on the subspace $\{\boldsymbol{\beta} : \boldsymbol{\beta}_{\mathcal{B}} = 0\}$, we take Taylor expansion of the penalized likelihood function at $\boldsymbol{\beta}^0_{\mathcal{A}}$. Since $\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}$ is local maximum and $\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_2 = O_p(\sqrt{s/n_k})$,

$$\boldsymbol{X}'_{k,\mathcal{A}}[\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)] - \boldsymbol{X}'_{k,\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k) \boldsymbol{X}_{k,\mathcal{A}}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) - \bar{\rho}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}; \lambda_k) + O_p(s^{3/2} n_k^{-1}) = 0 \quad (2.7)$$

Since $s = O(n_k^{1/3})$ and $\rho'(\beta_*/2; \lambda_k) = o(s^{-1/2} n_k^{-1/2})$, (2.7) gives

$$\{\boldsymbol{X}'_{k,\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k) \boldsymbol{X}_{k,\mathcal{A}}\}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) = \boldsymbol{X}'_{k,\mathcal{A}}[\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)] + O_p(\sqrt{n_k}).$$

Therefore,

$$\begin{aligned} & \{\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}\}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) \\ = \; & \{\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}\}\{\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}[\boldsymbol{X}'_{k,\mathcal{A}}\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)\} + O_p(\sqrt{n_k})]. \end{aligned}$$

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

$$\begin{aligned} & [\sum_{k=1}^{K}\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}](\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) \\ = \; & \sum_{k=1}^{K}\{\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}\}\{\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}[\boldsymbol{X}'_{k,\mathcal{A}}\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)\} + O_p(\sqrt{n/K})] \\ = \; & \sum_{k=1}^{K}\{\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}}\}\{\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}\}^{-1}[\boldsymbol{X}'_{k,\mathcal{A}}\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)\}] + O_p(\sqrt{nK}). \end{aligned}$$

Since $\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_2 = O(\sqrt{s/n_k})$, $\boldsymbol{X}'_{k,\mathcal{A}}\Sigma(\hat{\boldsymbol{\theta}}_k)\boldsymbol{X}_{k,\mathcal{A}} \xrightarrow{\text{P}} \boldsymbol{X}'_{k,\mathcal{A}}\Sigma(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}}$. Therefore,

$$\begin{aligned} & \{\sum_{k=1}^{K}\boldsymbol{X}'_{k,\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0_k)\boldsymbol{X}_{k,\mathcal{A}} + o_p(1)\}(\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) \\ = \; & \sum_{k=1}^{K}(1 + o_p(1))[\boldsymbol{X}'_{k,\mathcal{A}}\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)\}] + O_p(\sqrt{nK}). \end{aligned}$$

The above equation is equivalent to

$$\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}} = \{\boldsymbol{X}'_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)\boldsymbol{X}_{\mathcal{A}} + o_p(1)\}^{-1}\sum_{k=1}^{K}(1 + o_p(1))[\boldsymbol{X}'_{k,\mathcal{A}}\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)\}] + o_p(\sqrt{K/n}).$$

and

$$\begin{aligned} & \|\hat{\boldsymbol{\beta}}^{(c)}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_2 \\ = \; & \|\{\boldsymbol{X}'_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^0)\boldsymbol{X}_{\mathcal{A}} + o_p(1)\}^{-1}\sum_{k=1}^{K}(1 + o_p(1))[\boldsymbol{X}'_{k,\mathcal{A}}\{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}^0_k)\}]\|_2 + o_p(\sqrt{K/n}) \\ = \; & O_p(\sqrt{s/n}) + o_p(\sqrt{K/n}) = O_p(\sqrt{s/n}). \end{aligned}$$

The last step is to show $\|\boldsymbol{z}_k\|_\infty < \rho'(0+)$, where $\boldsymbol{z}_k = (n_k\lambda_k)^{-1}\{\boldsymbol{\xi}_{k,\mathcal{A}^c} - [\boldsymbol{\gamma}_{k,\mathcal{A}^c}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}) - \boldsymbol{\gamma}_{k,\mathcal{A}^c}(\boldsymbol{\beta}^0_{\mathcal{A}})]\}$. Consider event $E_{2k} = \{\|\boldsymbol{\xi}_{k,\mathcal{A}^c}\|_\infty \le c_1^{-1/2}u_{n,K}/K\}$, where $\boldsymbol{\xi}_k = \boldsymbol{X}_k^T\boldsymbol{y}_k -$

$\boldsymbol{X}_k^T \mu(\boldsymbol{\theta}_k^0)$. By Taylor expansion,

$$\begin{aligned} & \boldsymbol{\gamma}_{k,\mathcal{A}^c}(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}) - \boldsymbol{\gamma}_{k,\mathcal{A}^c}(\boldsymbol{\beta}_{\mathcal{A}}^0) \\ = \;& \boldsymbol{X}_{k,\mathcal{A}^c}^T \Sigma(\boldsymbol{\theta}_k^0) \boldsymbol{X}_{k,\mathcal{A}} (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0) + \boldsymbol{w}_k \\ = \;& \boldsymbol{X}_{k,\mathcal{A}^c}^T \Sigma(\boldsymbol{\theta}_k^0) \boldsymbol{X}_{k,\mathcal{A}} \{\boldsymbol{X}_{\mathcal{A}}^T \Sigma(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}\}^{-1} (\boldsymbol{\xi}_{k,\mathcal{A}} - \boldsymbol{\eta}_k(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}}) - \boldsymbol{r}_k) + \boldsymbol{w}_k, \end{aligned}$$

where $\boldsymbol{w}_k = (w_{k,s+1}, \ldots, w_{kp})$ and $w_{kj} = (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0)^T \nabla^2 \gamma_{kj}(\boldsymbol{\delta}_j)(\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0)$.

Then, by condition A4, A5 and under event E_{2k},

$$\begin{aligned} \|\boldsymbol{z}_k\|_\infty &= (n_k\lambda_k)^{-1}[\|\boldsymbol{\xi}_{k,\mathcal{A}^c}\|_\infty + \|\boldsymbol{X}_{k,\mathcal{A}^c}^T \mu(\hat{\boldsymbol{\theta}}_k) - \boldsymbol{X}_{k,\mathcal{A}^c}^T \mu(\boldsymbol{\theta}_k^0)\|_\infty] \\ &\leq (n_k\lambda_k)^{-1}\|\boldsymbol{\xi}_{k,\mathcal{A}^c}\|_\infty + (n_k\lambda_k)^{-1}[O(n_k)\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2 + O(n_k)\|\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_2^2] \\ &= o(1), \end{aligned}$$

and

$$\begin{aligned} & \mathrm{P}\{\cap_{k=1}^K E_{2k}\} \\ \geq \;& 1 - \sum_{k=1}^K \mathrm{P}(E_{2k}^c) \\ \geq \;& 1 - 2K(p-s)\exp\{-u_{n,K}^2/(nK)\} \longrightarrow 1. \end{aligned}$$

Now we will prove part (ii). Since $\rho'(\beta_*/2; \lambda_k) = o(s_n^{-1/2} n_k^{-1/2} K^{-1/2})$ and $s = o(n_k^{1/3}/K^{1/3})$, (2.7) gives

$$\boldsymbol{X}'_{k,\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0) \boldsymbol{X}_{k,\mathcal{A}} (\hat{\boldsymbol{\beta}}_{k,\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0) = \boldsymbol{X}'_{k,\mathcal{A}} [\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)] + o_p(\sqrt{n_k/K}).$$

Similar to the proof of part (i), we have

$$\begin{aligned} & \{\sum_{k=1}^K \boldsymbol{X}'_{k,\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0) \boldsymbol{X}_{k,\mathcal{A}} + o_p(1)\}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) \\ = \;& \sum_{k=1}^K \{1 + o_p(1)\}[\boldsymbol{X}'_{k,\mathcal{A}} \{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\}] + o_p(\sqrt{n}). \end{aligned}$$

Therefore,

$$\begin{aligned} & \{\sum_{k=1}^{K} \boldsymbol{X}'_{k,\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}_k^0) \boldsymbol{X}_{k,\mathcal{A}} + o_p(1)\}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) \\ = & \sum_{k=1}^{K} \{1 + o_p(1)\}[\boldsymbol{X}'_{k,\mathcal{A}} \{\boldsymbol{y}_k - \boldsymbol{\mu}(\boldsymbol{\theta}_k^0)\}] + o_p(\sqrt{n}). \end{aligned}$$

The above equation is equivalent to

$$\{\boldsymbol{X}'_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}} + o_p(1)\}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) = \{1 + o_p(1)\}[\boldsymbol{X}_{\mathcal{A}} \{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}^0)\}] + o_p(\sqrt{n}).$$

Thus,

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0 = \{\boldsymbol{X}'_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}} + o_p(1)\}^{-1} \{1 + o_p(1)\}[\boldsymbol{X}_{\mathcal{A}} \{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}^0)\}] + o_p(1/\sqrt{n}).$$

In addition,

$$\begin{aligned} & \boldsymbol{D}[\boldsymbol{X}_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}]^{1/2} (\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{(c)} - \boldsymbol{\beta}_{\mathcal{A}}^0) \\ = & \boldsymbol{D}[\boldsymbol{X}_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}]^{1/2} \{\boldsymbol{X}'_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}} + o_p(1)\}^{-1} \{1 + o_p(1)\}[\boldsymbol{X}_{\mathcal{A}} \{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}^0)\}] + o_p(1), \end{aligned}$$

and from condition A6, we have

$$\boldsymbol{D}[\boldsymbol{X}_{\mathcal{A}} \boldsymbol{\Sigma}(\boldsymbol{\theta}^0) \boldsymbol{X}_{\mathcal{A}}]^{-1/2} \boldsymbol{X}_{\mathcal{A}} [\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}^0)] \xrightarrow{\mathrm{D}} N(\boldsymbol{0}, \phi \boldsymbol{G}).$$

This complete the proof. □

Proof of Theorem 2.3:

We fist show that $\mathrm{P}(j \in \hat{\mathcal{A}}_k) \leq \bar{s}_k/p$, $j \in \mathcal{B}$, and $\mathrm{P}(j \in \hat{\mathcal{A}}_k) \geq \bar{s}_k/p$, $j \in \mathcal{A}$, $k = 1, \ldots, K$.

Because $E(|\mathcal{B} \cap \hat{\mathcal{A}}_k|) = E(|\hat{\mathcal{A}}_k|) - E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|) = \bar{s}_k - E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|)$ and $E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|)/E(|\mathcal{B} \cap \hat{\mathcal{A}}_k|) \geq |\mathcal{A}|/|\mathcal{B}|$, we have $E(|\mathcal{B} \cap \hat{\mathcal{A}}_k|) \leq \bar{s}_k/(1 + |\mathcal{A}|/|\mathcal{B}|)$ and $E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|) \geq \bar{s}_k/(1 + |\mathcal{B}|/|\mathcal{A}|)$. Therefore, $E(|\mathcal{B} \cap \hat{\mathcal{A}}_k|) \leq \bar{s}_k |\mathcal{B}|/p$ and $E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|) \geq \bar{s}_k |\mathcal{A}|/p$.

Using the exchangeability assumption, $\mathrm{P}(j \in \hat{\mathcal{A}}_k) = E(|\mathcal{B} \cap \hat{\mathcal{A}}_k|)/|\mathcal{B}|$, $j \in \mathcal{B}$ and $\mathrm{P}(j \in \hat{\mathcal{A}}_k) = E(|\mathcal{A} \cap \hat{\mathcal{A}}_k|)/|\mathcal{A}|$, $j \in \mathcal{A}$. Therefore, $\mathrm{P}(j \in \hat{\mathcal{A}}_k) \leq \bar{s}_k/p \leq s^*/p$, $j \in \mathcal{B}$ and $\mathrm{P}(j \in \hat{\mathcal{A}}_k) \geq \bar{s}_k/p \geq s_*/p$, $j \in \mathcal{A}$.

Since the observations in each subset are independent and $w \geq s^*K/p - 1$, $\mathrm{P}(j \in \hat{\mathcal{A}}^{(c)}) \leq 1 - F(w|K, s^*/p), j \in \mathcal{B}$ and $\mathrm{P}(j \in \hat{\mathcal{A}}^{(c)}) \geq 1 - F(w|K, s_*/p), j \in \mathcal{A}$. Therefore, $E(|\mathcal{B} \cap \hat{\mathcal{A}}^{(c)}|) = \sum_{j \in \mathcal{B}} \mathrm{P}(j \in \hat{\mathcal{A}}^{(c)}) \leq |\mathcal{B}|\{1 - F(w|K, s^*/p)\}$ and $E(|\mathcal{A} \cap \hat{\mathcal{A}}^{(c)}|) = \sum_{j \in \mathcal{A}} \mathrm{P}(j \in \hat{\mathcal{A}}^{(c)}) \geq |\mathcal{A}|(1 - F(w|K, s_*/p))$. □

Proof of Lemma 2.1:

We first state the LARS algorithm for LASSO here:

- Initialize, let the active set $A = \emptyset$, current estimation $\hat{\mu}_A = 0$ and current coefficient $\hat{\boldsymbol{\beta}}_A = 0$. $\boldsymbol{a} = 0$, $\gamma = 0$.

- Repeat the following steps until $|A| = n$.

  1. Calculate the correlation between variables and the current residual

  $$\hat{\boldsymbol{c}} = \boldsymbol{X}'_{A^c}\boldsymbol{y} - \gamma \boldsymbol{a} \quad \hat{C} = \max\{|\hat{c}_j|\}$$

  2. Let $A = \{j : |\hat{c}_j| = \hat{C}\}$ if $A = \emptyset$, $s_j = \mathrm{sgn}(\hat{c}_j)$ and $\boldsymbol{X}_A = (\dots, s_j\boldsymbol{x}_j, \dots)$, $j \in A$. Calculate the next moving direction $G_A = \boldsymbol{X}'_A\boldsymbol{X}_A$, $Q_A = (\mathbf{1}'_A G'_A \mathbf{1}_A)^{-1/2}$ and $\boldsymbol{w}_A = Q_A G_A^{-1}\boldsymbol{e}_A$, $\boldsymbol{u}_A = \boldsymbol{X}_A\boldsymbol{w}_A$.

  3. Calculate the size of tuning parameter. Let $\hat{d}_j = s_j w_j$, $j \in A$ and $\boldsymbol{a} = \boldsymbol{X}'_{A^c}\boldsymbol{u}_A$. Calculate

  $$\gamma_j = -\hat{\beta}_j/\hat{d}_j, \quad \tilde{\gamma} = \min_{\gamma_j > 0}(\gamma_j).$$

  and

  $$\hat{\gamma} = \min_{j \in A^c}{}^{+}\{(\hat{C} - \hat{c}_j)/(Q_A - a_j), (\hat{C} - \hat{c}_j)/(Q_A + a_j)\},$$

where $\min^+$ means theta the minimum is taken over only positive components.

4. If $\tilde{\gamma} \leq \hat{\gamma}$, update $\hat{\mu} \leftarrow \hat{\mu} + \tilde{\gamma}\boldsymbol{u}_A$, $A \leftarrow A - \tilde{j}$ where $\tilde{j}$ is the index for which the minimizing index in obtaining $\tilde{\gamma}$, and $\gamma = \tilde{\gamma}$. If $\tilde{\gamma} > \hat{\gamma}$, update $\hat{\mu} \leftarrow \hat{\mu} + \hat{\gamma}\boldsymbol{u}_A$, $A \leftarrow A + \tilde{j}$ where $\tilde{j}$ is the index for which the minimizing index in obtaining $\hat{\gamma}$ and $\gamma = \hat{\gamma}$.

Denote $comp(i)$ the computing steps at step i in each loop. Suppose linear search is used to find the maximum or minimum and schoolbook matrix multiplication algorithm is applied. We have $comp(1) = 2n(p - |A|)$.

In step 2, computing Q_A requires $|A|^2$ computing steps. When compute G_A^{-1}, Cholesky factorization is applied to update the inverse matrix. Details are given below. Get the block representation of G_A, the Cholesky factor of G_A, denoted by U and the inverse matrix of U $Y = U^{-1}$:

$$G_A = \begin{pmatrix} G_{11} & G_{12} \\ G_{12}' & G_{22} \end{pmatrix}, U = \begin{pmatrix} U_{11} & 0 \\ U_{12}' & U_{22} \end{pmatrix}, \text{ and } Y = \begin{pmatrix} Y_{11} & 0 \\ Y_{12}' & Y_{22} \end{pmatrix},$$

where $G_A = U^T U$ and G_{22} is a one-dimension matrix (a number) representing the newly added variable. Thus,

$$G_A^{-1} = \begin{pmatrix} Y_{11}'Y_{11} + Y_{12}Y_{12}' & Y_{12}Y_{22}' \\ Y_{22}'Y_{12}' & Y_{22}'Y_{22} \end{pmatrix},$$

where $G_{11}^{-1} = Y_{11}Y_{11}'$.

Since U_{11} and Y_{11} is known from the previous loop, we can update G_A^{-1} by the following equations: $U_{12} = Y_{11}'G_{12}, U_{22} = \sqrt{G_{22} - U_{12}'U_{12}}, Y_{22} = U_{22}^{-1}, Y_{12} = -Y_{11}U_{12}Y_{22}$, and compute $G_{11}^{-1} + Y_{12}Y_{12}'$, $Y_{12}Y_{22}'$ and $Y_{22}Y_{22}'$.

Thus,

$$comp(2) = 8|A|^2 - 10|A| + 7 + (2|A| - 1)n.$$

In step 3 and 4, we have $comp(3) = |A| + (2n-1)(p-|A|) + 2|A| + 7(p-|A|)$, and $comp(4) = 2|A| + 1$.

In all, one loop in LARS algorithm requires $8|A|^2 - 11|A| + (4n+6)p - 2n|A| + 8 - n$. Therefore, since $p \geq n$, at most n variables will be fitted and the LARS algorithm requires at least

$$\begin{aligned} & \sum_{|A|=1}^{n} 8|A|^2 - 11|A| + (4n+6)p - 2n|A| + 8 - n \\ = \quad & 5n^3/3 + 23n/6 + 4n^2(p - 7/8) + 6np. \end{aligned}$$

Each time dropping variable occurs, it will add additional $8|A|^2 - 11|A| + (4n+6)p - 2n|A| + 8 - n$ computing steps depending on the number of current active variables. The worst case would be $6n^2 + 4n(p-3) + 6p + 8$ computing steps each time and the solution path has n times downsize. The computing steps for the worst case would be $23n^3/3 + 71n/6 + 8n^2(p - 31/16) + 12np$.

As a result, as each sub-sample has n_k observations, for the best case, the computing steps for the combined estimator is $\sum_{k=1}^{K} 5n_k^3/3 + 23n_k/6 + 4n_k^2(p - 7/8) + 6n_k p$. Since $\sum_{k=1}^{K} n_k = n$, $5n^3/3 + 23n/6 + 4n^2(p - 7/8) + 6np \geq \sum_{k=1}^{K} 5n_k^3/3 + 23n_k/6 + 4n_k^2(p - 7/8) + 6n_k p$. The result follows immediately. Similarly, the combined estimator requires less computing steps for the worst case. □

Proof of Theorem 2.4:

We only need to show that under the assumptions, the worst case for split-and-conquer approach requires less computing steps than the best case for LARS algorithm

using the entire dataset. When $n_k = O(n_k)$, split-and-conquer approach requires at most $23n^3/(3K^2) + 71n/6 + 8n^2(p - 31/16)/K + 12np$ computing steps and LARS algorithm using the entire dataset needs at least $5n^3/3 + 23n/6 + 4n^2(p - 7/8) + 6np$ computing steps. It is equivalent to show that

$$\begin{aligned} & \{5n^3/3 + 23n/6 + 4n^2(p - 7/8) + 6np\} \\ & -\{23n^3/(3K^2) + 71n/6 + 8n^2(p - 31/16)/K + 12np\} \\ = \quad & (5 - 23/K^2)n^3/3 + \{4p(1 - 2/K) + (31/K - 7)/2\}n^2 - (8 + 6p)n \geq 0. \end{aligned}$$

When $K \geq 3$ and $p \geq 2$, we have $5-23/K^2 > 0$ and $4p(1-2/K)+(31/K-7)/2 > 0$. Thus, when $n \geq 4(4 + 3p)/\{1 + 8p(1 - 2/K) + 31/K - 7\}$, we have $(5 - 23/K^2)n^2/3 + \{4p(1 - 2/K) + (31/K - 7)/2\}n - (8 + 6p) > 0$. The result follows immediately. □

Chapter 3

Model Selection Consistency of OSCAR estimators

3.1 Introduction

In this chapter, we consider linear regression models:

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

where $\boldsymbol{y}$ is a $n \times 1$ vector, $\boldsymbol{X}$ is a $n \times p$ matrix and $\boldsymbol{\beta}$ are parameters; ε is a $n \times 1$ vector of errors. For simplicity, each explanatory variable is normalized, that is, $\sum_{i=1}^{n} x_{ij} = 0$ and $\|\boldsymbol{x}_j\|_2^2 = n, j = 1, \ldots, p$.

A substantial work has been proposed in the formulation of penalized least squares:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + p(\boldsymbol{\beta}),$$

where $p(\cdot)$ is a penalty function. With different choices of the penalty function $\rho(\cdot)$, we have various penalized least square estimators. For example, ridge regression proposed by Frank & Friedman (1993) is based on L_2 penalty function. Ridge estimator is a shrinkage method that can handle the low-rank design matrix situation but it does not lead to sparse coefficients that have many exact zero elements. On the contrary, Tibshirani (1996), Chen et al. (2001), Efron et al. (2004) consider L_1 penalty in Lasso or least angle regression (LARS). The Lasso estimators can result in exact zero estimation and thus perform both shrinkage and model selection. More recently, smoothly clipped absolute deviation (SCAD) penalty by Fan & Li (2001) and minimax concave

penalty (MCP) by Zhang (2010) extend the class of penalty functions broadly. The SCAD penalty corresponds to a nonconcave quadratic spline function with two knots. SCAD estimator retains the sparseness of Lasso estimators and are unbiased for large coefficients. As the SCAD penalty, the MCP also has the property of unbiasedness but releases the computational and analytical burden resulted from the non-convexity in the minimization problem in SCAD or SCAD-like penalties.

However, the penalized estimators mentioned above are not able to handle the case when high correlations exist among variables. Zou & Hastie (2005) have found that Lasso estimators do not encourage highly corrected variables to be selected all together but randomly select one of these variables. In practice, ignoring correlation structures among predictor can lead to wrongly interpreted models. For example, Chen et al. (2009) integrate genotype and gene expression data to predict complex quantitative phenotypes and identify genes that actively influence these traits. Since coregulated or functionally related genes are more likely to have similar expression patterns, gene expression data is also useful to find coexpression patterns and locate groups of co-transcribed genes (Lee et al., 2004). Therefore, a good estimator should discover all possible influential genes to discover gene functions and to find related transcription factors instead of selecting one of these highly correlated genes. After identifying features that actively influence the phenotype, further causality tests can be done to discover causal genes (Chen et al., 2009). In addition, selecting all highly correlated variables helps in data analysis problems with confounding variables. A confounding variable is defined as a variable that is correlated with both selected explanatory variables and the response variable. Without the risk of selecting wrong models, a conservative approach is to include all possible variables in the model for further causality tests.

The conservative approach of including highly correlated variables can be realized by the 'grouping' practice in the literature in the sense that they would have similar or exactly the same coefficients, see Zou & Hastie (2005) and Bondell & Reich (2008). The grouping property is essentially important in identifying relevant explanatory variables even causal effects without hurting any prediction accuracy. Indeed, grouping is an attractive feature in dealing with highly correlated variables. Let's consider a simple regression example with two highly correlated variables:

$$\boldsymbol{y} = \beta_1^* \boldsymbol{x}_1 + \beta_2^* \boldsymbol{x}_2 + \varepsilon, \tag{3.1}$$

where $\boldsymbol{x}_1^T \boldsymbol{x}_2 / n = 1$ and β_i^*, $i = 1, 2$ are true coefficients. If $\boldsymbol{x}_1$ is the causal effect and $\boldsymbol{x}_2$ is a confounding variable, we have $\beta_1^* \neq 0$ and $\beta_2^* = 0$. Conservatively, we would include both variables in the model and enforce their coefficients to be the same:

$$\boldsymbol{y} = \beta_1^0 \boldsymbol{x}_1 + \beta_2^0 \boldsymbol{x}_2 + \varepsilon, \tag{3.2}$$

where $\beta_1^0 = \beta_2^0 = (\beta_1^* + \beta_2^*)/2$. Model (3.1) and (3.2) are equivalent in prediction but model (3.2) is not at the risk of selecting wrong models. In the context of grouping highly correlated variables, we will consider model (3.2) as more appropriate target model. Inspired by this model, suppose the true model is

$$\boldsymbol{y} = \sum_{i=1}^{p} \boldsymbol{x}_i \beta_i^* + \boldsymbol{\varepsilon},$$

where β_i^* is the true coefficients which reflects causal effects. We reparameterize the model as following:

$$\boldsymbol{y} = \sum_{i=1}^{p} \boldsymbol{x}_i \beta_i^0 + \boldsymbol{\varepsilon},$$

where $|\beta_i^0| = \sum_{l \in G_i} |\beta_l^*| / |G_i|$ with $G_i = \{j : |\boldsymbol{x}_i^T \boldsymbol{x}_j / n| = 1, j = 1, \ldots, p\}$. Instead of recovering true coefficients $\boldsymbol{\beta}^* = (\beta_i^*, i = 1, \ldots, p)$, our goal is to recover $\boldsymbol{\beta}^0 = (\beta_i^0, i = 1, \ldots, p)$.

In order to achieve the grouping feature, many literatures have proposed doubly regularized estimators. Zou & Hastie (2005) propose Elastic net (Enet) penalty that combines L_1 norm and L_2 norm penalties. They prove that the Elastic net estimators have a grouping effect that is the upper bound of the difference between the coefficients of predictor i and predictor j is proportional to $\sqrt{1-|\rho_{ij}|}$, where $\rho_{ij} = \boldsymbol{x}_i'\boldsymbol{x}_j/n$ is the sample correlation. As a result, elastic net estimators have similar coefficients for highly correlated predictors because the difference between the coefficients would be small when the sample correlation is close to 1. Huang et al. (2010a) extend the Enet to Mnet penalty in which the L_1 norm penalty is replaced by MCP. Li & Li (2008) and Li & Li (2010) propose a graph-constrained estimation (Grace) procedure. The Grace penalty is a combination of L_1 norm penalty and a Laplacian quadratic penalty which is associated with the graph structure. Huang et al. (2011) propose sparse Laplacian shrinkage (SLS) estimation method in which the penalty function is a combination of MCP and a Laplacian quadratic penalty. The Laplacian matrix used in SLS is constructed through the adjacency measures among variables which is more general than the graph structure. They show that MCP can induce sparseness in $\boldsymbol{\beta}$ and the quadratic function of $\boldsymbol{\beta}$ encourages smoothness in the estimation. It is worth noting that all the methods mentioned above use quadratic penalty functions to enjoy the grouping property. More precisely, all these penalized estimators with grouping property can be written as:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_1 p_1(\boldsymbol{\beta}) + \lambda_2 \sum_{j<k} |a_{jk}|(\beta_j - s_{jk}\beta_k)^2,$$

where $p_1(\cdot)$ is a penalty that can introduces sparseness, e.g. L_1 norm penalty and MCP; and a_{jk}, s_{jk} are constants chosen by the user arbitrarily.

OSCAR estimators, proposed by Bondell & Reich (2008), further improve the

smoothing or grouping effect. OSCAR estimators are defined as penalized least squares where the penalty function combines L_1 norm and L_∞ norm penalties:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{j<k} \max\{|\beta_j|, |\beta_k|\}],$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are tuning parameters. Compared with elastic net, OSCAR estimators have an exact grouping property. In particular, the absolute coefficients of two variables are enforced to be exactly the same if the tuning parameters are larger than a bound that is proportional to $\sqrt{1 - |\rho_{ij}|}$. In other words, if $|\rho_{ij}|$ exceeds a threshold that is controlled by the tuning parameters, predictor i and predictor j will be grouped since their absolute coefficients are the same. The penalized least square function of OSCAR estimators can be rewritten as:

$$\hat{\boldsymbol{\beta}}^O = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{j<k} \big||\beta_j| - |\beta_k|\big|,$$

with properly adjusted tuning parameters λ_1 and λ_2. Compared with the penalty functions using quadratics, OSCAR estimators use L_1 norm based penalty function on the differences of the absolute coefficients rather than quadratic functions. As a result, OSCAR estimators enforce the coefficients' differences of highly correlated predictors to be exact 0.

The difficulty in studying OSCAR estimators arise in the nondifferentiation of L_1 norm penalty function on the coefficients' difference. All the other doubly regularized estimators such as Enet (adaptive elastic net), Grace and SLS estimators take advantage of the differentiable quadratic function. Thus, compared with the conditions for LASSO and MCP estimators, with minor adjustments on the design matrix, one can assess how well those penalized estimators perform in model selection. On the contrary, the consistency properties of OSCAR estimators can not be adapted from the LASSO

estimators' properties.

In this chapter, we consider a formal and more restrictive definition of sign consistency for model selection. An estimator $\hat{\boldsymbol{\beta}}$ is said to be *sign consistent* if

$$\text{sgn}(\hat{\beta}_i) = \text{sgn}(\beta_i^0), \text{sgn}(|\hat{\beta}_i| - |\hat{\beta}_j|) = \text{sgn}(|\beta_i^0| - |\beta_j^0|), \tag{3.3}$$

where $\boldsymbol{\beta}^0 = (\beta_i^0, i = 1, \ldots, p)$ is the target coefficient. Compared with the usual definition of sign consistency which requires that $\text{sgn}(\hat{\boldsymbol{\beta}}) = \text{sgn}(\boldsymbol{\beta}^0)$, the definition of (3.3) also requires that the penalized estimator should keep the magnitude order of the coefficients beside selecting the true model. Suppose a group structure exists based on variable correlations, that is $\{1, \ldots, p\} = \cup_k G_k$, where variables in the same group G_k are highly correlated and the groups G_k's are unknown non-overlapping subsets of $\{1, \ldots, p\}$. We generalize the definition of sign consistent: an estimator $\hat{\boldsymbol{\beta}}$ is said to be *group sign consistent* if

$$\text{sgn}(\hat{\beta}_i) = \text{sgn}(\beta_i^0), \text{sgn}(|\hat{\beta}_i| - |\hat{\beta}_j|) = \text{sgn}(|\beta_i^0| - |\beta_j^0|), i \in G_k, j \in G_l, k \neq l. \tag{3.4}$$

The group sign consistency means that the penalized estimator can select the true model and keep the magnitude order of the coefficients in different groups. Since penalized estimators are asymptotically unbiased, high correlation between two variables actually implies closer target coefficients. Here, we provide a distinguishable condition on the coefficients and show that OSCAR estimators are sign consistent with high probability when correlations are moderate. When the correlations between variables with nonzero coefficients are extremely high, we show that OSCAR estimators are group sign consistent.

The rest of the chapter is organized as follows. In section 3.2, we investigate the grouping property of OSCAR estimators when the sample size n intends to infinity

and we show OSCAR estimators are (group) sign consistent. In section 3.3, simulation studies are presented with comparison of other estimators. In section 3.4, we give final discussions on OSCAR estimators.

3.2 Model selection consistency

Let the target value of the regression coefficients be $\boldsymbol{\beta}^0$. Rewrite OSCAR estimator as

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1 + \lambda\delta \sum_{j<k} \max\{|\beta_j|, |\beta_k|\},$$

where $\lambda > 0$ and $\delta > 0$ are tuning parameters.

Without loss of generality, the coefficients are ordered by their magnitudes $|\beta_1^0| \geq |\beta_2^0| \geq \cdots \geq |\beta_{p-1}^0| \geq |\beta_p^0|$. We consider model selection properties of OSCAR estimators under a sparsity condition on the regression coefficients. Denote by $A = \{j : \beta_j^0 \neq 0\}$, the set of indices of nonzero coefficients. Let $s = |A|$ be the cardinality of A. In addition, because of the reparameterization, we assume that the coefficients magnitude are associated with the variable correlations, that is, the differences between two coefficients' magnitudes are proportional to the correlation between the corresponding variables. Denote $\beta_* = \min_{j \in A} |\beta_j^0|$ as the minimal signal and $\boldsymbol{\omega} = (\omega_j, j = 1, \ldots, p)$ with $\omega_j = (p - j)\delta + 1$ be the ordered weights.

Define subspaces $\mathcal{D}_g \subset \mathbb{R}^g$, $g \in \mathbb{N}$:

$$\mathcal{D}_g = \{\boldsymbol{v} = (v_1, \ldots, v_g)^T : v_j = \sum_{l:l<j, 1\leq l\leq g} d_{lj}(j) + \sum_{l:l>j, 1\leq l\leq g} d_{jl}(j), j = 1 \ldots g, \quad (3.5)$$

$$\text{with } \; d_{ls}(l) \geq 0, d_{ls}(s) \geq 0 \;\text{ and }\; d_{ls}(l) + d_{ls}(s) = 1, \forall 1 \leq l < s \leq g\}, \quad (3.6)$$

where $d_{ls}(\cdot)$ is a function on the set $\{l, s\}$.

We characterize OSCAR estimators in the following lemma. Lemma 3.1 comes from the KKT conditions of penalized least square function for OSCAR estimators.

Lemma 3.1 *Suppose* $\hat{\alpha}_1 > \hat{\alpha}_2 > \cdots > \hat{\alpha}_{\hat{K}} > 0$ *are the distinct values of* $\{|\hat{\beta}_j| \neq 0 : j = 1, \ldots, p\}$. *Then* $\hat{\boldsymbol{\beta}}$ *is OSCAR estimator if*

(i) $-\boldsymbol{X}_{\hat{G}_k}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n + \lambda\tilde{\omega}_{\hat{G}_k} \circ \mathrm{sgn}(\hat{\boldsymbol{\beta}}_{\hat{G}_k}) = 0, k = 1, \ldots, \hat{K}$,

where $\hat{G}_k = \{j : |\hat{\beta}_j| = \hat{\alpha}_k\}$ *and* $\tilde{\omega}_{\hat{G}_k} = \delta \boldsymbol{v}_{\hat{G}_k} + [(p - \sum_{l<k} |\hat{G}_l|)\delta + 1]\mathbf{1}_{\hat{G}_k}$ *for some* $\boldsymbol{v}_{\hat{G}_k} \in \mathcal{D}_{|\hat{G}_k|}$; *and*

(ii) $\|\boldsymbol{X}_{\hat{A}^c}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n\|_\infty \leq \lambda$, *where* $\hat{A}^c = \{j : |\hat{\beta}_j| = 0\}$.

To show the model selection consistency of OSCAR estimators, we make the following sub-Gaussian assumption on the error terms:

Condition A 3.1 *The sub-Gaussian assumption is made on the error terms:*

$$\sup_{\|\boldsymbol{u}\|_2=1} \mathrm{P}(\boldsymbol{u}'\boldsymbol{\varepsilon} > \sigma t) \leq e^{-t^2/2}, \quad t > 0.$$

We consider first in Section 3.2.1 the no grouping case, in which all nonzero coefficients are distinguishable with no ties. It implies that the correlations among variables are not high. In Section 3.2.2, we provide theoretical results for more complicated case involving high correlated variables in which unknown potential ties are presented among non-zero coefficients.

3.2.1 No grouping case

We introduce the following notations: $\Sigma = \boldsymbol{X}'\boldsymbol{X}/n$ and for any set S, Σ_S denotes the submatrix of Σ with row index in S and $\boldsymbol{X}_S$ denotes the submatrix of $\boldsymbol{X}$ with column index in S.

Let us consider the case that all nonzero coefficients are distinguishable. Without loss of generality, for $j \in A$, we assume $|\beta_1^0| > |\beta_2^0| > \cdots > |\beta_s^0| > 0$. In this case, we also make a simple assumption that Σ_A is invertible. Specifically, we assume:

Condition B 3.1 *Assume the following conditions on the design matrix:*

(i) $\|\Sigma_A^{-1}\|_\infty \leq C_1$,

(ii) $\|\Sigma_A^{-1}\boldsymbol{\omega}_A \circ \mathrm{sgn}(\boldsymbol{\beta}_A^0)/n\|_\infty \leq c_1$

(iii) $\|\boldsymbol{X}_{A^c}^T\boldsymbol{X}_A\Sigma_A^{-1}\boldsymbol{\omega}_A \circ \mathrm{sgn}(\boldsymbol{\beta}_A^0)/n\|_\infty \leq c_2$,

where $C_1 > 0$, $c_1 > 0$ *and* $c_2 > 0$ *are constants.*

Condition B3.1 are the standard regularity conditions on the design matrix. Condition B3.1(i) require the cross-product matrix for variables with nonzero coefficients are invertible and the L_∞ norm of the inverse matrix is bounded. Condition B3.1(ii) controls the bias introduced by the penalty. Condition B3.1(iii) put restrictions on the correlation between variables with nonzero coefficients and variables with zero coefficients.

Denote $\boldsymbol{\xi} = \Sigma_A^{-1}\boldsymbol{\omega}_A$. We evaluate the magnitude difference between two coefficients by

$$d_j = \{(|\beta_j^0| - |\beta_{j+1}^0|) - \lambda|\xi_j - \xi_{j+1}|\}/\sqrt{v_j},$$

where $v_j = \Sigma_A^{-1}(j,j) + \Sigma_A^{-1}(j+1,j+1) - 2\Sigma_A^{-1}(j,j+1)$ and $\Sigma_A^{-1}(i,j)$ denotes the (i,j) entry of matrix Σ_A^{-1}.

Condition B 3.2 *For* $0 < \epsilon < 1$, *assume*

(i) $\beta_* - \lambda c_1 \geq C_1\sigma\sqrt{2\log(s\epsilon^{-1})/n}$,

(ii) $\lambda \geq \sigma\sqrt{2\log(p\epsilon^{-1})/n}/(1-c_1)$,

(iii) $\min_{j\in\mathscr{A}} d_j \geq \sigma\sqrt{2\log\{(s-1)\epsilon^{-1}\}/n}$.

Conditions B3.2(i)-(ii) explains the relationship between minimal signal β_* and the tuning parameters. Basically, it requires that the minimal signal should be large enough to detect and the penalty is under control. Compared with LASSO, OSCAR penalty will put additional penalty that is as large as $(p-1)\delta$ times the absolute value of coefficients. The additional bias is controlled by c_1. If c_1 is large, we need to reduce λ and have β_* to be large enough to be detected. Condition B3.2(iii) requires the differences among nonzero coefficients distinguishable so that OSCAR estimators can keep the coefficients order. The difference level is adjusted by the bias ξ and correlation between two variables. Since we need to distinguish s coefficients, the normalized difference level needs to be as large as $O(\sigma\sqrt{2\log\{(s-1)\}/n})$. In terms of the target coefficients $\boldsymbol{\beta}^0$, the difference between biases $|\xi_j - \xi_{j+1}|$ is controlled by $C_1\delta$. Therefore, the correlation between two variables determines how large the difference is needed between their corresponding target coefficients. Specifically, when the correlation among two variables is high or v_j is small, the difference between their coefficients should be larger. On the contrary, if the correlation is low or v_j is high, this constraint can be relaxed. This condition is consistent with the grouping property of OSCAR estimators. As OSCAR estimators intend to give similar coefficients to highly correlated variables, the true coefficient difference should be larger for OSCAR estimators to detect. Nevertheless, since the cross-product matrix for variables with nonzero coefficients are invertible and the L_∞ norm of the inverse matrix is bounded, the correlations among variables in this case are at most moderate.

The following theorem shows the sign consistency property of OSCAR estimators under the moderate correlation scenario.

Theorem 3.1 *Assume Condition A, B3.1 and B3.2 hold. Then, we have*

$$\mathrm{P}\big(\mathrm{sgn}(\hat{\beta}_i) = \mathrm{sgn}(\beta_i^0), \mathrm{sgn}(|\hat{\beta}_i| - |\hat{\beta}_j|) = \mathrm{sgn}(|\beta_i^0| - |\beta_j^0|)\big) \geq 1 - 3\epsilon.$$

3.2.2 Grouping with high correlations case

When the variables are highly correlated and the covariance matrix Σ_A is not invertible, that is, Condition B3.1(i) cannot be satisfied, Theorem 3.1 is not applicable. But in this case, variables with nonzero coefficients can be grouped together according to their correlations. In particular, we assume, without loss of generalization, that there are K groups $G_k = \{j : j = s_{k-1} + 1, \ldots, s_k\}$, $k = 1, \ldots, K$ where $s_0 = 0$ and $s_K = s$. The target coefficients are organized such that $|\beta_1^0| \geq \cdots \geq |\beta_{s_1}^0| > |\beta_{s_1+1}^0| \geq \cdots \geq |\beta_{s_2}^0| > \cdots > |\beta_{s_{K-1}+1}| \geq \cdots \geq |\beta_{s_K}| > 0$, and $\beta_{s+1}^0 = \cdots = \beta_p^0 = 0$. If coefficients in the same group are tied, that is, $|\beta_{s_{k-1}+1}^0| = \cdots = |\beta_{s_k}^0|$, $k = 1, \ldots, K$, we have exact group structures. Here, we consider a more general scenario that variables in the same group do not have exactly the same coefficients but close coefficients. This generalizes the simulation settings considered by Bondell & Reich (2008).

We introduce and define a representing variable z_k and its corresponding coefficient b_k^0 for each group as following:

$$\begin{aligned} \boldsymbol{z}_k &= \sum_{j=s_{k-1}+1}^{s_k} \boldsymbol{x}_j \mathrm{sgn}(\beta_j^0)/|G_k|, \\ b_k^0/|G_k| &= \mathrm{argmin}_b \sum_{j \in G_k} \big||\beta_j^0| - b\big| = median(|\beta_j^0|, j \in G_k), \end{aligned}$$

where $|G_k| = s_k - s_{k-1}$ is the number of variables in group k. Thus, $b_1^0 > b_2^0 > \cdots > b_k^0 > 0$. Specifically, $\boldsymbol{z}_k$ is the average of variables in group k with sign adjusted and b_k^0 is the median of absolute coefficients in group k. We will show that with mild conditions imposed on representing variables, OSCAR estimators achieve group sign consistency.

In fact, this property reflects the equivalence between OSCAR and methods which average variables in the same group, e.g. Jornsten & Yu (2003).

Denote $\boldsymbol{Z} = (\boldsymbol{z}_k, k = 1, \ldots, K)$ as representing design matrix, $H_{\boldsymbol{Z}} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$, $P_{\boldsymbol{Z}}^{\perp} = I - H_{\boldsymbol{Z}}$ and $D = \text{diag}(|G_k|, k = 1, \ldots, K)$. Furthermore, the discrepancy between $\boldsymbol{b}^0 = (b_k^0, k = 1, \ldots, K)$ and $\boldsymbol{\beta}_A^0$ is evaluated by $\Delta = \sum_{k=1}^{K} \sum_{j \in G_k} \boldsymbol{x}_j \text{sgn}(\beta_j^0)(|\beta_j^0| - b_k^0/|G_k|)$. Furthermore, define average penalty weights for each group $\bar{\omega}_k = \sum_{j \in G_k} \omega_j/|G_k|$ and $\bar{\boldsymbol{\omega}} = (\bar{\omega}_k, k = 1, \ldots, K)$.

Condition C 3.1 *For $0 < \alpha < 1$, $0 < u_1 < 1$ and $0 < u_2 < 1$, assume the following conditions on the representative design matrix and coefficients:*

(i) $\|(\boldsymbol{Z}'\boldsymbol{Z}D/n)^{-1}\|_\infty \leq C_1$,

(ii) $\|(\boldsymbol{Z}'\boldsymbol{Z}D/n)^{-1}\bar{\boldsymbol{\omega}}\|_\infty \leq c_1$,

(iii) $\|\boldsymbol{X}'_{\mathscr{A}^c}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}/n)^{-1}\bar{\boldsymbol{\omega}}/n\|_\infty \leq c_2$,

(iv) $\big||\beta_j^0| - b_k^0/|G_k|\big| \leq (1-\alpha)\beta_*$, $j \in G_k$, $k = 1, \ldots, K$,

(v) $\|\boldsymbol{Z}^T\Delta/n\|_\infty \leq \lambda u_1$,

(vi) $\|\boldsymbol{X}_{A^c}^T(I - H_Z)\Delta/n\|_\infty \leq \lambda u_2$,

where $C_1 > 0$, $c_1 > 0$ and $c_2 > 0$ are constants.

Condition C3.1(i)-(iii) are parallel to Condition B1 but the restrictions are put on the representative design matrix. Also, Condition C3.1(iv)-(vi) requires that the coefficients for variables in the same group are close enough. If the coefficients in one group have the same absolute value, Condition C3.1(iv)-(vi) can be ignored.

Variables in the same group need to have high correlation to be grouped together by OSCAR estimators. We evaluate the within group correlation by the closeness between the subset average and the representative variable. More specifically, define

$$\phi_k^m = \min_{|B_k^m|=m} \|\frac{P_Z^{\perp}[\boldsymbol{z}(B_k^m) - \boldsymbol{z}_k]}{\sqrt{n}}\|_2^{-1}[\lambda\delta\frac{|G_k| - m}{2} - |\frac{(\boldsymbol{z}(B_k^m) - \boldsymbol{z}_k)^T}{n}\{P_{\boldsymbol{Z}}^{\perp}\Delta + \lambda\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z}/n)^{-1}\bar{\omega}\}|],$$

where $\boldsymbol{z}(B_k^m) = \sum_{j\in B_k^m} \boldsymbol{x}_j \text{sgn}(\beta_j^0)/m$ with $B_k^m \subset G_k$ and $|B_k^m| = m$. Again, when variables in one group have the same absolute coefficients, it can be reduced to $\|P_Z^{\perp}[\boldsymbol{z}(B_k^m) - \boldsymbol{z}_k]/\sqrt{n}\|_2^{-1}[\lambda\delta(|G_k| - m)/2 - |(\boldsymbol{z}(B_k^m) - \boldsymbol{z}_k)^T\{\lambda\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z}/n)^{-1}\bar{\omega}\}|/n]$. Moreover, denote $\boldsymbol{\xi} = (\boldsymbol{Z}^T\boldsymbol{Z}/n)^{-1}(\boldsymbol{Z}^T\Delta/n - \lambda\bar{\boldsymbol{\omega}})$ and $\Sigma_{\boldsymbol{Z}}^{-1} = (\boldsymbol{Z}^T\boldsymbol{Z}/n)^{-1}$. The magnitude difference between two group coefficients is evaluated by $d_k = [(b_k^0 - b_{k+1}^0) - |\xi_k - \xi_{k+1}|]/v_k$, $k = 1, \ldots, K-1$, where $v_k = \Sigma_{\boldsymbol{Z}}^{-1}(k,k) + \Sigma_{\boldsymbol{Z}}^{-1}(k+1,k+1) - 2\Sigma_{\boldsymbol{Z}}^{-1}(k,k+1)$ and $\Sigma_{\boldsymbol{Z}}^{-1}(k,l)$ denotes the (k,l) entry of matrix $\Sigma_{\boldsymbol{Z}}^{-1}$.

Condition C 3.2 *For $0 < \epsilon < 1$, assume*

(i) $\alpha\beta_* - \lambda c_1 = C_1(\sigma\sqrt{\log(K\epsilon^{-1})/n} + \lambda u_1)$,

(ii) $\phi_k^m = 2\sigma\sqrt{2[\log(p\epsilon^{-1}) + \log(\binom{|G_k|}{m}|G_k|)]/n}$,

(iii) $\lambda = \sigma\sqrt{2\log(p\epsilon^{-1})/n}/(1 - c_2 - u_2)$,

(iv) $\min_{k=1}^{K-1} d_k = \sigma\sqrt{2\log\{(K-1)\epsilon^{-1}\}/n}$.

Condition C3.2(ii) imposes conditions on the correlations for variables within one group. Roughly speaking, we need any subset average within a group to be close to the representative variable in order overcome the within group penalty difference $\lambda\delta(|G_k| - m)/$, $k = 1, \ldots, m$. In this case, OSCAR estimates for variables in the same group will have exactly the same absolute value. Compared with Theorem 3.1 in Bondell

& Reich (2008), this condition is much less restrictive. Here, we only require that each variable to be close to the representing variable for that group instead of pairwise high correlation. Also, Condition C2(ii) implies that the larger group size is, the more likely OSCAR estimator will force coefficients in that group to have identical absolute value. Condition C3.3(i),(iii) and (iv) are parallel to Condition B3.2 with adjustment to the additional bias introduced by the within group differences.

Theorem 3.2 *Assume Condition A3.1, C3.1 and C3.2 hold. Then we have*

$$\mathrm{P}\big(\mathcal{O}_1 \cap \mathcal{O}_2\big) \geq 1 - 4\epsilon,$$

where $\mathcal{O}_1 = \{\mathrm{sgn}(\hat{\beta}_i) = \mathrm{sgn}(\beta_i^0), \mathrm{sgn}(|\hat{\beta}_i| - |\hat{\beta}_j|) = \mathrm{sgn}(|\beta_i^0| - |\beta_j^0|), k \neq l,\ k, l = 1, \ldots, K\}$ *and* $\mathcal{O}_2 = \{|\hat{\beta}_i| = |\hat{\beta}_j|,\ i, j \in G_k,\ k = 1, \ldots, K\}$.

3.3 Numerical studies

3.3.1 Simulation studies

In this section, we present numerical results to study the finite sample performance of OSCAR estimators, compared with LASSO, SCAD and Elastic Net estimators. For all methods, BIC criterion is used to select the tuning parameters.

We consider the scenarios with moderately correlated variables and highly correlated variables. In Bondell & Reich (2008), the simulation studies focus on the settings with a small number of variables, such as $p = 8$, $p = 40$ and so on. To demonstrate the large sample property of OSCAR estimators, similar to Fan & Peng (2004) and Zou & Zhang (2009), we choose $p = n^v$, where $0 < v < 1$. The response variable is generated from

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

where ε are generated from IID $N(0, \sigma^2)$.

We evaluate the performance of different penalized estimators in two aspects: model selection accuracy and prediction. We calculate mean square errors (MSE), model size, model selection (Model Sel.) sensitivity (in %) and model selection specificity (in %). Here, model size is defined as the number of variables with nonzero coefficients estimation; model selection sensitivity is calculated as the ratio of the number variables with both true nonzero coefficients and nonzero coefficients estimation and the number of variables with true nonzero coefficients; model selection specificity is calculated as the ratio of the number variables with both true zero coefficients and zero coefficients estimation and the number of variables with true zero coefficients.

Example 3.1 In this example, no grouping case is simulated. We consider two choices of n, $n = 100, 400$, and two choices of p, $p = \lfloor\sqrt{n}\rfloor, \lfloor n^{2/3}\rfloor$ with a total of four settings. The covariance between variable i^{th} and j^{th} is $\Sigma_{ij} = \cos(|i-j|\pi/(2p))$ and $\boldsymbol{X}'\boldsymbol{X}/n = \Sigma$. The number of variables with nonzero coefficients $s = \lfloor\sqrt{s}\rfloor$ and the coefficients $\beta \approx \sqrt{2 * \log(p)/n}$. The variable correlations and coefficients magnitude are demonstrated in Figure 3.1. The numerical results are exhibited in Table 3.1 and Figure 3.3. As we can see from Table 3.1, OSCAR estimators outperform other estimators in terms of higher model selection sensitivity and model selection specificity for both examples. In this example, although explanatory variables are consecutively correlated, OSCAR estimators are able to distinguish variables with zero coefficients and variables with nonzero coefficients. Figure 3.3 shows that OSCAR estimators select variables with the largest absolute coefficient with the highest frequency. The selecting frequencies decrease as the absolute coefficients become smaller. Most variables with zero coefficients are excluded as they are never been selected by OSCAR estimators. On

the other hand, ENET estimators have higher model selection sensitivity than LASSO and SCAD estimators. However, ENET estimators sometimes are disturbed by the consecutive correlation and ignore a few true variables with nonzero coefficients but select variables with zero coefficients that are highly correlated with variables with nonzero coefficients. In addition, SCAD estimators will select only one of the highly correlated variables with nonzero coefficients. LASSO estimators are influenced by the correlation structures most as they cannot distinguish variables with zero and nonzero coefficients. All variables will be randomly picked by LASSO estimators.

Example 3.2 In this example, grouping with high correlations case is simulated. We consider two choices of n, $n = 100, 400$, and two choices of p, $p = \lfloor\sqrt{n}\rfloor, \lfloor n^{2/3}\rfloor$ with a total of four settings. The design matrix has $K = 4$ groups and each group has p/K variables, where variables $\{1, \ldots, s_1\} \in G_1, \ldots, \{s_k + 1, \ldots, K\} \in G_K$. $\boldsymbol{X}'\boldsymbol{X}/n = \Sigma$ and $\Sigma_{ij} = \cos(|i-j|\pi/(2p))$, if i, j are in the same group; otherwise $\Sigma_{ij} = \cos(|k-l|+|i-j|\pi/(2K))$, $i \in G_k$, $j \in G_l$. The number of variables with nonzero coefficients $s = \lfloor\sqrt{s}\rfloor$ and the coefficients $\beta \approx \sqrt{2 * \log(p)/n}$. Group 1 and group 2 have $s/2$ variables with nonzero coefficients separately. The variable correlations within and between groups are demonstrated in Figure 3.2. The numerical results are exhibited in Table 3.1 and Figure 3.4. According to Table 3.1, highly correlated variables are grouped together and correlations between variables in different groups are small. In this case, OSCAR estimators and ENET estimators perform similarly while SCAD estimators still keep only one variable with nonzero coefficient in each group and LASSO estimators do random selecting.

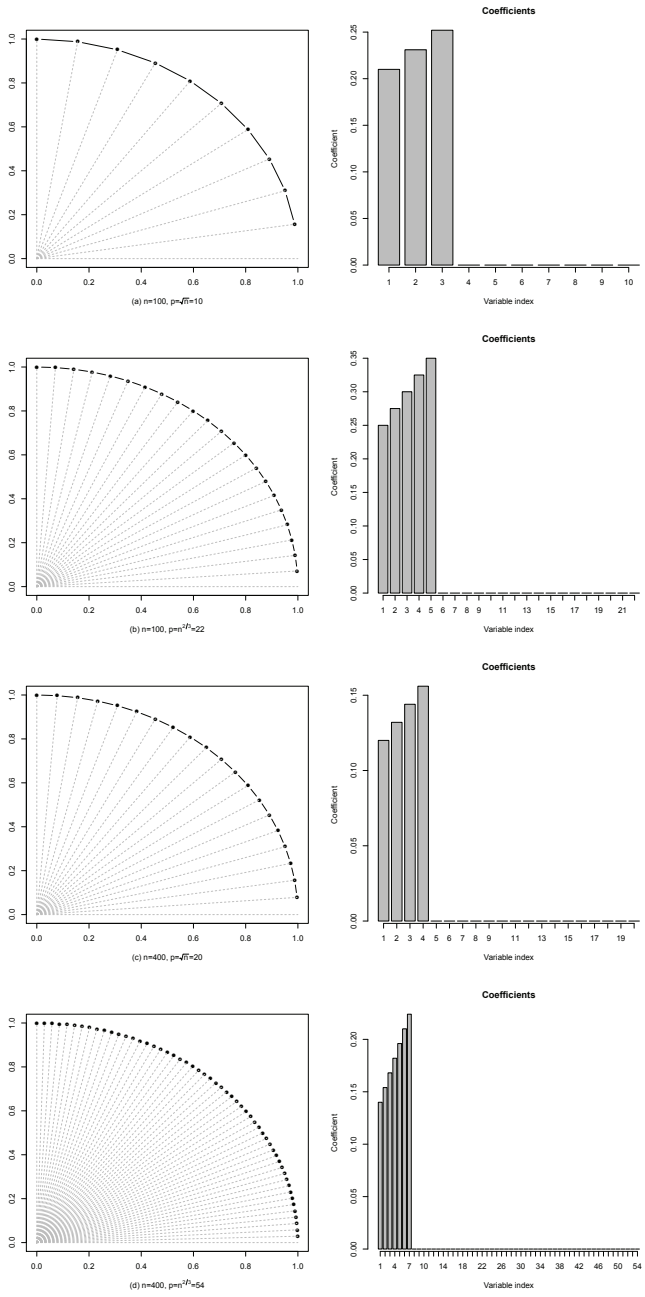

Figure 3.1: Simulation design demonstration of Example 3.1. Left panels: angles between vectors on two dimensional plane represent arccos(*correlation*); right panels: hight represents coefficients' magnitudes.

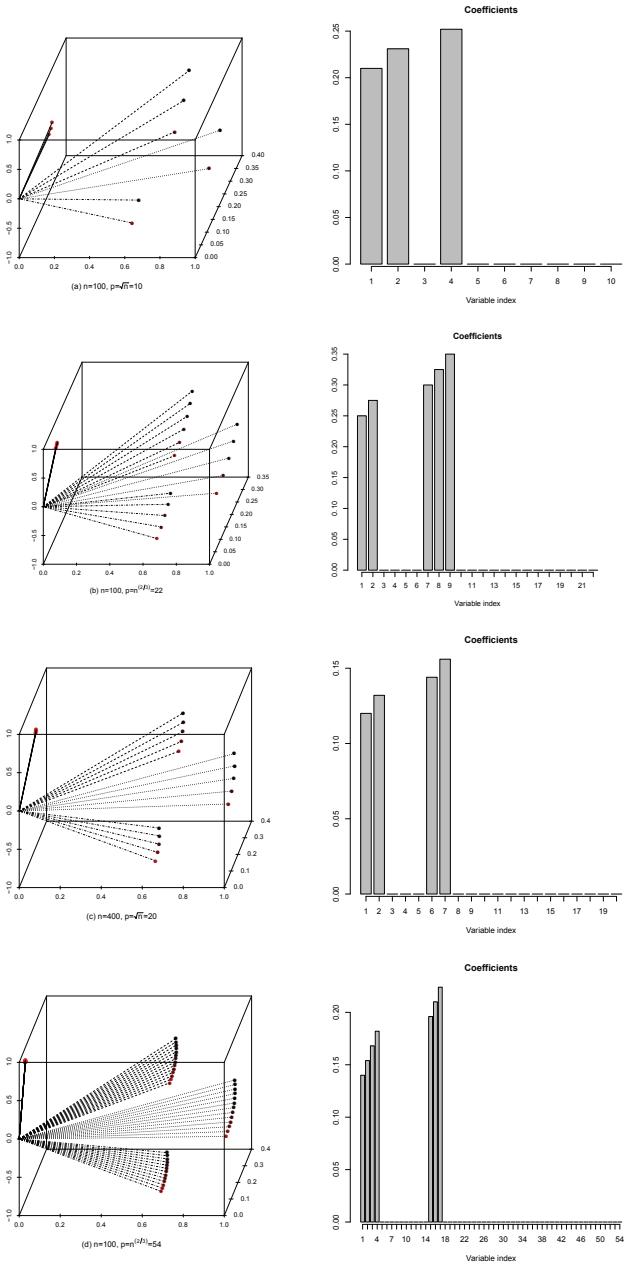

Figure 3.2: Simulation design demonstration of Example 3.2. Left panels: angles between vectors on three dimensional plane represent arccos(*correlation*); right panels: hight represents coefficients' magnitudes.

Table 3.1: Comparison of OSCAR, LASSO, SCAD and ENET estimators

	MSE	Model Size	Model Sel. Sensitivity	Model Sel. Specificity	MSE	Model Size	Model Sel. Sensitivity	Model Sel. Specificity
Example 3.1								
	$n = 100$, $p = \lfloor\sqrt{n}\rfloor$				$n = 100$, $p = \lfloor n^{2/3}\rfloor$			
OSCAR	0.98	2.54	72.50	94.79	0.97	5.81	92.10	92.91
	(0.14)	(0.69)	(23.97)	(8.49)	(0.14)	(1.84)	(16.52)	(7.76)
LASSO	0.97	1.97	47.83	92.43	0.97	2.40	29.00	94.38
	(0.15)	(0.63)	(19.65)	(8.93)	(0.63)	(0.09)	(11.99)	(5.99)
SCAD	0.96	1.00	31.67	99.21	0.96	1.00	20.00	99.97
	(0.15)	(0.07)	(7.28)	(3.27)	(0.15)	(0.07)	(0)	(0.42)
ENET	0.97	1.75	53.83	98.00	0.97	2.88	56.20	99.62
	(0.15)	(0.43)	(17.58)	(5.37)	(0.16)	(0.36)	(8.83)	(1.45)
	$n = 400$, $p = \lfloor\sqrt{n}\rfloor$				$n = 400$, $p = \lfloor n^{2/3}\rfloor$			
OSCAR	0.93	4.86	86.84	91.32	0.99	10.45	96.14	92.09
	(0.25)	(2.17)	(21.85)	(10.10)	(0.07)	(3.91)	(12.30)	(7.54)
LASSO	0.93	3.41	53.99	92.15	0.99	3.37	38.43	98.56
	(0.25)	(0.90)	(22.63)	(5.00)	(0.10)	(1.05)	(12.92)	(2.52)
SCAD	0.93	1.01	23.40	99.57	0.98	1.00	14.07	99.96
	(0.25)	(0.07)	(6.13)	(1.59)	(0.10)	(0.07)	(1.74)	(0.30)
ENET	0.93	2.45	54.92	98.44	0.99	4.97	66.64	99.35
	(0.26)	(0.58)	(17.51)	(3.63)	(0.10)	(0.51)	(11.39)	(1.47)
Example 3.2								
	$n = 100$, $p = \lfloor\sqrt{n}\rfloor$				$n = 100$, $p = \lfloor n^{2/3}\rfloor$			
OSCAR	0.96	2.71	54.67	84.64	0.96	7.99	73.90	74.76
	(0.15)	(1.45)	(33.26)	(17.06)	(0.15)	(4.33)	(29.41)	(17.72)
LASSO	0.96	2.62	30.17	75.43	0.96	3.92	37.40	87.91
	(0.15)	(0.75)	(21.03)	(11.82)	(0.15)	(0.88)	(24.50)	(7.20)
SCAD	0.97	1.30	22.17	90.93	0.96	1.95	15.70	93.15
	(0.15)	(0.48)	(15.77)	(9.30)	(0.15)	(0.26)	(8.24)	(2.81)
ENET	0.97	2.45	45.33	84.43	0.97	5.55	67.00	87.03
	(0.15)	(0.95)	(30.27)	(13.06)	(0.16)	(2.56)	(28.27)	(12.76)
	$n = 400$, $p = \lfloor\sqrt{n}\rfloor$				$n = 400$, $p = \lfloor n^{2/3}\rfloor$			
OSCAR	0.99	5.58	63.00	80.84	0.99	22.28	86.97	65.41
	(0.09)	(3.57)	(32.38)	(16.38)	(0.09)	(9.62)	(22.53)	(17.69)
LASSO	0.98	4.32	26.12	79.53	0.98	5.44	22.36	91.76
	(0.10)	(0.84)	(19.46)	(7.06)	(0.10)	(1.13)	(13.78)	(2.86)
SCAD	0.98	1.80	15.62	92.69	0.98	2.00	9.14	97.10
	(0.10)	(0.42)	(12.13)	(3.97)	(0.10)	(0.07)	(6.87)	(1.05)
ENET	0.99	4.35	54.75	86.50	1.00	9.53	74.00	90.76
	(0.10)	(2.14)	(30.22)	(12.57)	(0.10)	(4.78)	(20.24)	(9.89)

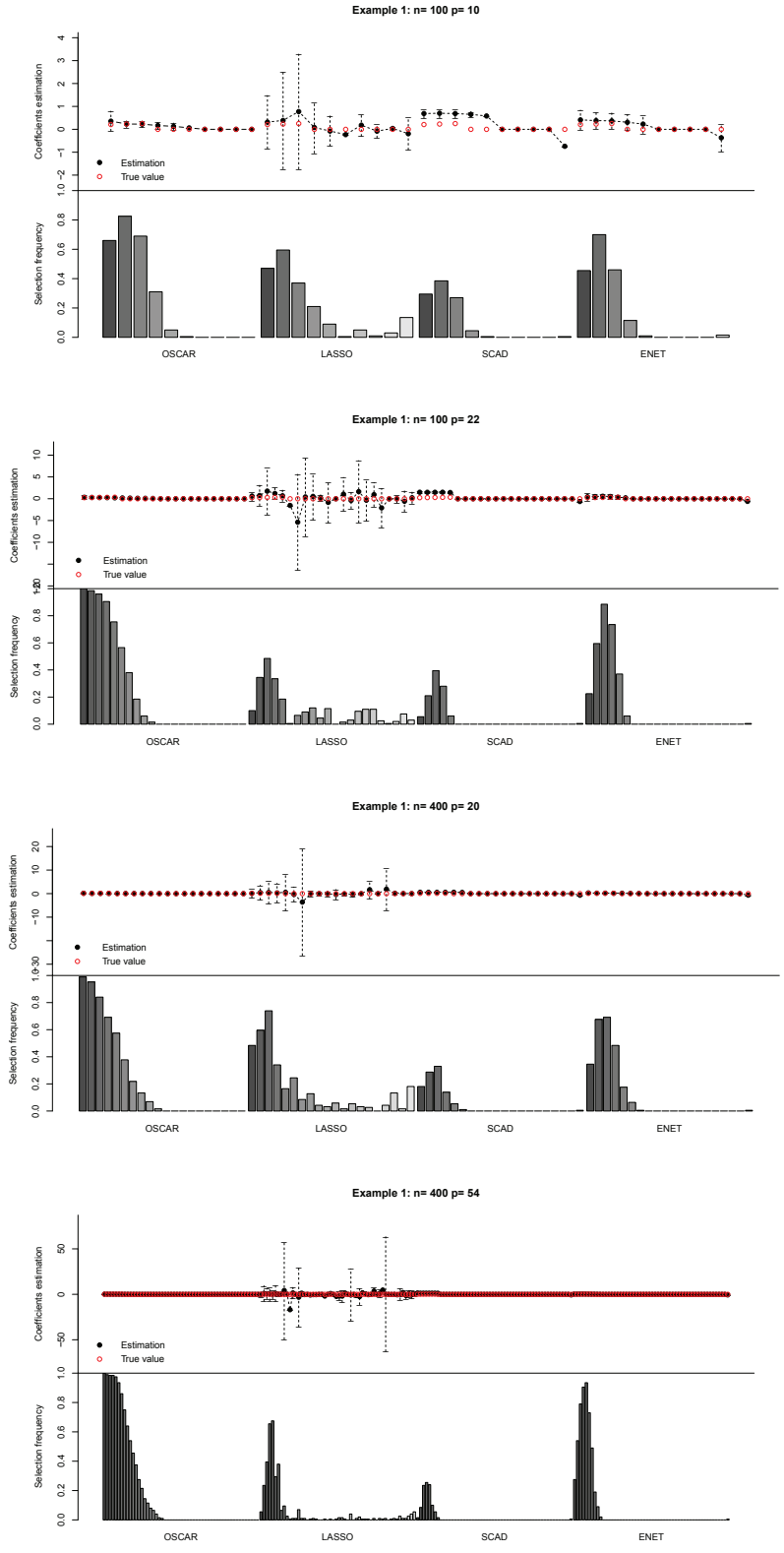

Figure 3.3: Coefficients estimation and variable selection frequency for OSCAR, LASSO, SCAD and ENET: consecutive correlation structure.

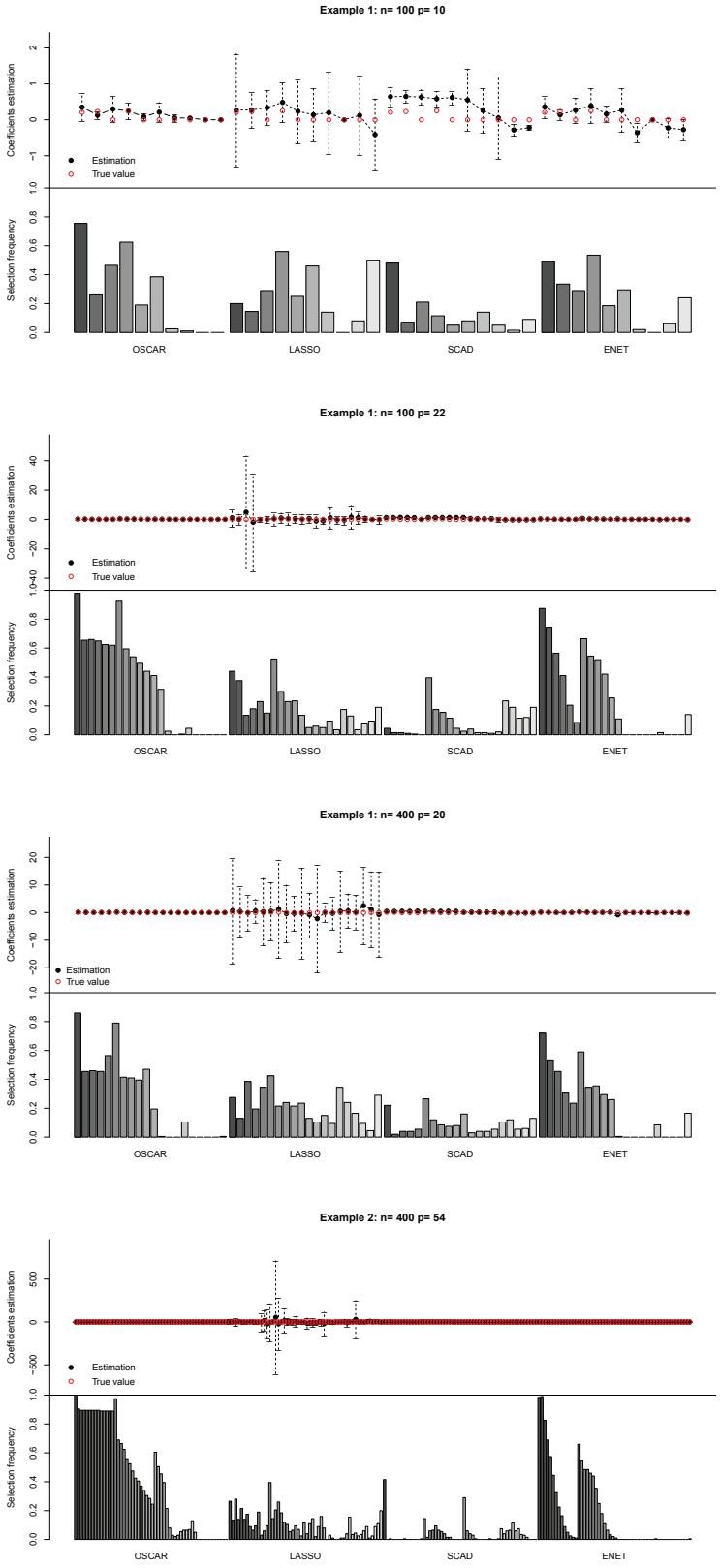

Figure 3.4: Coefficients estimation and variable selection frequency for OSCAR, LASSO, SCAD and ENET: clustered correlation structure.

3.3.2 Real data analysis

In this section, we illustrate the performance of OSCAR estimators using gene expression, genotype and phenotype (growth in the presence of drug) data from segregants obtained from a cross between two diverse strains of *Saccharomyces cerevisisae* in Chen et al. (2009). The goal is to select features from a large pool of markers and transcripts to predict the growth in the presence of drug. In the dataset, we have 104 segregants, 813 transcripts in gene expressions and 154 different genotypes. We apply OSCAR, LASSO, SCAD and Elastic Net approaches to fit a linear regression model. The tuning parameters are chosen by 10-fold cross-validation criterion. Since cross-validation errors represent prediction errors, we report the mean cross-validation error as well. The fitted values and model selection information are exhibited in Figure 3.5. In addition, OSCAR selects 25 genes, LASSO selects 74 genes, SCAD selects 13 genes and ENET selects 45 genes. Together with Figure 3.5, we can see that SCAD performs well in prediction as it has the smallest cross-validation error. However, SCAD selects too few genes and may miss the causal gene because of high correlation among the genes. On the other hand, LASSO selects too many genes leading to a overfitting model. ENET and OSCAR perform similarly while OSCAR has slightly less cross-validation errors.

3.4 Discussion

In this chapter, we investigate the model selection property of OSCAR (octagonal shrinkage and clustering algorithm for regression) estimators when the number of observations increases, and provide a set of mild conditions under which model selection consistency can be achieved. These theoretical results provide insights of the characteristics of OSCAR estimators. It is shown that OSCAR estimators are able to select the

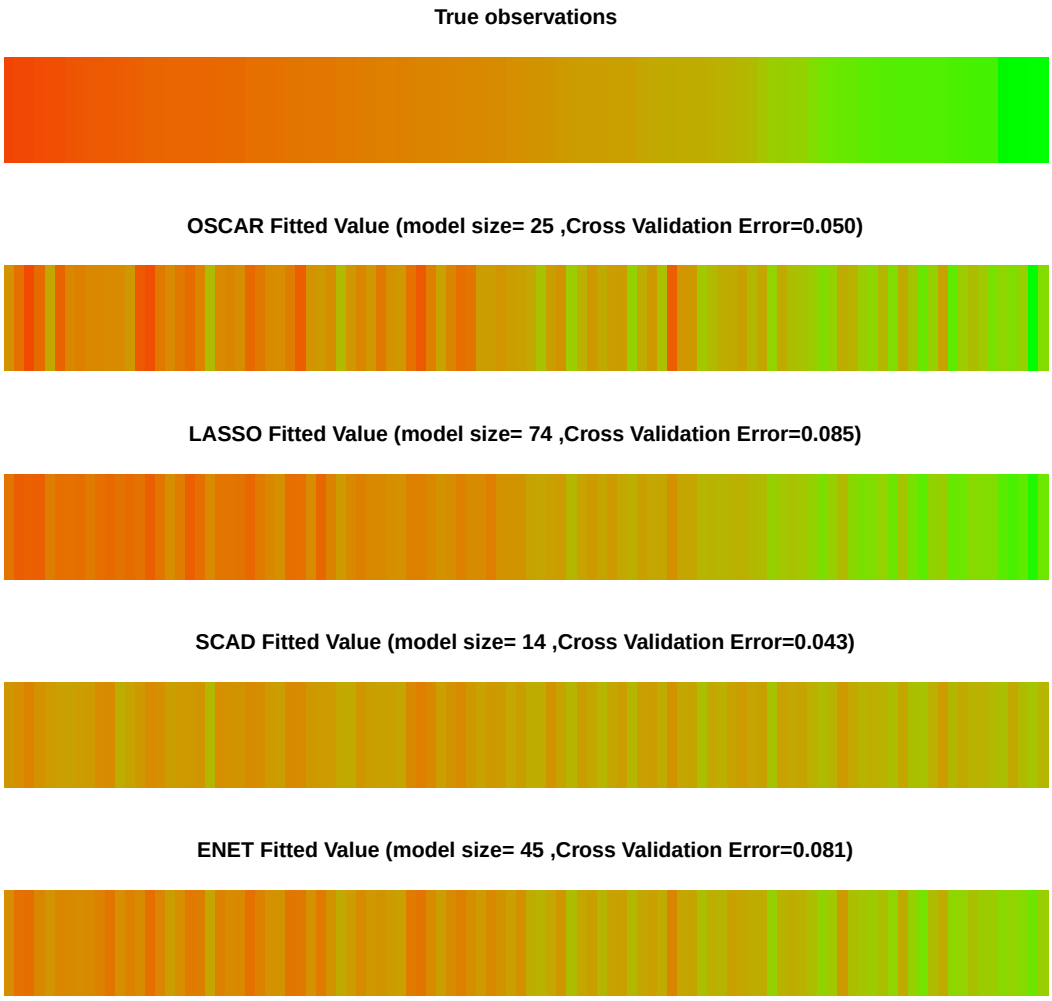

Figure 3.5: The top bar represents growth in the presence of drug; each column is associated with a different segregant (matched horizontal positions within the panel) sorted by growth from low (red) to high (green). The fitted growth rates are presented in the bottom 4 bars.

true model under reasonable conditions. The conditions reveal the relationship between coefficients' magnitudes and correlations among variables. Simulation studies further show that, compared with other penalized estimators, OSCAR estimators perform best when noise variables are highly correlated with variables with nonzero coefficients. The drawback of OSCAR estimators is that it is computationally intensive, especially when the number of parameters is large. An efficient computing algorithm would be desirable to make OSCAR estimators more applicable.

3.5 Appendix

Proof of Lemma 3.1:

By definition, $\hat{\boldsymbol{\beta}}$ is an OSCAR solution if $\mathbf{0} \in \partial P_n(\hat{\boldsymbol{\beta}})$, where $P_n(\boldsymbol{\beta}) = \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 + \lambda\delta\sum_{i<j}\max\{|\beta_i|, |\beta_j|\}]$ and $\partial P_n(\hat{\boldsymbol{\beta}})$ is the subgradient of $P_n(\cdot)$ at $\hat{\boldsymbol{\beta}}$.

We only need to calculate the subgradient for $\max\{|\beta_i|, |\beta_j|\}$. When $|\beta_i| > |\beta_j| > 0$,

$$\lim_{\gamma\to 0}\big[\max\{|\hat{\beta}_i + \gamma b_i|, |\hat{\beta}_j + \gamma b_j|\} - \max\{|\hat{\beta}_i|, |\hat{\beta}_j|\}\big]/\gamma = b_i \mathrm{sgn}(\beta_i).$$

When $|\beta_i| = |\beta_j| > 0$,

$$\lim_{\gamma\to 0}\big[\max\{|\hat{\beta}_i + \gamma b_i|, |\hat{\beta}_j + \gamma b_j|\} - \max\{|\hat{\beta}_i|, |\hat{\beta}_j|\}\big]/\gamma \geq \mathrm{sgn}(\beta_i) d_i b_i + \mathrm{sgn}(\beta_j) d_j b_j,$$

for $d_i \geq 0$, $d_j \geq 0$ and $d_i + d_j = 1$.

When $|\beta_i| = |\beta_j| = 0$,

$$\lim_{\gamma\to 0}\big[\max\{|\hat{\beta}_i + \gamma b_i|, |\hat{\beta}_j + \gamma b_j|\} - \max\{|\hat{\beta}_i|, |\hat{\beta}_j|\}\big]/\gamma \geq d_i b_i + d_j b_j,$$

where $|d_i| \leq 1$, $|d_j| \leq 1$ and $|d_i| + |d_j| \leq 1$.

Together with the well known subdifferential of L_1 norm penalty, the conclusion follows immediately. □

Before we prove the theorems, we establish the following lemma to characterize the $g-$dimensional vector $\boldsymbol{v} \in \mathcal{D}_g$ defined in (3.5).

Lemma 3.2 *For any $g-$dimensional vector $\boldsymbol{v} = (v_1, \ldots, v_g)^T$ such that $v_1 \geq v_2 \geq \cdots \geq v_g \geq 0$, then $\boldsymbol{v} \in \mathcal{D}_g$ if and only if*

$$\binom{m}{2} \leq \sum_{i=1}^{m} v_i \leq m(g-1) - \binom{m}{2}, \forall 1 \leq m \leq g.$$

Proof: We simplify the notation by rewriting $d_{ls}(l) = d_{ls}$, $\forall 1 \leq l < s \leq g$. Firstly, if $\boldsymbol{v} \in \mathcal{D}_g$, there exists $0 \leq d_{ls} \leq 1$, $1 \leq l \leq s \leq g$ such that

$$v_j = \sum_{l:l>j} d_{jl} + \sum_{l:l<j} (1 - d_{lj}).$$

Therefore,

$$\sum_{j=1}^{m} v_j = \sum_{j=1}^{m} \sum_{l=m+1}^{g} d_{jl} + \binom{m}{2}.$$

Since $0 \leq d_{ls} \leq 1$, $1 \leq l \leq s \leq g$,

$$\binom{m}{2} \leq \sum_{i=1}^{m} v_i \leq m(g-1) - \binom{m}{2}.$$

On the other hand, if

$$\binom{m}{2} \leq \sum_{i=1}^{m} v_i \leq m(g-1) - \binom{m}{2}.$$

There exists a solution of $0 \leq d_{ls} \leq 1$, $1 \leq l \leq s \leq g$ such that

$$\sum_{j=1}^{m} \sum_{l=m+1}^{g} d_{jl} = \sum_{j=1}^{m} v_j - \binom{m}{2} \leq m(g-m).$$

□

Proof of Theorem 3.1: Define oracle estimator $\hat{\boldsymbol{\beta}}^{\mathcal{O}}$ with

$$\begin{cases} \hat{\boldsymbol{\beta}}_A^{\mathcal{O}} = \Sigma_A^{-1} \boldsymbol{X}_A^T \boldsymbol{y}/n - \lambda \Sigma_A^{-1} \boldsymbol{\omega}_A \circ \operatorname{sgn}(\boldsymbol{\beta}_A^0) \\ \hat{\boldsymbol{\beta}}_{A^c}^{\mathcal{O}} = 0. \end{cases}$$

Consider three events $E_1 = \{\|\boldsymbol{X}_A^T\varepsilon/n\|_\infty < C_1^{-1}(\beta_* - \lambda c_1)\}$, $E_2 = \{\|\boldsymbol{X}_{A^c}^T(I - H_A/n)\varepsilon/n\|_\infty \leq \lambda(1-c_2)\}$, $H_A = \boldsymbol{X}_A\Sigma_A^{-1}\boldsymbol{X}_A^T$ and $E_3 = \{|\eta_j - \eta_{j+1}| \leq d_j\sqrt{v_j}, j = 1,\ldots,s-1\}$, where $\boldsymbol{\eta} = \Sigma_A^{-1}\boldsymbol{X}_A^T\varepsilon \circ \mathrm{sgn}(\boldsymbol{\beta}_A^0)/n$. Under these events, we first show that $\hat{\boldsymbol{\beta}}^{\mathcal{O}} = \mathrm{argmin}_{\boldsymbol{\beta}}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda\sum_{i=1}^n \omega_i|\beta_i|$ and $\|\hat{\boldsymbol{\beta}}_A^{\mathcal{O}} - \boldsymbol{\beta}_A^0\|_\infty < \beta_*$. Then if $\hat{\boldsymbol{\beta}}^{\mathcal{O}}$ satisfies $\mathrm{sgn}(|\hat{\beta}_i^{\mathcal{O}}| - |\hat{\beta}_j^{\mathcal{O}}|) = \mathrm{sgn}(|\hat{\beta}_i^0| - |\hat{\beta}_j^0|)$, it means $\hat{\boldsymbol{\beta}}^{\mathcal{O}}$ is the OSCAR estimator.

We need to verify that $\hat{\boldsymbol{\beta}}^{\mathcal{O}}$ satisfies the KKT conditions for optimization function $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda\sum_{i=1}^n \omega_i|\beta_i|$. Since $\hat{\boldsymbol{\beta}}_A^{\mathcal{O}} = \Sigma_A^{-1}\boldsymbol{X}_A^T\boldsymbol{y}/n - \lambda\Sigma_A^{-1}\boldsymbol{\omega}_A \circ \mathrm{sgn}(\boldsymbol{\beta}_A^0) = \boldsymbol{\beta}_A^0 + \Sigma_A^{-1}\boldsymbol{X}_A^T\varepsilon/n - \lambda\Sigma_A^{-1}\boldsymbol{\omega}_A \circ \mathrm{sgn}(\boldsymbol{\beta}_A^0)$, by condition B3.1(i) and (ii) and under event E_1,

$$\begin{aligned} & \|\hat{\boldsymbol{\beta}}_A^{\mathcal{O}} - \boldsymbol{\beta}_A^0\|_\infty \leq \|\Sigma_A^{-1}\boldsymbol{X}_A^T\varepsilon/n\|_\infty + \lambda\|\Sigma_A^{-1}\boldsymbol{\omega}_A \circ \mathrm{sgn}(\boldsymbol{\beta}_A^0)\|_\infty \\ \leq \;& C_1\|\boldsymbol{X}_A^T\varepsilon/n\|_\infty + \lambda c_1 < \beta_*. \end{aligned}$$

Therefore, $\mathrm{sgn}(\hat{\boldsymbol{\beta}}_A^{\mathcal{O}}) = \mathrm{sgn}(\boldsymbol{\beta}_A^0)$ and

$$\Sigma_A\hat{\boldsymbol{\beta}}_A^{\mathcal{O}} - \boldsymbol{X}_A^T\boldsymbol{y}/n + \lambda\boldsymbol{\omega}_A \circ \mathrm{sgn}(\hat{\boldsymbol{\beta}}_A) = 0. \tag{3.7}$$

In addition, by condition B3.1(iii) and under event E_2,

$$\begin{aligned} & \|\boldsymbol{X}_{A^c}^T(\boldsymbol{y} - \boldsymbol{X}_A\hat{\boldsymbol{\beta}}_A^{\mathcal{O}}\|_\infty \\ = \;& \|\boldsymbol{X}_{A^c}(I - \boldsymbol{X}_A\Sigma_A^{-1}\boldsymbol{X}_A^T/n)\varepsilon/n + \lambda\boldsymbol{X}_{A^c}^T\boldsymbol{X}_A\Sigma_A^{-1}\boldsymbol{\omega}_A \circ \mathrm{sgn}(\boldsymbol{\beta}_A^0)/n\|_\infty \\ \leq \;& \lambda(1-c_2) + \lambda c_2 = \lambda \end{aligned}$$

Together with (3.7), we conclude that $\hat{\boldsymbol{\beta}}^{\mathcal{O}} = \mathrm{argmin}_{\boldsymbol{\beta}}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda\sum_{i=1}^n \omega_i|\beta_i|$.

Since $\hat{\beta}_j^{\mathcal{O}} = \beta_j^0 + \eta_j\mathrm{sgn}(\beta_j^0) - \lambda\xi_j\mathrm{sgn}(\beta_j^0)$, we have $|\hat{\beta}_j^{\mathcal{O}}| = |\beta_j^0| + \eta_j - \lambda\xi_j$ and $|\hat{\beta}_j^{\mathcal{O}}| - |\hat{\beta}_{j+1}^{\mathcal{O}}| = |\beta_j^0| - |\beta_{j+1}^0| + (\eta_j - \eta_{j+1}) - \lambda(\xi_j - \xi_{j+1})$. Therefore, under event E_3,

$$\begin{aligned} & \big|(|\hat{\beta}_j^{\mathcal{O}}| - |\hat{\beta}_{j+1}^{\mathcal{O}}|) - (|\beta_j^0| - |\beta_{j+1}^0|)\big| \leq |\eta_j - \eta_{j+1}| + \lambda|\xi_j - \xi_{j+1}| \\ < \;& |\beta_j^0| - |\beta_{j+1}^0|, \end{aligned}$$

and $\mathrm{sgn}(|\hat{\beta}_j^{\mathcal{O}}| - |\hat{\beta}_{j+1}^{\mathcal{O}}|) = \mathrm{sgn}(|\beta_j^0| - |\beta_{j+1}^0|)$. So, we have shown that $\hat{\boldsymbol{\beta}}^{\mathcal{O}}$ is the OSCAR estimator and $\mathrm{sgn}(|\hat{\beta}_i| - |\hat{\beta}_j|) = \mathrm{sgn}(|\beta_i^0| - |\beta_j^0|)$ under events E_1, E_2 and E_3. Therefore, $\mathrm{P}\big(\mathrm{sgn}(|\hat{\beta}_i| - |\hat{\beta}_j|) = \mathrm{sgn}(|\beta_i^0| - |\beta_j^0|)\big) \geq \mathrm{P}(E_1 \cap E_2 \cap E_3) \geq 1 - \mathrm{P}(E_1^c) - \mathrm{P}(E_2^c) - \mathrm{P}(E_3^c)$. Compute the probabilities: by condition B2,

$$\mathrm{P}(E_1^c) \leq \sum_{j=1}^{s} \mathrm{P}(|\boldsymbol{x}_j^T \varepsilon|/n \geq C_1^{-1}(\beta_* - \lambda c_1)) \leq s \exp\{-[nC_1^{-2}(\beta_* - \lambda c_1)^2]/[2\sigma^2]\} = \epsilon,$$

and

$$\begin{aligned} \mathrm{P}(E_2^c) &\leq \sum_{j=s+1}^{p} \mathrm{P}(\|\boldsymbol{x}_j^T (I - H_A/n)\varepsilon/n\|_\infty \geq \lambda(1 - c_2)) \\ &\leq (p-s) \exp\{-[n\lambda^2(1-c_2)^2]/[2\sigma^2]\} \leq (p-s)\epsilon/p, \end{aligned}$$

and since the variance of $|\eta_j - \eta_{j+1}|$ is $v_j\sigma^2/n$,

$$\begin{aligned} \mathrm{P}(E_3^c) &\leq \sum_{j=1}^{s-1} \mathrm{P}(|\eta_j - \eta_{j+1}| \geq d_j\sqrt{v_j}) \\ &\leq \sum_{j=1}^{s-1} \mathrm{P}\big(|\eta_j - \eta_{j+1}|/\sqrt{v_j} \geq \min_{j=1}^{s-1}(d_j)\big) \\ &\leq (s-1) \exp -[\sqrt{n} \min_{j=1}^{s-1}(d_j)]^2/[2\sigma^2] \leq \epsilon. \end{aligned}$$

The conclusion follows immediately. □

Proof of Theorem 3.2: Define $\tilde{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}}\big\{\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2/(2n) + \lambda \sum_{k=1}^{K} \bar{\omega}_k \beta_k\big\} = (\boldsymbol{Z}^T\boldsymbol{Z}/n)^{-1}\boldsymbol{y} - \lambda(\boldsymbol{Z}^T\boldsymbol{Z}/n)^{-1}\bar{\boldsymbol{\omega}}$ and the oracle estimator $\hat{\boldsymbol{\beta}}^{\mathcal{O}}$ with

$$\begin{cases} \hat{\boldsymbol{\beta}}_{G_k}^{\mathcal{O}} = \tilde{\beta}_k \mathrm{sgn}(\boldsymbol{\beta}_{G_k}^0)/|G_k|)^T \\ \hat{\boldsymbol{\beta}}_{A^c}^{\mathcal{O}} = 0. \end{cases}$$

Consider events $E_1 = \big\{\|\boldsymbol{Z}^T\varepsilon/n\|_\infty \leq C_1^{-1}(\alpha\beta_* - \lambda c_1) - \lambda u_1\big\}$, $E_2 = \{\|\boldsymbol{X}_{A^c}^T(I - H_{\boldsymbol{Z}})\varepsilon/n\|_\infty < \lambda(1 - c_2 - u_2)\}$, $E_3 = \{|(\boldsymbol{z}(B_k^m) - \boldsymbol{z}_k)^T P_{\boldsymbol{Z}}\varepsilon/n| \leq \lambda\delta(|G_k| - m)/2 - |(\boldsymbol{z}(B_k^m) - \boldsymbol{z}_k)^T[P_{\boldsymbol{Z}}\Delta + \lambda\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z}/n)^{-1}\bar{\omega}]|/n, \boldsymbol{z}(B_k^m) = \sum_{j \in B_k^m} \boldsymbol{x}_j \mathrm{sgn}(\beta_j^0)/m, \forall B_k^m \subset$

$G_k, |B_k^m| = m\}$, and $E_4 = \{|\eta_k - \eta_{k+1}| < b_k^0 - b_{k+1}^0 - |\xi_k - \xi_{k+1}|, k = 1, \ldots, K-1\}$, where $\boldsymbol{\eta} = (\boldsymbol{Z}^T\boldsymbol{Z}/n)^{-1}\boldsymbol{Z}^T\varepsilon/n$.

For better presentation, we decompose $\boldsymbol{y} - \boldsymbol{X}_A\hat{\boldsymbol{\beta}}_A^{\mathcal{O}}$ as follows

$$\begin{aligned}
& \boldsymbol{y} - \boldsymbol{X}_A\hat{\boldsymbol{\beta}}_A^{\mathcal{O}} \\
= \; & \boldsymbol{y} - \sum_k \sum_{j\in G_k} \boldsymbol{x}_j \mathrm{sgn}(\beta_j^0)|\beta_j^0| + \sum_k \sum_{j\in G_k} \boldsymbol{x}_j \mathrm{sgn}(\beta_j^0)(|\beta_j^0| - b_k^0/|G_k|) \\
& + \sum_k \sum_{j\in G_k} \boldsymbol{x}_j \mathrm{sgn}(\beta_j^0)(b_k^0/|G_k| - \tilde{\beta}_k/|G_k|) \\
= \; & \varepsilon + \Delta + \sum_{k=1}^K \boldsymbol{z}_k(b_k^0 - \tilde{\beta}_k).
\end{aligned}$$

We first show that $|\tilde{\beta}_k - b_k^0|/|G_k| < \alpha\beta_*$ and $\mathrm{sgn}(\tilde{\beta}_k - \tilde{\beta}_{k+1}) = \mathrm{sgn}(b_k^0 - b_{k+1}^0)$. Then we prove that $\hat{\boldsymbol{\beta}}^{\mathcal{O}}$ satisfies the KKT conditions for OSCAR estimators

$$\begin{cases} -\boldsymbol{x}_j^T(y - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathcal{O}})/n = \lambda\tilde{\omega}_j \mathrm{sgn}(\hat{\beta}_j^{\mathcal{O}}), j \in A \\ |\boldsymbol{x}_j^T(y - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathcal{O}})|/n \le \lambda, j \in A^c, \end{cases}$$

where $\tilde{\omega}_{G_k} = \mathbf{1}_{|G_k|} + \delta\boldsymbol{v}_{G_k}$, $\boldsymbol{v}_{G_k} \in \mathcal{D}_{|G_k|}$, $k = 1, \ldots, K$.

Therefore, by condition C3.1 and under event E_1

$$\begin{aligned}
& \|D^{-1}(\tilde{\boldsymbol{\beta}} - \boldsymbol{b}^0)\|_\infty \\
= \; & \|(\boldsymbol{Z}^T\boldsymbol{Z}D/n)^{-1}[\boldsymbol{Z}^T\varepsilon/n + \boldsymbol{Z}^T\Delta/n] - \lambda(\boldsymbol{Z}^T\boldsymbol{Z}D/n)^{-1}\bar{\boldsymbol{\omega}}\|_\infty \\
\le \; & C_1\|\boldsymbol{Z}^T\varepsilon/n\|_\infty + C_1\lambda u_1 + \lambda c_1 \\
< \; & \alpha\beta_*,
\end{aligned}$$

and $|\hat{\beta}_j^{\mathcal{O}} - \beta_j^0| = |\tilde{\beta}_k/|G_k| - \beta_j^0| \le |\tilde{\beta}_k - b_k^0|/|G_k| + \left|b_k^0/|G_k| - \beta_j^0\right| < \beta_*$. Thus $\mathrm{sgn}(\hat{\boldsymbol{\beta}}_A^{\mathcal{O}}) = \mathrm{sgn}(\boldsymbol{\beta}_A^0)$.

Similar to the proof of theorem 3.1, we have $\tilde{\beta}_k = b_k^0 + \eta_k + \xi_k$ and under event E_4

$$\begin{aligned}
& |(\tilde{\beta}_k - \tilde{\beta}_{k+1}) - (b_k^0 - b_{k+1}^0)| \\
= \; & |(\eta_k - \eta_{k+1}) + (\xi_k - \xi_{k+1})| \le b_k^0 - b_{k+1}^0,
\end{aligned}$$

thus $\text{sgn}(\tilde{\beta}_k - \tilde{\beta}_{k+1}) = \text{sgn}(b_k^0 - b_{k+1}^0)$.

To show that the oracle estimator satisfies KKT condition in Lemma 3.1

$$-\boldsymbol{X}_{G_k}^T(y - \boldsymbol{X}_A\hat{\boldsymbol{\beta}}_A^{\mathcal{O}})/n + \lambda\delta\boldsymbol{v}_{G_k}\text{sgn}(\hat{\boldsymbol{\beta}}_{G_k}^{\mathcal{O}}) + \lambda[(p-s_k)\delta+1]\text{sgn}(\hat{\boldsymbol{\beta}}_{G_k}^{\mathcal{O}}) = 0,$$

for some $\boldsymbol{v}_{G_k} \in \mathcal{D}_{|G_k|}$, by Lemma 3.2, we only need to prove for $1 \le m \le |G_k|$

$$(m-1)/2 \le \|\tilde{\boldsymbol{X}}_{G_k}^T(y - \boldsymbol{X}_A\hat{\boldsymbol{\beta}}_A^{\mathcal{O}})/n - \lambda[(p-s_k)\delta+1]\mathbf{1}_{|G_k|}\|_{(1,m)}/m \le \lambda\delta(|G_k| - m/2 - 1/2) \quad (3.8)$$

where $\tilde{\boldsymbol{X}}_{G_k} = (\boldsymbol{x}_j\text{sgn}(\beta_j^0), j \in G_k)$ and $\|\boldsymbol{u}\|_{(1,m)} = \max_{|B|=m}\|\boldsymbol{u}_m\|_1$ with B being an index subset.

Since we already know $-\boldsymbol{z}_k^T(\boldsymbol{y} - \boldsymbol{X}_A\hat{\boldsymbol{\beta}}_A^{\mathcal{O}})/n + \lambda\bar{\omega}_k = 0$ which implies

$$\|\tilde{\boldsymbol{X}}_{G_k}^T(y - \boldsymbol{X}_A\hat{\boldsymbol{\beta}}_A^{\mathcal{O}})/n - \lambda[(p-s_k)\delta+1]\mathbf{1}_{|G_k|}\|_{(1,|G_k|)}/|G_k| = \lambda\delta(|G_k|-1)/2,$$

we only need to prove that for $1 \le m \le |G_k| - 1, \forall B_k^m \subset G_k$ and $|B_k^m| = m$

$$\begin{aligned} & |(\boldsymbol{z}(B_k^m) - \boldsymbol{z}_k)^T(\boldsymbol{y} - \boldsymbol{X}_A\hat{\boldsymbol{\beta}}_A^{\mathcal{O}})|/n \\ \le\ & \lambda\delta(|G_k| - m)/2, \end{aligned}$$

where $\boldsymbol{z}(B_k^m) = \sum_{j\in B_k^m}\boldsymbol{x}_j\text{sgn}(\beta_j^0)/m$.

Under event E_3,

$$\begin{aligned} & |(\boldsymbol{z}(B_k^m) - \boldsymbol{z}_k)^T(\boldsymbol{y} - \boldsymbol{X}_A\hat{\boldsymbol{\beta}}_A^{\mathcal{O}})|/n \\ =\ & |(\boldsymbol{z}(B_k^m) - \boldsymbol{z}_k)^T[P_{\boldsymbol{Z}}^{\perp}\varepsilon + P_{\boldsymbol{Z}}\Delta + \lambda\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z}/n)^{-1}\bar{\omega}]|/n \\ \le\ & |(\boldsymbol{z}(B_k^m) - \boldsymbol{z}_k)^T P_{\boldsymbol{Z}}^{\perp}\varepsilon/n| + |(\boldsymbol{z}(B_k^m) - \boldsymbol{z}_k)^T[P_{\boldsymbol{Z}}^{\perp}\Delta + \lambda\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z}/n)^{-1}\bar{\omega}]|/n \\ \le\ & \lambda\delta(|G_k| - m)/2. \end{aligned}$$

Finally, we show the KKT conditions are satisfied for variables with zero coefficients,

by condition C3.1,

$$\begin{aligned}
& \|\boldsymbol{X}_{A^c}^T(\boldsymbol{y}-\boldsymbol{X}_A\hat{\boldsymbol{\beta}}_A^{\mathcal{O}})/n\|_\infty \\
= \;& \|\boldsymbol{X}_{A^c}^T(I-H_{\boldsymbol{Z}})\varepsilon/n+\boldsymbol{X}_{A^c}^T(I-H_{\boldsymbol{Z}})\Delta/n+\lambda\boldsymbol{X}_{A^c}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z}/n)^{-1}\bar{\boldsymbol{\omega}}/n\|_\infty \\
\leq \;& \|\boldsymbol{X}_{A^c}^T(I-H_{\boldsymbol{Z}})\varepsilon/n\|_\infty+\|\boldsymbol{X}_{A^c}^T(I-H_{\boldsymbol{Z}})\Delta/n\|_\infty+\lambda\|\boldsymbol{X}_{A^c}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z}/n)^{-1}\bar{\boldsymbol{\omega}}/n\|_\infty \\
\leq \;& \lambda(1-c_2-u_2)+\lambda u_2+\lambda c_2\leq\lambda.
\end{aligned}$$

Therefore we have conclude that $\hat{\boldsymbol{\beta}}^{\mathcal{O}}$ is OSCAR estimator and satisfies $\mathcal{O}_1=\{\mathrm{sgn}(|\hat{\beta}_i|-|\hat{\beta}_j|)=\mathrm{sgn}(|\beta_i^0|-|\beta_j^0|),k\neq l,\ k,l=1,\ldots,K\}$ and $\mathcal{O}_2=\{|\hat{\beta}_i|=|\hat{\beta}_j|,\ i,j\in G_k,\ k=1,\ldots,K\}$. In addition,

$$\begin{aligned}
& \mathrm{P}(\mathcal{O}_1\cap\mathcal{O}_2)\geq\mathrm{P}(E_1\cap E_2\cap E_3\cap E_4) \\
\geq \;& 1-\mathrm{P}(E_1^c)-\mathrm{P}(E_2^c)-\mathrm{P}(E_3^c)-\mathrm{P}(E_4^c).
\end{aligned}$$

Since $\|\boldsymbol{z}_k\|_2\leq\sqrt{n}$, calculate the probabilities: by condition C3.2

$$\begin{aligned}
& \mathrm{P}(E_1^c)=\mathrm{P}(\|\boldsymbol{Z}^T\varepsilon/n\|_\infty\geq C_1^{-1}(\alpha\beta_*-\lambda c_1)-\lambda u_1\big\}) \\
\leq \;& \mathrm{P}(\|\boldsymbol{Z}^T\varepsilon/n\|_\infty\geq C_1^{-1}(\alpha\beta_*-\lambda c_1)-\lambda u_1) \\
\leq \;& \sum_{k=1}^{K}\mathrm{P}(|\boldsymbol{z}_k^T\varepsilon/n|\geq\alpha\beta_*-\lambda c_1-\lambda u_1) \\
\leq \;& \epsilon,
\end{aligned}$$

and

$$\begin{aligned}
& \mathrm{P}(E_2^c)\leq\mathrm{P}\big(\|\boldsymbol{X}_{A^c}^T(I-H_{\boldsymbol{Z}})\varepsilon/n\|_\infty\geq\lambda(1-c_2-u_2)\big) \\
\leq \;& \sum_{j=s+1}^{p}\mathrm{P}\big(|\boldsymbol{x}_j^T(I-H_{\boldsymbol{Z}})\varepsilon/n|\geq\lambda(1-c_2-u_2)\big) \\
\leq \;& (p-s)\epsilon/p,
\end{aligned}$$

and

$$
\begin{aligned}
& \mathrm{P}(E_3^c) \\
\leq\ & \mathrm{P}(|(\boldsymbol{z}(B_k^m) - \boldsymbol{z}_k)^T P_{\boldsymbol{Z}}^{\perp} \varepsilon / n| \geq \\
& \lambda\delta(|G_k| - m)/2 - |(\boldsymbol{z}(B_k^m) - \boldsymbol{z}_k)^T [P_{\boldsymbol{Z}}^{\perp}\Delta + \lambda \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z}/n)^{-1}\bar{\omega}]|/n, \\
& B_k^m \subset G_k, m = 1, \ldots, |G_k| - 1, k = 1, \ldots, K) \\
\leq\ & \sum_{k=1}^{K} \sum_{m=1}^{|G_k|-1} \binom{|G_k|}{m} \exp\{-n(\phi_k^m)^2/(2\sigma^2)\} \\
\leq\ & s\epsilon/p,
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{P}(E_4^c) &\leq \sum_{k=1}^{K-1} \mathrm{P}(|\eta_k - \eta_{k+1}| \geq b_k^0 - b_{k+1}^0 - |\xi_k - \xi_{k+1}|) \\
&\leq \sum_{k=1}^{K-1} \mathrm{P}(|\eta_k - \eta_{k+1}|/v_k \geq d_k) \leq \epsilon.
\end{aligned}
$$

□

Chapter 4

Group OSCAR parameter estimation and model selection in presence of unknown group structures

4.1 Introduction

In this chapter, also consider the linear regression model

$$\boldsymbol{y} = \sum_{i=1}^{p} \boldsymbol{x}_i \beta_i + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} = (y_1 \dots y_n)^T$ is the response vector, $\boldsymbol{x}_1 = (x_{11} \dots x_{1n})^T, \dots, \boldsymbol{x}_p = (x_{p1} \dots x_{pn})^T$ are the vectors of p explanatory variables, $\beta_1, \dots, \beta_p$ are the corresponding regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ is the vector of independent random errors. Classical literature typically considers the case when p is finite and small. But in recent years, with increasing availability of large size data and computing power, there is a tremendous amount of publications dealing with data of large p. Variable selection plays an important role in this development. In this paper, we address a model selection problem where the variables are naturally grouped but only part of the group structure is known. There are several papers that have taken group structures into consideration in recent literatures. However, most of them only consider the situation that the group structure is completely given and known. See, e.g. Yuan & Lin (2006), Zhao et al. (2008) and so on. In practice, there are situations where the underlying group structures are not known. Bondell & Reich (2008) study the "unknown predictive clusters" of potentially

highly correlated explanatory variables. Here, we consider model selection problems with both types of groups, known or unknown structures, to make full use of available group information.

Our research is motivated by a project of nuclear detection sponsored by the US Department of Homeland Security through the Command, Control, and Interoperability Center for Advanced Data Analysis (CCICADA), a DHS Center of Excellence. In this project, we need to analyze and process a variety of information in customs forms from large volumes of shipping containers. The data compiled from the custom forms, referred as "manifest data", is used to detect high-risk containers and to learn about important variables. The manifest data have special features and require special considerations. For instance, most of the information contained in the manifest data is given by categorical variables, which are often represented by dummy variables that form natural groups; see, e.g., Yuan & Lin (2006) who described this type of groups as the known groups of "derived input variables". In addition, some of the groups of categorical variables can be highly correlated. For example, although 'Voyage Number' and 'Inbond Entry Type' are two different variables, they are highly correlated because of redundant information. Such and other complications could cause problems in model fitting. Nevertheless, the group structures of variables provide an important source of regularization which can properly handle these issues and help model building.

Recently, penalized regression is a subject that has generated a lot of publications and it has emerged as a successful technique for model selection, see, e.g. ridge regression (Frank & Friedman, 1993), L_1 norm penalty (Tibshirani, 1996; Chen et al., 2001; Efron et al., 2004), SCAD (Fan & Li, 2001) and MCP (Zhang, 2010). The penalized

regression technique have also been extended to handle the case when explanatory variables are grouped. When the group structures in the explanatory variables is known, Yuan & Lin (2006) generalize the LASSO method, the LARS algorithm and the nonnegative garrotte penalty for selecting grouped variables. They define an L_2 norm of the coefficients associated with a group of variables as the component of the penalty functions. Zhao et al. (2008) propose composite absolute penalties (CAP) families which combine the norm penalties at the between-group and within-group levels and group selection occurs for nonoverlapping groups. Li et al. (2010) address this problem differently using the derivative of the conditional mean to achieve groupwise dimension reduction for the explanatory variables. Huang et al. (2009) use a specially designed group bridge approach to carry out variable selection at the between-group and within-group levels simultaneously. Wang et al. (2009) consider the same problem in Cox regression with known group structure by reparameterizing the coefficients and imposing a L_1 norm penalty. All these methods assume the group structure is completely known.

Group structures can be retrieved from different sources. On one hand, prior knowledge can provide information about group structures. For example, dummy variables created to represent different factors of one categorical variable are considered as one group or the experts who collect the data may suggest to split the explanatory variables into several groups. On the other hand, the data itself contain group structures which are not available beforehand. For instance, identical or highly correlated variables may need to be grouped together because of their similarity. In other words, the underlying group information is implied in the correlation patterns among explanatory variables.

In order to identify the group structure and incorporate the correlation patterns, supervised clustering can be used to determine useful groups of explanatory variables. One approach that has been used by Jornsten & Yu (2003) is to create a new explanatory variable by averaging the explanatory variables in the same group. Equivalently, one can also achieve this goal by assigning identical coefficients or imposing smoothness among coefficients corresponding to highly correlated explanatory variables. Bondell & Reich (2008) construct the OSCAR penalty function by combining L_1 norm and pairwise L_∞ norm of the parameters to enforce highly correlated variables with identical coefficients. Huang et al. (2010b) establish the sparse Laplacian shrinkage (SLS) estimator which has a generalized grouping property with regards to the graph represented by the Laplacian quadratic in the penalty function. SLS has a grouping property like other general L_2 penalties such as elastic net by Zou & Hastie (2005) which combines the L_1 and L_2 penalties.

In this chapter, we propose a penalized approach that is able to capture group features in the data from both sources for variable selection. Unlike most aforementioned publications, we do not assume the group structure is completely known. Also, unlike OSCAR, this method does not discard any prior group information but utilizes both known group structure and correlation patterns in the existing data. It can be shown that the group lasso (Yuan & Lin, 2006), CAP (Zhao et al., 2008) or OSCAR (Bondell & Reich, 2008) are all special cases of our general model.

The rest of this chapter is organized as follows. In Section 4.2 and Section 4.3, we define the general model and consider some special cases along with an iterative algorithm. In Section 4.4, the large sample properties of the penalized regression approach are investigated. In Section 4.5, simulation studies and detailed analysis of the manifest

data are presented. In Section 4.6, further discussions are provided.

4.2 Variable Selection via Penalized Regression

4.2.1 A general formulation for penalized regression with group structures

Suppose there are p explanatory variables in all. Let $G_j \subseteq (1, ..., p)$, $j = 1, ..., J$ be the J non-overlapping subsets of the indices $(1, ..., p)$ associated with groups of variables. Denoted by $\boldsymbol{G} = \{G_1, ..., G_K\}$ the given group structure, $|G_j| = p_j$ and $\sum_{j=1}^{J} p_j = p$, where $|G_j|$ is the cardinality of G_j. Here, we assume $\boldsymbol{G}$ is given or known. In this chapter, we also assume that there are potentially high correlations among some of these J groups. For instance, let us say s groups, $G_{j_1}, ..., G_{j_s}$ among $\boldsymbol{G} = \{G_1, ..., G_J\}$ have high correlations. In the context of linear models, they provide exactly the same or similar information for regression analysis. Thus, it is reasonable to merge them into one new group $G_k^* = \cup_{i=1}^{s} G_{j_i}$. Therefore, the division $\boldsymbol{G} = \{G_1, ..., G_J\}$ reflects the prior knowledge about group structures while $\boldsymbol{G^*} = \{G_1^*, ..., G_K^*\}$, $K \leq J$, incorporates both prior information and the correlation patterns in the design matrix. Thus, $\boldsymbol{G}$ refines $\boldsymbol{G}^*$. In the case when $G^* \equiv G$, it reduces to the setting consider by Yuan & Lin (2006), Zhao et al. (2008) and so on. In the case when $|G_k| = p_k \equiv 1$, it becomes the setting considered by Bondell & Reich (2008).

Without loss of generality, we assume that the $\boldsymbol{x}$'s are standardized so that $\sum_{j=1}^{n} x_{ij} = 0$ and $\sum_{j=1}^{n} x_{ij}^2 = n$, $i = 1, \ldots, p$. Also denote the corresponding coefficients for each

group as $\boldsymbol{\beta}_{G_j} = (\beta_l, l \in G_j)$. In order to handle model selection and correlation problems simultaneously, we propose the penalized least squares estimator as:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \rho_1(\boldsymbol{\beta}_{G_j}, j = 1, ..., J) + \lambda_2 \sum_{i \neq j} \rho_2(c(G_i, G_j)), \quad (4.1)$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ are tuning parameters; ρ_1, ρ_2 are penalty functions and $c(G_i, G_j)$ is a function which reflects the relationship between G_i and G_j. The two penalty terms perform different functions in the regularization. The first part ρ_1 induces sparsity in model selection while the second term ρ_2 encourages highly correlated groups to have the same impact on the response variable, that is, they will have identical groupwise coefficients. It means G_i and G_j are merged into one super group.

It is worth noting that (4.1) is a general model and contains several important special cases as ρ_1, ρ_2 and c functions can be chosen flexibly. For example, if set $\rho_1 = \sum_j (\boldsymbol{\beta}_{G_j}^T B_j \boldsymbol{\beta}_{G_j})^{1/2}$ with symmetric positive definite matrix B_j's and set $\rho_2 \equiv 0$ (i.e. ignoring the correlation patterns in the predictors), the estimator is (4.1) is Group Lasso estimator (Yuan & Lin, 2006). The CAP (Zhao et al., 2008) can be obtained from (4.1) by setting $\rho_1 = \sum_j (\|\boldsymbol{\beta}_{G_j}\|_{n_j})^{n_0}$ and $\rho_2 \equiv 0$, where $\|\boldsymbol{\beta}_{G_j}\|_{n_j}$ is L_{n_j} norm. In addition, Group bridge estimators (Huang et al., 2009) can be obtained from (4.1) by defining ρ_1 as $\sum_j c_j \|\boldsymbol{\beta}_{G_j}\|_1^\gamma$ and $\rho_2 \equiv 0$, where c_j and γ are constants. Furthermore, Bondell & Reich (2008) consider another special case of (4.1). They choose ρ_1 as the L_1 norm penalty of $\boldsymbol{\beta}$ and ρ_2 as the summation of L_∞ norm for pairwise $\boldsymbol{\beta}$. Finally, if MCP is chosen as ρ_1 and Laplacian quadratic is used as ρ_2 for the combination, we have Sparse Laplacian Shrinkage estimator (Huang et al., 2010b).

4.2.2 Group OSCAR

To incorporate both prior group structure and correlation patterns, we further reparameterize β_{kj} as

$$\boldsymbol{\beta}_{G_j} = \gamma_j \theta_{G_j}, j = 1, ..., J, \tag{4.2}$$

where $\gamma_i \geq 0$. The parameter γ_j controls all the $\boldsymbol{\beta}_{G_j}$ in G_j while $\theta_i, i \in G_j$ show the differences of coefficients within the certain group. Thus, the impact on the response variable of a certain group can be written as $\boldsymbol{X}_{G_j}\boldsymbol{\beta}_{G_j} = \gamma_j \boldsymbol{X}_{G_j}\boldsymbol{\theta}_{G_j}$, $j = 1, \ldots, J$. The correlations between two groups G_i and G_j are reflected by the correlation between $\boldsymbol{X}_{G_i}\boldsymbol{\theta}_{G_i}$ and $\boldsymbol{X}_{G_j}\boldsymbol{\theta}_{G_j}$. Similar practice of creating a new explanatory variables by averaging the grouped variables can be found in Jornsten & Yu (2003); Dettling & Bühlmann (2004) as well.

Note that since the group representing variables $\boldsymbol{z}_j = \boldsymbol{X}_{G_j}\boldsymbol{\theta}_{G_j}$, $j = 1, \ldots, J$ are rescaled by θ's, it is important to remove the scale effect in order to identify true group correlation structures. To tackle this issue, we further impose a constraint

$$\|\boldsymbol{z}_j\|_2^2 = \|\boldsymbol{X}_{G_j}\boldsymbol{\theta}_{G_j}\|_2^2 = n. \tag{4.3}$$

Therefore, γ_j can be expressed explicitly in terms of $\boldsymbol{\beta}_{G_j}$:

$$\gamma_j = \|\boldsymbol{X}_{G_j}\boldsymbol{\beta}_{G_j}\|_2/\sqrt{n}. \tag{4.4}$$

Inspired by OSCAR (Bondell & Reich, 2008), we define Group OSCAR estimator as

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda_1 \sum_{j=1}^{J} \gamma_j + \lambda_1 \delta \sum_{i<j} \max(\gamma_i, \gamma_j) \\
&= \text{argmin}_{\boldsymbol{\beta}} \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2}{2n} + \lambda_1 \sum_{j=1}^{J} \frac{\|\boldsymbol{X}_{G_j}\boldsymbol{\beta}_{G_j}\|_2}{\sqrt{n}} + \frac{\lambda_1 \delta}{\sqrt{n}} \sum_{i<j} \max(\|\boldsymbol{X}_{G_i}\boldsymbol{\beta}_{G_i}\|_2, \|\boldsymbol{X}_{G_j}\boldsymbol{\beta}_{G_j}\|_2),
\end{aligned}$$

where $\lambda_1 > 0$ and $\delta > 0$ are tuning parameters.

The grouping effect of the estimator is quantified by the following theorem.

Theorem 4.1 *Let* $\hat{\boldsymbol{z}}_j = \boldsymbol{X}_{G_j}\hat{\boldsymbol{\theta}}_{G_j}$ *and* $\hat{\boldsymbol{z}}_{j'} = \boldsymbol{X}_{G_{j'}}\hat{\boldsymbol{\theta}}_{G_{j'}}$. *Define the group correlation between* G_j *and* $G_{j'}$ *as* $\hat{\phi}_{jj'} = cor(\hat{\boldsymbol{z}}_j, \hat{\boldsymbol{z}}_{j'}), j \neq j'$ *where* $cor(\mathbf{u}, \mathbf{v})$ *is sample correlation for vectors* $\mathbf{u}$ *and* $\mathbf{v}$.

Then, for a given pair of groups G_j *and* $G_{j'}$, *suppose both* $\hat{\gamma}_j > 0$ *and* $\hat{\gamma}_{j'} > 0$ *are distinct from other* $\hat{\gamma}$*'s. Thus, if* λ_1 *and* δ *are chosen such that*

$$\lambda_1\delta > \|\boldsymbol{y}\|\sqrt{2(1-\hat{\phi}_{jj'})/n}$$

we have $\hat{\gamma}_j = \hat{\gamma}_{j'}$.

The group correlation $\hat{\phi}_{jj'}$ defined in Theorem 4.1 amounts to the weighted correlation of the explanatory variables in two groups. If there are highly correlated variables in two groups with nonzero $\theta's$, their group correlation are likely to be large and hence likely to be merged into one group. On the other hand, even highly correlated explanatory variables exist in two groups, if at least one of them has zero θ, the group correlation may still be small and does not lead to grouping. This property is a result from an interaction of the two-level selection and grouping effect. This result is an extension of Bondell & Reich (2008) to our setting. In the special case with $p_1 = p_2 = \cdots = p_K = 1$, this theorem is the same as Theorem 1 in Bondell & Reich (2008).

4.3 Asymptotic properties

In this section, we investigate the asymptotic properties of the penalized estimator. Without loss of generality, assume the true regression coefficients are $\boldsymbol{\beta}^0 = (\beta_i^0, i =$

$1, \ldots, p)$ and

$$\gamma_1^0 = \cdots = \gamma_{J_1} > \gamma_{J_1+1}^0 = \cdots = \gamma_{J_2}^0 > \cdots > \gamma_{J_{K-1}+1} = \ldots \gamma_{J_K} > 0, \gamma_{J_K+1} = \cdots = \gamma_J = 0,$$

where $\gamma_j^0 = \|\boldsymbol{X}_{G_j}\boldsymbol{\beta}_{G_j}^0\|_2/\sqrt{n}$. That is, the first J_K groups $G_1, \ldots, G_{J_K}$ have nonzero coefficients and can be merged into K hyper groups.

Denote by $A = \{i : \beta_i^0 \neq 0\}$ which is the set of indices of nonzero coefficients and $s = |A|$ be the cardinality of A. Let $\alpha_1^0 > \alpha_2^0 > \cdots > \alpha_K^0 > 0$ be the set of distinct values of $\{\gamma_j^0 : j = 1, \ldots, J_K\}$. For presentation convenience, define a notation for hyper-groups and merged groups:

$$\mathscr{G}_k^h = \{j : \gamma_j^0 = \alpha_k^0\}, G_k^* = \{G_j : \gamma_j^0 = \alpha_k^0\} k = 1, \ldots, K.$$

Denote $\omega_j = (J - j)\delta + 1$ and $\bar{\boldsymbol{\omega}} = (\bar{\omega}_k, k = 1, \ldots, K)$, where $\bar{\omega}_k = \sum_{j \in \mathscr{G}_k^h} \omega_j / |\mathscr{G}_k^h|$.

Recall the definition of subspaces $\mathcal{D}_g \subset \mathbb{R}^g$, $g \in \mathbb{N}$ in Chapter 3:

$$\begin{aligned} \mathcal{D}_g = \{ \quad & \boldsymbol{v} = (v_1, \ldots, v_g)^T : v_j = \sum_{l:l<j, 1 \leq l \leq g} d_{lj}(j) + \sum_{l:l>j 1 \leq l \leq g} d_{jl}(j), j = 1, \ldots, g, \\ & \text{with} \;\; d_{ls}(l) \geq 0, d_{ls}(s) \geq 0 \;\; \text{and} \;\; d_{ls}(l) + d_{ls}(s) = 1, \forall 1 \leq l < s \leq g\}, \end{aligned}$$

where $d_{ls}(\cdot)$ is a function on the set $\{l, s\}$.

We characterize Group OSCAR estimators in the following lemma.

Lemma 4.1 *Suppose $\hat{\alpha}_1 > \hat{\alpha}_2 > \cdots > \hat{\alpha}_{\hat{K}} > 0$ are the distinct values of $\{|\hat{\gamma}_j| \neq 0 : j = 1, \ldots, J\}$. Then $\hat{\boldsymbol{\beta}}$ is Group OSCAR estimator if*

$$-\boldsymbol{X}_{\hat{G}_k^*}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n + \lambda_1 \bar{\rho}(\boldsymbol{\beta}_{\hat{G}_k^*}) = 0, k = 1, \ldots, \hat{K},$$

and

$$\|\boldsymbol{X}_{G_j}(\boldsymbol{X}_{G_j}^T\boldsymbol{X}_{G_j})^{-1}\boldsymbol{X}_{G_j}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n\|_2 \leq \lambda_1/\sqrt{n}, \text{for} \;\; \hat{\gamma}_j = 0,$$

where $\hat{G}_k^* = \{G_j : |\hat{\gamma}_j| = \hat{\alpha}_k\}$ *and* $\bar{\rho}(\boldsymbol{\beta}_{\hat{G}_k^*}) = (\tilde{\omega}_j \boldsymbol{X}_{G_j}^T \boldsymbol{X}_{G_j} \hat{\boldsymbol{\beta}}_{G_j} / (\|\boldsymbol{X}_{G_j} \hat{\boldsymbol{\beta}}_{G_j}\|_2 \sqrt{n}), j \in \hat{\mathscr{G}}_k^h)$ *with* $\tilde{\omega}_{\hat{\mathscr{G}}_k^h} = \delta \boldsymbol{v}_{\hat{\mathscr{G}}_k^h} + [(J - \sum_{l<k} |\hat{\mathscr{G}}_l^h|)\delta + 1]\mathbf{1}_{\hat{\mathscr{G}}_k^h}$ *for some* $\boldsymbol{v}_{\hat{\mathscr{G}}_k^h} \in \mathcal{D}_{|\hat{\mathscr{G}}_k^h|}$.

Now we investigate the estimation and model selection properties of Group OSCAR estimators. Note that the correlation between the estimated group representative variables $\boldsymbol{X}_{G_i} \hat{\boldsymbol{\beta}}_{G_i} / \|\boldsymbol{X}_{G_i} \hat{\boldsymbol{\beta}}_{G_i}\|_2$ and $\boldsymbol{X}_{G_j} \hat{\boldsymbol{\beta}}_{G_j} / \|\boldsymbol{X}_{G_j} \hat{\boldsymbol{\beta}}_{G_j}\|_2$ is different from the correlation between true group representative variables $\boldsymbol{X}_{G_j} \boldsymbol{\beta}^0_{G_j} / \|\boldsymbol{X}_{G_j} \boldsymbol{\beta}^0_{G_j}\|_2$ and $\boldsymbol{X}_{G_i} \boldsymbol{\beta}^0_{G_i} / \|\boldsymbol{X}_{G_i} \boldsymbol{\beta}^0_{G_i}\|_2$. Thus, to ensure model selection consistency, the estimated representative variable and true representative variable should be close enough. We impose the following condition on the design matrix:

Condition A 4.1 *Assume there exist constants* $\kappa_k > 0$, $k = 1, \ldots, K$ *such that for any* $\boldsymbol{u} \in \mathbb{R}^s$

$$\|\boldsymbol{X}_A \boldsymbol{u}\|_2 \geq \sum_{k=1}^{K} \sum_{j \in \mathscr{G}_k^h} \kappa_k \|\boldsymbol{X}_{G_j} \boldsymbol{u}_{G_j}\|_2.$$

We further define a subspace in $\mathbb{R}^s$ in which group subvectors are close to the true coefficients for corresponding groups:

$$\begin{aligned} \mathcal{U} \quad = \quad & \{\boldsymbol{u}, \kappa_k \|\boldsymbol{X}_{G_j}(\boldsymbol{u}_{G_j} - \boldsymbol{\beta}^0_{G_j})\|_2 / \sqrt{n} \leq \lambda \bar{\omega}_k / 2 + \sqrt{(\lambda \bar{\omega}_k)^2/4 + 2\kappa_k \bar{\omega}_k \|\boldsymbol{X}_{G_j} \boldsymbol{\beta}^0_{G_j}\|_2 / \sqrt{n}}, \\ & j \in \mathscr{G}_k^h, k = 1, \ldots, K\}. \end{aligned}$$

Since both estimated and true representative variables are normalized, the differences among them can be quantified by the angle between these two, i.e.

$$\theta_j = \arccos\left[(\boldsymbol{X}_{G_i} \boldsymbol{u}_{G_i} / \|\boldsymbol{X}_{G_i} \boldsymbol{u}_{G_i}\|_2)^T (\boldsymbol{X}_{G_j} \boldsymbol{\beta}^0_{G_j} / \|\boldsymbol{X}_{G_j} \boldsymbol{\beta}^0_{G_j}\|_2)\right], j = 1, \ldots, J_K.$$

For $\boldsymbol{u} \in \mathcal{U}$, we have that $\sin(\theta_j/2) \leq c_k \sqrt{\lambda \bar{\omega}_k}, j \in \mathscr{G}_k^h$, where $c_k = \min\{c : 2/c^2 + \sqrt{\lambda \bar{\omega}_k}/c \leq \|\boldsymbol{X}_{G_j} \boldsymbol{\beta}^0_{G_j}\|_2 / \sqrt{n}, j \in \mathscr{G}_k^h\}$. Therefore, if estimated coefficients are in $\mathcal{U}$, estimated representative variables and true representative variables are close enough.

We further impose the following conditions:

Condition B 4.1 *Denote*

$$\boldsymbol{F_u} = (\boldsymbol{f}_k, k = 1, \ldots, K),$$

where $|\mathscr{G}_k^h|\boldsymbol{f}_k = \sum_{j\in\mathscr{G}_k^h}(\boldsymbol{X}_{G_j}\boldsymbol{u}_{G_j}/\|\boldsymbol{X}_{G_j}\boldsymbol{u}_{G_j}\|_2$. *Assume*

1. *For* $i \in \mathscr{G}_k^h$, $j \in \mathscr{G}_l^h$, $k \neq l$,

$$\begin{aligned} & \gamma_i^0 - \gamma_j^0 \\ > \ & \frac{\bar{\omega}_k/\kappa_k - \bar{\omega}_l/\kappa_l}{2} + \frac{\sqrt{(\lambda\bar{\omega}_k)^2/4 + 2\kappa_k\lambda\bar{\omega}_k\gamma_i^0}}{\kappa_k} - \frac{\sqrt{(\lambda\bar{\omega}_l)^2/4 + 2\kappa_l\lambda\bar{\omega}_l\gamma_j^0}}{\kappa_l} \end{aligned}$$

2. $\exists\ 0 < t_0 < 1$, *such that*

$$\sup_{\boldsymbol{u}\in\mathcal{U}} \|(P_{\boldsymbol{X}_{G_i}} - P_{\boldsymbol{X}_{G_j}})(P_{\boldsymbol{X}_A} - P_{\boldsymbol{F_u}})\boldsymbol{F}^0\boldsymbol{\gamma}^0/\sqrt{n}\|_2 \leq t_0\lambda_1\delta/4,$$

and

$$\sup_{\boldsymbol{u}\in\mathcal{U}} \|(P_{\boldsymbol{X}_{G_i}} - P_{\boldsymbol{X}_{G_j}})\boldsymbol{F_u}(\boldsymbol{F_u}^T/n)^{-1}\bar{\boldsymbol{\omega}}\|_2/\sqrt{n} \leq t_0\delta/4,$$

$i, j \in \mathscr{G}_k^h$, $k = 1, \ldots, K$

3. *For groups with zero coefficients,*

$$\sup_{j>J_K}\sup_{\eta} \|\boldsymbol{X}_{G_j}(\boldsymbol{X}_{G_j}^T\boldsymbol{X}_{G_j})^{-1}\boldsymbol{X}_{G_j}^T\boldsymbol{X}_A(\boldsymbol{X}_A^T\boldsymbol{X}_A)^+\eta\|_2 \leq C_1 < 1,$$

where $\eta = (\omega_i\boldsymbol{X}_{G_i}^T\boldsymbol{X}_{G_i}\boldsymbol{u}_{G_i}/(\|\boldsymbol{X}_{G_i}\boldsymbol{u}_{G_i}\|_2\sqrt{n}), i \in \mathscr{G}_k^h, k = 1, \dot{)}$, $\boldsymbol{u} \in \mathcal{U}$ *and* $\omega_{\mathscr{G}_k^h} = \delta\boldsymbol{v}_{\mathscr{G}_k^h} + [(J - J_{k-1})\delta + 1]\mathbf{1}_{\mathscr{G}_k^h}$ *for some* $\boldsymbol{v}_{\mathscr{G}_k^h} \in \mathcal{D}_{|\mathscr{G}_k^h|}$.

Condition B 4.2 *Assume the following conditions on the tuning parameters: for* $0 < \epsilon < 1$,

1. *For* $k = 1, \ldots, K$, *let* $c_k = \max_{j \in \mathscr{G}_k^h}\{\Sigma_{G_j}^{-1}(d,d), d = 1, \ldots, p_j\}$ *and* $p_k^* = \max\{p_j, j \in \mathscr{G}_k^h\}$, *assume*

$$\lambda_1 \bar{\omega}_k - \sigma c_k \sqrt{p_k^*} \geq \sigma \sqrt{2\log(2|\mathscr{G}_k^h|K/\epsilon)/n}.$$

2. *For* $i, j \in \mathscr{G}_k^h$, $k = 1, \ldots, K$, *assume* $\exists c_{ij} > 0$, *such that*

$$\text{trace}(P_{\boldsymbol{X}_{G_i}} P^{\perp}_{\boldsymbol{X}_{G_j}} + P^{\perp}_{\boldsymbol{X}_{G_i}} P_{\boldsymbol{X}_{G_j}}) \leq c_{ij}\lambda_1 \delta/\sqrt{s}$$

and

$$(1 - t_0)\lambda_1\delta - \sigma\lambda_1\delta c_{ij} \geq \sigma\sqrt{2\log[2\binom{|\mathscr{G}_k^h|}{2}K/\epsilon]/n}$$

3. *For* $j = J_K + 1, \ldots, J$, $\exists c_j > 0$ *such that*

$$\sqrt{\text{trace}(P_{\boldsymbol{X}_{G_j}} P^{\perp}_{\boldsymbol{X}_A})} \leq c_j\sqrt{p_j},$$

and

$$\lambda_1(1 - C_1) - \sigma c_j\sqrt{p_j} \geq \sigma\sqrt{2\log[2(J - J_k)/\epsilon]/n}.$$

Theorem 4.2 *Assume Conditions A4.1, B4.1 and B4.2 hold. Then we have*

$$\text{P}\big(\text{sgn}(\hat{\gamma}_i - \hat{\gamma}_j) = \text{sgn}(\gamma_i^0 - \gamma_j^0))\big) \geq 1 - 3\epsilon.$$

4.4 Computation

4.4.1 Computing algorithm

To find the Group OSCAR estimator

$$(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}) = \text{argmin}_{\boldsymbol{\gamma},\boldsymbol{\theta}}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda_1 \sum_{j=1}^{J} \gamma_j + \lambda_1\delta \sum_{i<j} \max(\gamma_i, \gamma_j),$$

$$\text{subject to} \quad \|\boldsymbol{X}_{G_j}\boldsymbol{\theta}_{G_j}\|_2^2 = n,$$

an iterative algorithm is applied here. More specifically, we first estimate θ's and find out the representative variable for each group. Then we fix the updated θ's and update γ's.

Since the objective function for Group OSCAR is convex, the optimal solution can be characterized by the subgradient equations. More specifically, for group G_j, we have

$$\boldsymbol{X}_{G_j}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n \propto \boldsymbol{X}_{G_j}^T\boldsymbol{X}_{G_j}\hat{\boldsymbol{\beta}}_{G_j}/\|\boldsymbol{X}_{G_j}\hat{\boldsymbol{\beta}}_{G_j}\|_2.$$

We are interested in the normalized group representative variable $\boldsymbol{z}_j = \boldsymbol{X}_{G_j}\boldsymbol{\theta}_{G_j} \propto \boldsymbol{X}_{G_j}\boldsymbol{\beta}_{G_j}$, according to the formula above, we have

$$\boldsymbol{X}_{G_j}\hat{\boldsymbol{\beta}}_{G_j} \propto \boldsymbol{X}_{G_j}(\boldsymbol{X}_{G_j}^T\boldsymbol{X}_{G_j})^{-1}\boldsymbol{X}_{G_j}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n, j = 1\ldots, J.$$

Therefore,

$$\hat{\boldsymbol{z}}_j = \sqrt{n}\frac{P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}}{\|P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}\|_2},$$

where $P_{\boldsymbol{X}_{G_j}} = \boldsymbol{X}_{G_j}(\boldsymbol{X}_{G_j}^T\boldsymbol{X}_{G_j})^{-1}\boldsymbol{X}_{G_j}^T$ is the the projection matrix on to the space spanned by $\boldsymbol{X}_{G_j}$ and $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the residual.

The detailed algorithm is

a. Initialization: $m = 1$, set $\gamma_j^{(0)} = 0, j = 1, ..., J$ and $\boldsymbol{r}^{(0)} = \boldsymbol{y}$.

b. Iteration: during the m^{th} iteration

Step 1: Find out the representative variable for each group:

$$\boldsymbol{z}_j^{(m)} = \sqrt{n}\frac{P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}^{(m-1)}}{\|P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}^{(m-1)}\|_2}, j = 1, \ldots, J$$

Step 2: Update γ's:

$$\hat{\boldsymbol{\gamma}}^{(m)} = \text{argmin}_{\boldsymbol{\gamma}} \|\boldsymbol{y} - \sum_{j=1}^{J} \boldsymbol{z}_j^{(m)} \gamma_j\|_2^2/(2n) + \lambda_1 \sum_{j=1}^{J} \gamma_j + \lambda_1 \delta \sum_{i<j} \max(\gamma_i, \gamma_j),$$

and

$$\boldsymbol{r}^{(m)} = \boldsymbol{y} - \sum_{j=1}^{J} \boldsymbol{z}_j^{(m)} \gamma_j^{(m)}.$$

m=m+1

c. Repeat part b. until convergence

4.4.2 Choosing the tuning parameters

The selection of tuning parameters (λ, δ) can be done by cross-validation. Since there are two tuning parameters, we first pick a grid of values for the grouping tuning parameter δ. For each δ, a 5-fold or 10-fold cross validation is applied. Suppose the dataset is divided into N partitions: $(y_1, y_2, \ldots, y_n) \overset{\text{d}}{=} (D_1, D_2, ..., D_N)$. Denote $\boldsymbol{\beta}(D_i)$ is the estimated coefficients without data D_i. Then the cross validation criterion is formed as:

$$\text{CV}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{j \in D_i^c} \{y_j - \boldsymbol{x}_j \boldsymbol{\beta}(D_i)\}^2.$$

Tuning parameters (λ, δ) which minimize the criterion $\text{CV}(\boldsymbol{\beta})$ are selected. However, the cross validation approach is computationally intensive, especially when the data set is large. Alternative techniques are AIC, BIC criteria or generalized cross validation score.

In addition, $(\hat{\theta}_{kj} = 0, j = 1, ..., p_k) \Leftrightarrow \hat{\gamma}_k = 0$. Then, according to Zou et al. (2007), we can approximate the number of effective parameters for a given λ by $d(\lambda) = |\mathscr{B}(\lambda)|$, where $\mathscr{B}(\lambda)$ is the active set in the Lasso regression step for θ's.

An AIC-type criterion is:

$$\text{AIC}(\lambda) = \log(\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda)\|_2^2/n) + 2d(\lambda)/n.$$

An BIC-type criterion is:

$$\text{BIC}(\lambda) = \log(\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda)\|_2^2/n) + \log(n)d(\lambda)/n.$$

A generalized cross validation score is:

$$\text{GCV}(\lambda) = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda)\|_2^2/\{n(1 - d(\lambda)/n)2\}.$$

For a given δ, the selection of tuning parameter λ is chosen to minimize the AIC(λ), BIC(λ) or GCV(λ). Usually, the AIC(λ) tends to select more variables while BIC(λ) performs better in shrinkage.

4.5 Numerical Studies

4.5.1 Simulation study

We use simulations to evaluate the performance of the proposed estimator. To demonstrate the grouping property, two scenarios are considered. For generating models in Example 4.1, correlations of some variables on within-group level are relatively large and so are group correlations of some groups in the model. Only highly correlated groups have nonzero coefficients. In Example 4.2, the correlations of variables within the same group are moderate but correlations between some groups in the model are relatively large. However, we have both highly correlated groups and independent groups have nonzero coefficients. In addition, Example 4.3 works as a control example in which correlations of variables on both within-group and between-group levels are small or moderate. Simulation setting details are provided below.

Example 4.1 There are 5 groups with 15 variables and each group consists of 3 variables. Firstly, we generate 6 variables for group 1 and group 2, $\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{13}$ and $\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{23}$ from standard normal distributions with covariance $cov(\boldsymbol{x}_{ki}, \boldsymbol{x}_{k'j}) = 0.9^{|i-j|}, k, k' = 1, 2, i, j = 1, \ldots, 3$ and $\boldsymbol{X}_{G_1} = \boldsymbol{X}_{G_2}$. Then 9 variables are generated from standard normal distribution with covariance $cov(\boldsymbol{x}_{ki}, \boldsymbol{x}_{k'j}) = 0.6^{|k-k'|} * 0.9^{|i-j|}, k, k' = 3, 4, 5, i, j = 1, \ldots, 3$ for group 3, group 4 and group 5. Thus, G_1 and G_2 are highly correlated with group correlation 0.95. But they are independent of the other three groups G_3, G_4 and G_5. Groups G_3, G_4 and G_5 are also moderately correlated with each other as well. Moreover, the within-group correlations in each group could be as large as 0.9.

We assume the true model for the response $\boldsymbol{y}$ is

$$\boldsymbol{y} = \sum_{k=1}^{2} \sum_{i=1}^{p_k} \boldsymbol{x}_{ki} \beta_{ki} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}$ are independently generated from $N(0, 1)$ and $\boldsymbol{\beta}^0_{G_1} = \boldsymbol{\beta}^0_{G_2} = (0.3, 0.3, 0.3)$, with all other $\beta^0_{kj} = 0$. Or, alternatively, the true model with only the significant covariates is

$$\boldsymbol{y} = \boldsymbol{X}_{G_1} \boldsymbol{\beta}^0_{G_2} + \boldsymbol{X}_{G_2} \boldsymbol{\beta}^0_{G_2} + \boldsymbol{\varepsilon}.$$

In each simulation, we generate $n = 110$ response observations from this model. The first 100 observations are used to fit a regression model, and the last 10 observations are used to evaluate the predictions the fitted model. This simulation is repeated 200 times.

Example 4.2 There are 8 groups with 40 variables and each group consists of 5 variables. For G_1 and G_2, we first generate 10 variables $(z_{11}, z_{21}), \ldots, (z_{15}, z_{25})$ from bivariate normal distribution with correlation=0.95. For the $G_3, \ldots, G_8$, $\boldsymbol{z}_1, ..., \boldsymbol{z}_6$ vectors

are generated from the standard normal distribution with $cov(z_k, z_{k'}) = 0.7^{|k-k'|}, k, k' = 3, \ldots, 8$ and let $z_{kj} = z_k, j = 1, \ldots, 5$. Then we generate $w_{k1}, \ldots, w_{k5}, k = 1, \ldots, 8$ vectors from standard normal distribution independently .

Then The explanatory variables are obtained by $\boldsymbol{x}_{kj} = 0.7 \times w_{kj} + 0.3 \times z_{kj}$. Thus, G_1 and G_2 are highly correlated with correlation around 0.9 and the absolute correlations among $G_3, \ldots, G_8$ range from 0 to 0.2. However, the within-group correlations that are in the range of [0,0.7] are moderate.

The response vector with $n = 110$ observations is generated by

$$\boldsymbol{y} = \sum_{k=1}^{8} \sum_{i=1}^{p_k} \boldsymbol{x}_{ki} \beta_{ki}^0 + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}$ are independently generated from $N(0, 1)$ and $\boldsymbol{\beta}_{G_1}^0 = \boldsymbol{\beta}_{G_2}^0 = (0.3, \ldots, 0.3)$, $\boldsymbol{\beta}_{G_3}^0 = (0.5, \ldots, 0.5)$ with all other $\beta_{kj} = 0$. Again, the first 100 observations are used to fit a regression model and the last 50 observations are used to evaluate the prediction of the fitted model. This simulation is repeated 200 times.

Example 4.3 There are 4 groups with 40 variables and each group consists of 10 variables. For group 1 and group 2, 10 variables are generated from $N(0, 1)$ with covariance $cov(\boldsymbol{x}_{ki}, \boldsymbol{x}_{kj}) = 0.5^{|i-j|}, k = 1, 2$. In other groups, variables are generated from $N(0, 1)$ independently of each other. Thus, between-groups correlations are small which do not exceed 0.15 and the within-groups correlations are also moderate which range from 0 to 0.6.

In this example, we assume the true model for the responses is

$$\boldsymbol{y} = \sum_{k=1}^{4} \sum_{i=1}^{p_k} \boldsymbol{x}_{ki} \beta_{ki} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}$ are independently generated from $N(0, 1)$ and $\boldsymbol{\beta}_{\mathscr{G}_1} = \boldsymbol{\beta}_{\mathscr{G}_2} = (0.3, 0.5, 0.7, 0.9, 0, \ldots, 0)$

Table 4.1: Simulation: Frequency (%) of occasions on which exact true groups are selected, group sensitivity, group specificity, number of groups selected and prediction error (PE) over 200 replications (with standard deviation in parentheses).

	% Group Sensitivity	% Group Specificity	# Groups	PE
		Example 1		
Lasso	11.00(2.08)	96.33 (1.24)	0.33 (0.60)	1.00 (0.45)
GLasso	32.00(2.51)	77.67 (2.22)	1.31 (0.53)	1.00 (0.44)
Gbridge	39.50 (2.05)	42.33 (2.50)	2.52 (0.85)	1.02 (0.46)
GOSCAR	78.00 (4.16)	57.00 (2.65)	2.85 (0.90)	1.01 (0.44)
		Example 2		
Lasso	3.00 (0.96)	91.40 (1.66)	0.52 (0.99)	1.12 (0.50)
GLasso	11.00 (1.64)	77.80 (1.33)	1.44 (0.66)	1.12 (0.50)
Gbridge	51.33 (1.98)	41.00 (2.23)	4.49 (1.19)	1.18 (0.62)
GOSCAR	40.33 (12.24)	77.60 (1.95)	2.33 (0.94)	1.11 (0.49)
		Example 3		
Lasso	100 (0)	52.25 (35.55)	2.96 (0.71)	1.22 (0.27)
GLasso	100 (0)	97.00 (11.90)	2.06 (0.24)	1.29 (0.29)
Gbridge	100 (0)	86.00 (24.12)	2.28 (0.48)	1.16 (0.24)
GOSCAR	100 (0)	100 (0)	2.00 (0)	1.14 (0.24)

with all other $\beta_{kj} = 0$. Thus, the true model with only the significant covariates is

$$\boldsymbol{y} = \sum_{j=1}^{4} \boldsymbol{x}_{1j}\beta_{1j} + \sum_{j=1}^{4} \boldsymbol{x}_{2j}\beta_{2j} + \boldsymbol{\varepsilon}$$

We generate $n = 150$ response observations from the model. The first 100 observations are used to fit a regression model, and the last 50 observations are used to evaluate the prediction performance of the fitted model. This simulation is also repeated 200 times

In order to demonstrate the performance of our proposed method, we also analyze all simulated data with Lasso, Group Lasso and Group OSCAR methods. The tuning parameters are chosen by BIC criteria for all the methods. The results are compared in terms of group selection. In the group selection, if any variable in a certain group is selected, we consider this group is selected regarding two-level selection property. We calculate the frequencies of occasions on which only true groups are selected. To illustrate the selection errors, we define group selection sensitivity and group selection

specificity. Group sensitivity is defined as the ratio of number of selected groups that are truly in the model to the total number of true groups in the model. Group specificity is defined as the ratio of number of groups that are excluded and not in the model to the total number of groups that are not in the model. Here, in the performance evaluation, the calculation is based on the known groups.

Furthermore, we also report the average numbers of selected groups and the average numbers of selected variables to illustrate model complexity. Since Lasso does not have grouping property, the numbers of selected groups are obtained based on the prior known group structure G to make the results comparable. The prediction error for the holdout $m = 10$ observations in each simulation is defined as $\sum_{i=1}^{m}(y_i - (\hat{y}_i))^2/m$, where $\hat{y}_i$ is the fitted value that is obtained from the selected model with least square estimates.

It can be seen from Table 4.1 that when high correlations exist among explanatory variables, weighted group ridge performs better than any other methods in the sense of group selection and prediction errors. In fact, if both between-group and within-group correlations are large, Group OSCAR has better performance than other methods. Lasso would randomly pick among highly correlated variables which explains its low variable sensitivity. It also tends to select more groups as their selected variables are distributed in different groups with low group specificity. This is because Lasso is designed to select individual variables and ignore the information contained in the prior group structure. Group Lasso and Group OSCAR take the prior group structures into account. Therefore, they achieve good results in terms of group selection.

4.5.2 Manifest data analysis

We analyze the manifest data to demonstrate the application of the penalized least square estimator. The denotation of a nuclear weapon on the U.S. soil is among the most dangerous types of terrorist attacks. Standardized shipping containers, which transport 95% of the U.S. imports by tonnage, are highly vulnerable vehicles for delivering nuclear and radiological weapons. The cost of an exploded bomb at a major U.S. shipping port has been estimated to be a trillion dollars. To counter the potential threat, substantial efforts have been made in devising strategies for inspecting containers and interdicting illicit nuclear materials. Practical issues and challenges exist in carrying out this important task. Due to the enormous size of traffic and a large number of entry sites, there are now 307 ports of entry representing 621 official air, sea and land border crossing sites. Entering these ports everyday via $57,000$ trucks, $2,500$ aircrafts, and 580 sea vessels are the cargoes that much of this country's commercial life depends on. As the result, we have two competing priorities in the inspecting process. On one hand, we must detect any illicit nuclear materials to safeguard the national security. On the other hand, we need to move the cargoes as fast possible from their port of entries to reduce the waiting cost.

It is important to make quantitative evaluations of manifest data. We address this problem by constructing a linear regression model to enhance the effectiveness of the real-time inspection system. Our primary goal is to identify high-risk shipment, i.e. get the risk score for each shipment and determine the effects of different sources of information in the manifest data. Table 2.3 provides the definition and description of the variables contained in the manifest data. These seven categorical variables cannot be used directly in the linear regression and dummy variables for each categorical variable

are created which results in $p = 216$ variables in total. These dummy variables naturally form a group. Moreover, some of these categorical variables are highly correlated which can be reflected by high correlations of dummy variables among these groups. The response variable is the risk score for each shipment. Here, we focus on the data on February 28, 2009 with 24373 shipments.

However, the risk scores of the 24373 shipments are not accessible, due to security concerns. To illustrate our approach, potential influential characteristics are selected to generate the risk scores. Thus, it is possible to test our approach in identifying risk factors. We treat the knowledge based on the representation of different levels for categorical variables as explicit group structure while the correlations patterns are implicit group information. For further evaluation, we also select 24000 observations as the training set and then predict the risk score for future shipments. Only information about $\mathscr{G}$ is known. The group labels used are:

$$(\underbrace{1,\ldots,1}_{|\mathscr{G}_1|=8},\underbrace{2,\ldots,2}_{|\mathscr{G}_2|=68},\underbrace{3,\ldots,3}_{|\mathscr{G}_3|=8},\underbrace{4,\ldots,4}_{\mathscr{G}_4|=13},\underbrace{5,\ldots,5}_{|\mathscr{G}_5|=69},\underbrace{6,\ldots,6}_{|\mathscr{G}_6|=34},\underbrace{7,\ldots,7}_{|\mathscr{G}_7|=16})$$

Example 4.4 The risk scores are generated by

$$\boldsymbol{y} = \sum_{k=1}^{7}\sum_{i=1}^{p_k} \boldsymbol{x}_{ki}\beta_{ki} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}$ are independently generated from $N(0,1)$. The coefficients are $\boldsymbol{\beta}_{\mathscr{G}_1} = (0,\ldots,0)$, $\boldsymbol{\beta}_{\mathscr{G}_2} = (\underbrace{0.3,\ldots,0.3}_{13},0\ldots,0)$ $\boldsymbol{\beta}_{\mathscr{G}_3} = (0,\ldots,0)$, $\boldsymbol{\beta}_{\mathscr{G}_4} = (0,\ldots,0)$, $\boldsymbol{\beta}_{\mathscr{G}_5} = (0,\ldots,0)$, $\boldsymbol{\beta}_{\mathscr{G}_6} = (\underbrace{0.3,\ldots,0.3}_{13},0\ldots,0)$, $\boldsymbol{\beta}_{\mathscr{G}_7} = (\underbrace{0,\ldots,0}_{8},\underbrace{0.8,\ldots,0.8}_{8})$

Similar to simulation study, we also compare our method with Lasso. OSCAR is eliminated in the real data analysis because of its extremely computational burden. Here, we still use least square estimates with the selected model to predict the risk score for the future 373 shipments. The results displayed in Table 4.2 confirm what

Table 4.2: Simulation: Frequency (%) of occasions on which exact true groups are selected, group sensitivity, group specificity, number of groups selected and prediction error (PE) over 100 replications (with standard deviation in parentheses).

	% Group Sensitivity	% Group Specificity	# Groups	PE
Lasso	96.00 (1.97)	16.75 (2.19)	6.21 (1.39)	0.96 (0.21)
GLasso	66.67 (0)	30.75 (1.06)	4.77 (0.42)	1.03 (0.06)
Gbridge	96.00 (1.97)	9.25 (2.15)	6.51 (1.40)	1.38 (0.31)
GOSCAR	97.83 (8.24)	100.00 (0)	2.94 (0.25)	0.99 (0.08)

we have found in the simulation studies. By incorporating the correlation patterns and group structures, Group OSCAR beats the other two methods in terms of group selection and prediction errors. Lasso is bad at group selection since it always include less groups.

4.6 Discussion

In this chapter, we propose a new penalized regression model which performs model selection and variable grouping functions simultaneously. It explicitly incorporate the correlation structures in the explanatory variables. The definition of correlation among groups is important in the general formulation of our estimator defined in (4.1). As a specific model, the L_ν norm for groupwise coefficients are used in this paper. However, it is possible to generalize the correlation definition by establishing other connections in explanatory variables.

The proposed penalized regression approach can also be applied to other regression models when there exists strong correlation patterns in the explanatory variables, such as general linear regression models or Cox regression. the penalized regression criterion

can be written as:

$$\widehat{\boldsymbol{\beta}} = \text{argmin}\boldsymbol{\beta}\frac{1}{2n}l(y_i;\sum_{j=1}^{p}\boldsymbol{x}_i\beta_j) + \lambda_1\rho_1(\boldsymbol{\beta}_d G_j, j = 1, ..., J) + \lambda_2\sum_{i\neq j}\rho_2(c(G_i, G_j))$$

where l is a given loss function ρ_1, ρ_2 are penalty functions and $c(G_i, rG_j)$ denotes the correlations between G_i and G_j. Although for these general models, new computational algorithms need to developed.

Another advantage of Group OSCAR over regular OSCAR methods is that since the number of groups is much smaller than the number of variables with nonzero coefficients, Group OSCAR can reduce the computing burden a lot.

4.7 Appendix

Proof of Theorem 4.1:

Denote r_k and $r_{k'}$ are the ranks of $\hat{\gamma}_k$ and $\hat{\gamma}_{k'}$ respectively, i.e. $r_k = \#\{\hat{\gamma}_l : \hat{\gamma}_l \leq \hat{\gamma}_k\}$ and $r_{k'} = \#\{\hat{\gamma}_l : \hat{\gamma}_l \leq \hat{\gamma}_{k'}\}$.

Now suppose $\hat{\gamma}_k \neq \hat{\gamma}_{k'}$. Then $r_k \neq r_{k'}$. Since both $\hat{\gamma}_k > 0$ and $\hat{\gamma}_{k'} > 0$ and they are different from the other $\hat{\gamma}$'s, given $(\hat{\theta}_{kj})$, we can differentiate the penalized least square function and obtain

$$-\frac{1}{n}\hat{\boldsymbol{x}}_k^T(\boldsymbol{y} - \hat{\boldsymbol{X}}\hat{\boldsymbol{\gamma}}) + \lambda_1\{\delta(r_k - 1) + 1\} = 0,$$

$$-\frac{1}{n}\hat{\boldsymbol{x}}_{k'}^T(\boldsymbol{y} - \hat{\boldsymbol{X}}\hat{\boldsymbol{\gamma}}) + \lambda_1\{\delta(r_{k'} - 1) + 1\} = 0,$$

where $\hat{\boldsymbol{x}}_k = \sum_{i=1}^{p_k}\hat{\theta}_{ki}\boldsymbol{x}_{ki}, k = 1, \ldots, K$ and $\hat{\boldsymbol{X}} = (\hat{\boldsymbol{x}}_k, k = 1, \ldots, K)$.

Therefore

$$\begin{aligned} & |\lambda_1\delta(r_k - r_{k'})| \\ \leq\ & |\frac{1}{n}(\hat{\boldsymbol{x}}'_k - \hat{\boldsymbol{x}}'_{k'})(\boldsymbol{y} - \hat{\boldsymbol{X}}\hat{\boldsymbol{\gamma}})| \\ \leq\ & \frac{1}{n}\|(\hat{\boldsymbol{x}}'_k - \hat{\boldsymbol{x}}'_{k'}\|_2\|\boldsymbol{y} - \hat{\boldsymbol{X}}\hat{\boldsymbol{\gamma}}\|_2 \\ \leq\ & \frac{1}{n}\|(\hat{\boldsymbol{x}}'_k - \hat{\boldsymbol{x}}'_{k'})\|_2\|\boldsymbol{y}\|_2 \\ \leq\ & \sqrt{2(1-\phi_{kk'})}\|\boldsymbol{y}\|_2/\sqrt{n}. \end{aligned}$$

Therefore, when $\lambda_1\delta > \sqrt{2(1-\hat{\phi}_{kk'})}(\|\boldsymbol{y}\|_2/\sqrt{n}$, contradiction occurs. We must have $\hat{\gamma}_k = \hat{\gamma}_{k'}$. □

Lemma 4.2 *When $\lambda_1 = \delta = 0$, the proposed method is equivalent to Group Lasso:*

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda_2\sum_{k=1}^{K}\|\boldsymbol{X}_k\boldsymbol{\beta}_k\|_2/\sqrt{n}. \tag{4.5}$$

And $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{\mathscr{A}}, \boldsymbol{0})$ if and only if

(i) for $1 \leq k \leq s$, $(\boldsymbol{X}_{\mathscr{A}}^T\boldsymbol{X}_{\mathscr{A}}/n)(\hat{\boldsymbol{\beta}}_{\mathscr{A}} - \boldsymbol{\beta}^0_{\mathscr{A}}) = \boldsymbol{X}_{\mathscr{A}}^T\varepsilon/n - \lambda_2\bar{P}(\hat{\boldsymbol{\beta}}_{\mathscr{A}})$,

(ii) for $s+1 \leq k \leq K$, $\|-\boldsymbol{X}_k^T\boldsymbol{y}/n + \boldsymbol{X}_k^T\boldsymbol{X}_{\mathscr{A}}\hat{\boldsymbol{\beta}}_{\mathscr{A}}\|_{k^*} \leq \lambda_2/\sqrt{n}$,

where $\bar{P}(\hat{\boldsymbol{\beta}}_k) = \boldsymbol{X}_k^T\boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k/(\|\boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k\|_2\sqrt{n})$, $\|\cdot\|_{k^*}$ is the dual norm of $\|\boldsymbol{u}\|_k \equiv \|\boldsymbol{X}_k\boldsymbol{u}\|_2$ and $\|\boldsymbol{u}\|_{k^*} = \|\boldsymbol{X}_k(\boldsymbol{X}_k^T\boldsymbol{X}_k)^{-1}\boldsymbol{u}\|_2$.

Proof: When $\delta = 0$, it is Group Lasso estimator. We obtain the KKT conditions for $\delta = 0$ first. Introduce slack variables v_j such that $\|\boldsymbol{X}_j\boldsymbol{\beta}_j\|_2 \leq v_j$, $j = 1, \ldots, J$. Thus, (4.5) is equivalent to

$$\min_{\boldsymbol{\beta},\boldsymbol{v}}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda_1\sum_{j=1}^{J}v_j/\sqrt{n},$$

with constraints $\|\boldsymbol{X}_j\boldsymbol{\beta}_j\|_2 \leq v_j$. Or equivalently, $(\boldsymbol{\beta}_j, v_j) \in \boldsymbol{C}_j$, $\boldsymbol{C}_j = \{(x,t) : \|x\|_j \leq t\}$.

According to Bach (2008) and Bach et al. (2011), the Lagrangian dual problem with dual variables $\boldsymbol{\eta}_1 = (\boldsymbol{\eta}_{1j}, j = 1, \ldots, J)$ and $\boldsymbol{\eta}_2 = (\eta_{2j}, j = 1, \ldots, J)$ is

$$\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\eta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda_2 \sum_{j=1}^{J} v_j/\sqrt{n} - \sum_{j=1}^{J}(\boldsymbol{\eta}_{1j}^T\boldsymbol{\beta}_j + \eta_{2j}v_j),$$

where $(\boldsymbol{\eta}_{1j}, \eta_{2j}) \in \boldsymbol{C}_{j^*}$, $\boldsymbol{C}_{j^*}$ is the dual cone of $\boldsymbol{C}_j$.

The derivatives with respect to primal variables are

$$\bigtriangledown_{\boldsymbol{\beta}}\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\eta}) = -\boldsymbol{X}^T\boldsymbol{y}/n + \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}/n - \boldsymbol{\eta}_1,$$

and

$$\bigtriangledown_{\boldsymbol{v}}\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\eta}) = \lambda_2/\sqrt{n} - \boldsymbol{\eta}_2.$$

The KKT conditions for reduced variables $\boldsymbol{\beta}$ and $\boldsymbol{\eta}_1$ are

$\forall k \;\; \|\eta_{1j}\|_{j^*} \leq \lambda_2/\sqrt{n}$

$\forall k \;\; -\boldsymbol{X}_j^T\boldsymbol{y}/n + \boldsymbol{X}_j^T\boldsymbol{X}\boldsymbol{\beta}/n = \boldsymbol{\eta}_{1j}$

$\forall k \;\; \boldsymbol{\eta}_{1j}^T\boldsymbol{\beta}_j + \|\boldsymbol{X}_k\boldsymbol{\beta}_j\|_2\lambda_2/\sqrt{n}=0$

The last equation is satisfied if $\boldsymbol{\beta}_j = \boldsymbol{0}$ or $\boldsymbol{\eta}_{1j} = -\lambda_2\boldsymbol{X}_j^T\boldsymbol{X}_j\boldsymbol{\beta}_j/(\|\boldsymbol{X}_j\boldsymbol{\beta}_j\|_2\sqrt{n})$. Together with the first and second KKT condition, the lemma is proved for $\delta = 0$. □

Proof of Lemma 4.1:

When $\delta \neq 0$, the weights $\tilde{\omega}_j$ will be different for each group coefficients $\|\boldsymbol{X}_{G_j}\boldsymbol{\beta}_{G_j}\|_2/\sqrt{n}$ rather than 1. Together with the subgradients of the L_∞ norm in Chapter 3, the results follow immediately. □

To prove theorem 4.2, we first define Oracle Group OSCAR estimator $\hat{\boldsymbol{\beta}}^{\mathcal{O}}$ as following: $\hat{\boldsymbol{\beta}}^{\mathcal{O}}_{A^c} = 0$ and

$$\hat{\boldsymbol{\beta}}_A = \text{argmin}_{\boldsymbol{\beta}_A} \|\boldsymbol{y} - \boldsymbol{X}_A\boldsymbol{\beta}_A\|_2^2 + \lambda \sum_{k=1}^{K} \sum_{j \in \mathscr{G}_k^h} \bar{\omega}_k \|\boldsymbol{X}_{G_j}\boldsymbol{\beta}_{G_j}\|_2/\sqrt{n},$$

subject to $\|\boldsymbol{X}_{G_i}\boldsymbol{\beta}_{G_i}\|_2 = \|\boldsymbol{X}_{G_j}\boldsymbol{\beta}_{G_j}\|_2$ for $i, j \in \mathscr{G}_k^h$, $k = 1, \ldots, K$.

We can see that $\hat{\boldsymbol{\beta}}^{\mathcal{O}}_A$ is the converging point of the Oracle Algorithm:

a. Initialization: $m = 1$, set $\gamma_j^{(0)} = 0, j = 1, ..., J_K$ and $\boldsymbol{r}^{(0)} = \boldsymbol{y}$.

b. Iteration: during the m^{th} iteration

Step 1: Find out the representative variable for each group:

$$\boldsymbol{z}_j^{(m)} = \sqrt{n}\frac{P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}^{(m-1)}}{\|P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}^{(m-1)}\|_2}, j = 1, \ldots, J$$

and let

$$\boldsymbol{f}_k^{(m)} = \sum_{j \in \mathscr{G}_k^h} \boldsymbol{z}_j^{(m)}, k = 1, \ldots, K.$$

Step 2: Update γ's. Let $\boldsymbol{F}^{(m)} = (\boldsymbol{f}_k^{(m)})$ and

$$\boldsymbol{b}^{(m)} = [(\boldsymbol{F}^{(m)})^T\boldsymbol{F}^{(m)}]^{-1}[(\boldsymbol{F}^{(m)})^T\boldsymbol{y} + \lambda_2\bar{\boldsymbol{\omega}}],$$

and $\gamma_j^{(m)} = b_k^{(m)}$, $k = 1, \ldots, K$. Then,

$$\boldsymbol{r}^{(m)} = \boldsymbol{y} - \sum_{j=1}^{J} \boldsymbol{z}_j^{(m)}\gamma_j^{(m)}.$$

m=m+1

c. Repeat part b. until convergence

We first find out how close true representing variables and estimated representing variables are for each group. The results are stated in the following lemma.

Lemma 4.3 *Let* $\hat{\Delta}_A = \hat{\boldsymbol{\beta}}_A^{\mathcal{O}} - \boldsymbol{\beta}_A^0$. *Under the event* $\{\lambda\bar{\omega}_k > \|\boldsymbol{X}_{G_j}(\boldsymbol{X}_{G_j}^T\boldsymbol{X}_{G_j})^{-1}\boldsymbol{X}_{G_j}^T\varepsilon\|_2/\sqrt{n}, j \in \mathscr{G}_k^h, k = 1,\ldots,K\}$, *we have*

$$\|\boldsymbol{X}_{G_j}\hat{\Delta}_{G_j}^{\mathcal{O}}\|_2/\sqrt{n} \le \bar{\omega}_k/\kappa_k/2 + (1/\kappa_k)\sqrt{(\lambda\bar{\omega}_k)^2/4 + 2\kappa_k\lambda\bar{\omega}_k\gamma_i^0}, j \in \mathscr{G}_k^h, k = 1,\ldots,K$$

and thus

$$\max_{j\in\mathscr{G}_k^h} \|\boldsymbol{X}_{G_j}\hat{\Delta}_{G_j}^{\mathcal{O}}\|_2/\|\boldsymbol{X}_{G_j}\boldsymbol{\beta}_{G_j}^0\|_2 \le c_k\sqrt{\lambda\bar{\omega}}, k = 1,\ldots,K.$$

Proof: For any $\Delta_A \in \mathbb{R}^s$, define

$$F(\Delta_A) = \|\boldsymbol{X}_A\Delta_A\|_2^2 - \varepsilon^T\boldsymbol{X}_A\Delta_A/n + \lambda_1\sum_{k=1}^{K}\sum_{j\in\mathscr{G}_h^k}(\|\boldsymbol{X}_{G_j}(\boldsymbol{\beta}_{G_j}^0 + \Delta_{G_j}^0)\|_2 - \|\boldsymbol{X}_{G_j}\boldsymbol{\beta}_{G_j}^0\|_2)/\sqrt{n}.$$

Consider the cone

$$\mathcal{C} = \{\Delta_A : \|\boldsymbol{X}_{G_j}\Delta_{G_j}\|_2 \le \bar{\omega}_k/\kappa_k/2 + (1/\kappa_k)\sqrt{(\lambda\bar{\omega}_k)^2/4 + 2\kappa_k\lambda\bar{\omega}_k\gamma_i^0}, j \in \mathscr{G}_k^h, k = 1,\ldots,K\}.$$

We will show that $\hat{\Delta}_A \in \mathcal{C}$. If $\hat{\Delta}_A \notin \mathcal{C}$, the segment connecting $\hat{\Delta}$ and $\mathbf{0}$ will cross the boundary of $\mathcal{C}$, $\partial\mathcal{C}$. Then there exist $\tilde{\Delta}_A \in \partial\mathcal{C}$ such that $\tilde{\Delta}_A = t\hat{\Delta}_A$, $0 \le t \le 1$. And we have

$$F(\tilde{\Delta}_A) \le tF(\hat{\Delta}_A) + (1-t)F(\mathbf{0}_A) = tF(\hat{\Delta}_A) \le 0.$$

The last inequality is because by the definition of $\hat{\boldsymbol{\beta}}_A^{\mathcal{O}}$, $F(\hat{\Delta}_A) < 0$.

However, for any $\Delta_A \in \partial\mathcal{C}$, under the events

$$\{\lambda\bar{\omega}_k/2 > \|\boldsymbol{X}_{G_j}(\boldsymbol{X}_{G_j}^T\boldsymbol{X}_{G_j})^{-1}\boldsymbol{X}_{G_j}^T\varepsilon\|_2/\sqrt{n}, j \in \mathscr{G}_k^h, k = 1,\ldots,K\},$$

we have

$$
\begin{aligned}
& F(\Delta) \\
\geq\ & \sum_k \sum_j \kappa_k \|\boldsymbol{X}_{G_j}\Delta_{G_j}\|_2^2/(2n) - \sum_k \sum_j (\boldsymbol{X}_{G_j}^T\varepsilon)^T\Delta_{G_j}/n \\
& +\lambda_1 \sum_k \sum_j (\|\boldsymbol{X}_{G_j}(\boldsymbol{\beta}_{G_j}^0 + \Delta_{G_j}^0)\|_2 - \|\boldsymbol{X}_{G_j}\boldsymbol{\beta}_{G_j}^0\|_2)/\sqrt{n} \\
\geq\ & \sum_k \sum_j \kappa_k \|\boldsymbol{X}_{G_j}\Delta_{G_j}\|_2^2/(2n) - \sum_k \sum_j \|\boldsymbol{X}_{G_j}^T\varepsilon\|_{j*}\|\Delta_{G_j}\|_j/n - \lambda_1 \sum_k \sum_j \bar{\omega}_k \frac{\|\boldsymbol{X}_{G_j}\boldsymbol{\beta}_{G_j}^0\|_2}{\sqrt{n}} \\
>\ & \sum_k \sum_j [\kappa_k \|\boldsymbol{X}_{G_j}\Delta_{G_j}\|_2^2/(2n) - (\lambda_1\bar{\omega}_k/2)\|\boldsymbol{X}_{G_j}\Delta_{G_j}\|_2/\sqrt{n} - \lambda_1\bar{\omega}_k\|\boldsymbol{X}_{G_j}\boldsymbol{\beta}_{G_j}^0\|_2/\sqrt{n}] \\
\geq\ & 0
\end{aligned}
$$

when $\|\boldsymbol{X}_{G_j}\Delta_{G_j}\|_2/\sqrt{n}$ is large enough.

A simple calculation will give that $\bar{\omega}_k/\kappa_k/2 + (1/\kappa_k)\sqrt{(\lambda\bar{\omega}_k)^2/4 + 2\kappa_k\lambda\bar{\omega}_k\gamma_i^0}$ is sufficient. Contradiction. Then we have $\hat{\Delta}_A \in \mathcal{C}$. □

Proof of Theorem 4.2:

Consider the following events $E_1 = \{\lambda\bar{\omega}_k/2 > \|\boldsymbol{X}_{G_j}(\boldsymbol{X}_{G_j}^T\boldsymbol{X}_{G_j})^{-1}\boldsymbol{X}_{G_j}^T\varepsilon\|_2/\sqrt{n}, j \in \mathscr{G}_k^h, k = 1,\ldots,K\}$, $E_2 = \{\|(P_{\boldsymbol{X}_{G_i}} - P_{\boldsymbol{X}_{G_j}})P_{\boldsymbol{X}_A}\varepsilon/n\|_2 \leq (1-t_0)\lambda_1\delta/2\}$ and $E_3 = \{\|P_{\boldsymbol{X}_{G_j}}(I - P_{\boldsymbol{X}_A})\varepsilon/n\|_2 \leq \lambda_1(1-C_1), j = J_K+1,\ldots,J\}$.

According to Lemma 4.3, we have $|\hat{\gamma}_j^{\mathcal{O}} - \gamma_j^0| \leq \bar{\omega}_k/\kappa_k/2 + (1/\kappa_k)\sqrt{(\lambda\bar{\omega}_k)^2/4 + 2\kappa_k\lambda\bar{\omega}_k\gamma_i^0}$. Therefore, under Condition B4.1(1),

$$\operatorname{sgn}(\hat{\gamma}_i^{\mathcal{O}} - \hat{\gamma}_j^{\mathcal{O}}) = \operatorname{sgn}(\gamma_i^0 - \gamma_j^0).$$

Then, we will show that the Oracle estimator restricted on the subspace $\mathbb{R}^s$ is the Group OSCAR estimator. It is equivalent to show that

$$(\|P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}\|_2/\sqrt{n}, j \in \mathscr{G}_k^h) = \lambda_1\delta\boldsymbol{v}_{\mathscr{G}_k^h} + \lambda_1[(J - J_{k-1})\delta + 1]\mathbf{1}_{\mathscr{G}_k^h}$$

where $\boldsymbol{r} = (\boldsymbol{y} - \boldsymbol{X}_A\hat{\boldsymbol{\beta}}_A^{\mathcal{O}})/n$ and $\boldsymbol{v}_{\mathscr{G}_k^h} \in \mathcal{D}_{|\mathscr{G}_k^h|}$. By the definition of the Oracle Algorithm,

we know that

$$\sum_{j\in\mathscr{G}_k^h}(\|P_{\boldsymbol{X}_{G_j}}\|_2/\sqrt{n})/|\mathscr{G}_k^h| = \lambda_1\bar{\omega}_k.$$

According to Chapter 3 in this dissertation, we only need to show that for $m = 1,\ldots,|\mathscr{G}_k^h|$, $k = 1,\ldots,K$, $\forall B_k^m \subset \mathscr{G}_k$ with $|B_k^m| = m$

$$\Big|\sum_{j\in B_k^m}\|P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}\|_2/m - \sum_{j\in\mathscr{G}_k^h}\|P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}\|_2/|\mathscr{G}_k^h|\Big|/\sqrt{n} \le \lambda_1\delta(|\mathscr{G}_k^h| - m)/2.$$

Denote $\hat{\boldsymbol{z}}_j = \sqrt{n}\frac{P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}}{\|P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}\|_2}, j = 1,\ldots,J_K$, $\hat{\boldsymbol{f}}_k = \sum_{j\in\mathscr{G}_k^h}\hat{\boldsymbol{z}}_j, k = 1,\ldots,K$, and $\hat{\boldsymbol{F}} = (\hat{\boldsymbol{f}}_k, k = 1,\ldots,K)$. Then, we can rewrite the residual $\boldsymbol{r}$ as

$$\boldsymbol{r} = (I - P_{\hat{\boldsymbol{F}}})(\varepsilon/n + \boldsymbol{F}^0\gamma^0) - \lambda_1\hat{\boldsymbol{F}}(\hat{\boldsymbol{F}}^T\hat{\boldsymbol{F}}/n)^{-1}\bar{\boldsymbol{\omega}}.$$

Since the space spanned by $\hat{\boldsymbol{F}}$ is in the space spanned by $\boldsymbol{X}_A$, we have $P_{\hat{\boldsymbol{F}}}^{\perp} = P_{\boldsymbol{X}_A}^{\perp} + P_{\boldsymbol{X}_A} - P_{\hat{\boldsymbol{F}}}$. Therefore,

$$P_{\boldsymbol{X}_{G_j}}\boldsymbol{r} = P_{\boldsymbol{X}_{G_j}}(P_{\boldsymbol{X}_A} - P_{\hat{\boldsymbol{F}}})(\varepsilon/n + \boldsymbol{F}^0\boldsymbol{\gamma}^0/n) - \lambda_1 P_{\boldsymbol{X}_{G_j}}\hat{\boldsymbol{F}}(\hat{\boldsymbol{F}}^T\hat{\boldsymbol{F}}/n)^{-1}\bar{\boldsymbol{\omega}}.$$

Since

$$\begin{aligned}
&\Big|\sum_{j\in B_k^m}\|P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}\|_2/m - \sum_{j\in\mathscr{G}_k^h}\|P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}\|_2/|\mathscr{G}_k^h|\Big|/\sqrt{n}\\
\le\ &\sum_{j\in\mathscr{G}_k^h}\sum_{i\in B_k^m}(\|P_{\boldsymbol{X}_{G_i}}\boldsymbol{r}\|_2 - \|P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}\|_2)/(m|\mathscr{G}_k^h|),
\end{aligned}$$

we only need to show that

$$\begin{aligned}
&\big|\|P_{\boldsymbol{X}_{G_i}}\boldsymbol{r}\|_2 - \|P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}\|_2\big|\\
\le\ &\big|\|(P_{\boldsymbol{X}_{G_i}} - P_{\boldsymbol{X}_{G_j}})\boldsymbol{r}\|_2 \le \lambda_1\delta/2.
\end{aligned}$$

Under condition B4.1(2) and events E_2, the conclusion follows immediately.

Finally, we will show the condition for zero groups in Lemma 4.1 is satisfied. We have shown that $\hat{\boldsymbol{\beta}}_A^{\mathcal{O}}$ is the Group OSCAR estimator restricted on the nonzero groups, then

$$\boldsymbol{X}_A(\boldsymbol{y} - \boldsymbol{X}_A\boldsymbol{\beta}_A^{\mathcal{O}}) = \lambda_1\rho(\hat{\ }^{-}{}_A^{\mathcal{O}}\boldsymbol{\beta},$$

where $\bar{\rho}(\hat{\boldsymbol{\beta}}_A^{\mathcal{O}}) = (\omega_i \boldsymbol{X}_{G_i}^T \boldsymbol{X}_{G_i}\hat{\boldsymbol{\beta}}_{G_i}^{\mathcal{O}}/(\|\boldsymbol{X}_{G_i}\hat{\boldsymbol{\beta}}_{G_i}^{\mathcal{O}}\|_2\sqrt{n}), i \in \mathscr{G}_k^h, k = 1, \dot{)}$,and $\omega_{\mathscr{G}_k^h} = \delta \boldsymbol{v}_{\mathscr{G}_k^h} + [(J - J_{k-1})\delta + 1]\mathbf{1}_{\mathscr{G}_k^h}$ for some $\boldsymbol{v}_{\mathscr{G}_k^h} \in \mathcal{D}_{|\mathscr{G}_k^h|}$.

Then $\forall j = J_K + 1, \ldots, J$

$$\begin{aligned}
& \|P_{\boldsymbol{X}_{G_j}}\boldsymbol{r}\|_2/\sqrt{n} \\
= \; & \|P_{\boldsymbol{X}_{G_j}}(I - P_{\boldsymbol{X}_A})\varepsilon/n + \lambda_1 P_{\boldsymbol{X}_{G_j}}\boldsymbol{X}_A(\boldsymbol{X}_A^T\boldsymbol{X}_A)^{+}\bar{\rho}(\hat{\boldsymbol{\beta}}_A^{\mathcal{O}})\|_2/\sqrt{n} \\
\leq \; & \|P_{\boldsymbol{X}_{G_j}}(I - P_{\boldsymbol{X}_A})\varepsilon/n\|_2/\sqrt{n} + \lambda_1 C_1 \\
\leq \; & \lambda_1.
\end{aligned}$$

It completes the proof that the Oracle estimator is the Group OSCAR estimator. Now we will compute the probability of the three events.

For event E_1, denote $\tilde{\varepsilon}_j = P_{\boldsymbol{X}_{G_j}}$. Then $\tilde{\varepsilon}_j \sim N(\mathbf{0}, \sigma(\boldsymbol{X}_{G_j}^T\boldsymbol{X}_{G_j}))$ and

$$E(\|\tilde{\varepsilon}_j\|_2) \leq \sqrt{\sum_{d=1}^{p_j}\Sigma_{G_j}^{-1}(d,d)} \leq c_k\sqrt{p_j} \leq c_k\sqrt{p_k^*}.$$

Then

$$\begin{aligned}
\mathrm{P}(E_1^c) \; & \leq \; \sum_{k=1}^{K}\sum_{j\in\mathscr{G}_k^h}\mathrm{P}(\|\tilde{\varepsilon}_j\|_2 \geq \lambda\bar{\omega}_k) \\
& \leq \; \sum_{k=1}^{K}\sum_{j\in\mathscr{G}_k^h} 2\exp\{-n(\lambda\bar{\omega}_k - c_k\sqrt{p_k^*})^2/2\} \\
& \leq \; \epsilon.
\end{aligned}$$

For event E_2, $(P_{\boldsymbol{X}_{G_i}} - P_{\boldsymbol{X}_{G_j}})P_{\boldsymbol{X}_A}\varepsilon \sim N(\mathbf{0}, \sigma\Sigma_{ij})$, where $\Sigma_{ij} = P_{\boldsymbol{X}_A}(P_{\boldsymbol{X}_{G_i}}P_{\boldsymbol{X}_{G_j}}^{\perp} +$

$P^{\perp}_{\boldsymbol{X}_{G_i}} P_{\boldsymbol{X}_{G_j}})P_{\boldsymbol{X}_A}$. Similar to event E_1,

$$\begin{aligned} & E(\|(P_{\boldsymbol{X}_{G_i}} - P_{\boldsymbol{X}_{G_j}})P_{\boldsymbol{X}_A}\varepsilon\|_2) \le \sqrt{\text{trace}(\Sigma_{ij})} \\ \le & \sqrt{\text{trace}(P_{\boldsymbol{X}_A})\text{trace}(P_{\boldsymbol{X}_{G_i}}P^{\perp}_{\boldsymbol{X}_{G_j}} + P^{\perp}_{\boldsymbol{X}_{G_i}}P_{\boldsymbol{X}_{G_j}})} \\ \le & \sqrt{s}\lambda_1\delta c_{ij}/\sqrt{s} \\ \le & \lambda_1\delta c_{ij}. \end{aligned}$$

Then

$$\begin{aligned} \mathrm{P}(E_2^c) & \le \sum_{k=1}^{K}\sum_{i<j,i,j\in\mathscr{G}_k^h} \mathrm{P}(\|(P_{\boldsymbol{X}_{G_i}} - P_{\boldsymbol{X}_{G_j}})P_{\boldsymbol{X}_A}\varepsilon\|_2/\sqrt{n} \ge (1-t_0)\lambda_1\delta/2) \\ & \le \sum_{k=1}^{K}\sum_{i<j,i,j\in\mathscr{G}_k^h} 2\exp\{-n((1-t_0)\lambda_1\delta/2 - \lambda_1\delta c_{ij})/2\} \\ & \le \epsilon. \end{aligned}$$

For event E_3, $P_{\boldsymbol{X}_{G_j}}(I - P_{\boldsymbol{X}_A})\varepsilon \sim N(\mathbf{0}, \sigma P_{\boldsymbol{X}_{G_j}}(I - P_{\boldsymbol{X}_A})P_{\boldsymbol{X}_{G_j}})$ and

$$\begin{aligned} & E(\|P_{\boldsymbol{X}_{G_j}}(I - P_{\boldsymbol{X}_A})\varepsilon\|_2) \le \sqrt{\text{trace}(P_{\boldsymbol{X}_{G_j}}(I - P_{\boldsymbol{X}_A})P_{\boldsymbol{X}_{G_j}})} \\ \le & \sqrt{P_{\boldsymbol{X}_{G_j}}(I - P_{\boldsymbol{X}_A})} \le c_j\sqrt{p_j}. \end{aligned}$$

Then

$$\begin{aligned} \mathrm{P}(E_3^c) & \le \sum_{j=J_K+1}^{J} \mathrm{P}(\|P_{\boldsymbol{X}_{G_j}}(I - P_{\boldsymbol{X}_A})\varepsilon\|_2/\sqrt{n} \ge \lambda_1(1-C_1)) \\ & \le \sum_{j=J_K+1}^{J} 2\exp\{-n[\lambda_1(1-C_1) - c_j\sqrt{p_j}]/2\} \\ & \le \epsilon. \end{aligned}$$

This complete the proof. □

Bibliography

Andrews, G. (2000). *Foundations of multithreaded, parallel, and distributed programming*, vol. 1. Addison-Wesley.

Bach, F. (2008). Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research* 9 1179–1225.

Bach, F., Jenatton, R., Mairal, J. & Obozinski, G. (2011). Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning* 19–53.

Bondell, H. & Reich, B. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* 64 115–123.

Breheny, P. & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics* 5 232–253.

Chen, B., Causton, H., Mancenido, D., Goddard, N., Perlstein, E. & Pe'er, D. (2009). Harnessing gene expression to identify the genetic basis of drug resistance. *Molecular systems biology* 5.

Chen, S., Donoho, D. & Saunders, M. (2001). Atomic decomposition by basis pursuit. *SIAM review* 129–159.

Dettling, M. & Bühlmann, P. (2004). Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis* 90 106–131.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *Annals of statistics* 32 407–451.

Fan, J., Guo, S. & Hao, N. (2010). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Arxiv preprint arXiv:1004.5178* .

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96 1348–1360.

Fan, J. & Lv, J. (2011). Non-concave penalized likelihood with np-dimensionality. *IEEE transaction on information theory* 57 5467–5484.

Fan, J. & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* 32 928–961.

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

Frank, I. & Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics* 35 109–135.

Huang, J., Breheny, P., Ma, S. & Zhang, C. (2010a). The mnet method for variable selection. *Department of Statistics and Actuarial Science, The University of Iowa* .

Huang, J., Ma, S., Li, H. & Zhang, C. (2010b). The sparse Laplacian shrinkage estimator for high-dimensional regression. *Technical Report* Department of Statistics and Actuarial Science, University of Iowa.

Huang, J., Ma, S., Li, H. & Zhang, C. (2011). The sparse laplacian shrinkage estimator for high-dimensional regression. *The Annals of Statistics* 39 2021–2046.

Huang, J., Ma, S., Xie, H. & Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika* 96 339.

Jornsten, R. & Yu, B. (2003). Simultaneous gene clustering and subset selection for sample classification via MDL. *Bioinformatics* 19 1100.

Lee, H., Hsu, A., Sajdak, J., Qin, J. & Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome research* 14 1085–1094.

Li, C. & Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* 24 1175–1182.

Li, C. & Li, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The annals of applied statistics* 4 1498.

Li, L., Li, B. & Zhu, L. (2010). Groupwise dimension reduction. *Journal of the American Statistical Association* In press.

Liu, D. (2012). Combination of confidence distributions and an efficient approach for meta-analysis of heterogeneous studies. *Ph.D thesis* Department of Statistics and Biostatistics, Rutgers University.

Mackey, L., Talwalkar, A. & Jordan, M. (2011). Divide-and-conquer matrix factorization. *arXiv preprint arXiv:1107.0789* .

Meinshausen, N. & Buhlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 417–473.

Shah, R. & Samworth, R. (2012). Variable selection with error control: Another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* to appear.

Singh, K., Xie, M. & Strawderman, W. (2005). Combining information from independent sources through confidence distributions. *Annals of statistics* 159–183.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58 267–288.

Wang, S., Nan, B., Zhu, N. & Zhu, J. (2009). Hierarchically penalized Cox regression with grouped variables. *Biometrika* 96 307–322.

Xie, M., Singh, K. & Strawderman, W. (2011). Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association* 106 320–333.

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

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38 894–942.

Zhao, P., Rocha, G. & Yu, B. (2008). Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics* 3468–3497.

Zhao, P. & Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research* 7 2541–2563.

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

Zou, H., Hastie, T. & Tibshirani, R. (2007). On the degrees of freedom of the lasso. *Annals of Statistics* 35 2173–2192.

Zou, H. & Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics* 37 1733.

Vita

Xueying Chen

2008-2013 Ph.D. in Statistics, Rutgers University, the State University of New Jersey, Piscataway, NJ

2004-2008 B.Sc. in Probability and Statistics, Peking University, Beijing, China

2009-2012 Teaching Assistant, Department of Statistics and Biostatistics, Rutgers University

2008-2009 Graduate Fellow, Department of Statistics and Biostatistics, Rutgers University