2013

Theories on Group Variable Selection in Multivariate Regression Models

Seung-Yeon Ha

THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

THEORIES ON GROUP VARIABLE SELECTION IN MULTIVARIATE REGRESSION MODELS

By

SEUNG-YEON HA

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Summer Semester, 2013

Seung-Yeon Ha defended this dissertation on July 1, 2013.

The members of the supervisory committee were:

Yiyuan She
Professor Directing Thesis

Giray Ökten
University Representative

Fred Huffer
Committee Member

Debajyoti Sinha
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with the university requirements.

I dedicate this dissertation to God.
Also, my mom, husband, and lovely daughter, Haryn, without your support I could not finish this work. Thank you for your love.

ACKNOWLEDGMENTS

Many thanks are due to many people. My major professor have introduced the interesting topic to me and always gives me the insight into what we are studying, which motivates me for the further research. I will always be grateful for his support and guidance. The other members of my committee advised me to study more deeply on the topic and this paper would not be the same without their helps. Many thanks.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

We study group variable selection on multivariate regression model. Group variable selection is selecting the non-zero rows of coefficient matrix, since there are multiple response variables and thus if one predictor is irrelevant to estimation then the corresponding row must be zero. In a high dimensional setup, shrinkage estimation methods are applicable and guarantee smaller MSE than OLS according to James-Stein phenomenon (1961). As one of shrinkage methods, we study penalized least square estimation for a group variable selection.

Among them, we study L_0 regularization and $L_0 + L_2$ regularization with the purpose of obtaining accurate prediction and consistent feature selection, and use the corresponding computational procedure Hard TISP and Hard-Ridge TISP (She, 2009) to solve the numerical difficulties. These regularization methods show better performance both on prediction and selection than Lasso (L_1 regularization), which is one of popular penalized least square method. L_0 acheives the same optimal rate of prediction loss and estimation loss as Lasso, but it requires no restriction on design matrix or sparsity for controlling the prediction error and a relaxed condition than Lasso for controlling the estimation error. Also, for selection consistency, it requires much relaxed incoherence condition, which is correlation between the relevant subset and irrelevant subset of predictors. Therefore L_0 can work better than Lasso both on prediction and sparsity recovery, in practical cases such that collinearity is high or sparsity is not low.

We study another method, $L_0 + L_2$ regularization which uses the combined penalty of L_0 and L_2. For the corresponding procedure Hard-Ridge TISP, two parameter work independently for selection and shrinkage (to enhance prediction) respectively, and therefore it gives better performance on some cases (such as low signal strength) than L_0 regularization. For L_0 regularization, λ works for selection but it is tuned in terms of prediction accuracy. $L_0 + L_2$ regularization gives the optimal rate of prediction and estimation errors without any restriction, when the coefficient of l_2 penalty is appropriately assigned. Furthermore, it can achieve a better rate of estimation error with an ideal choice of block-wise weight to l_2 penalty.

CHAPTER 1

INTRODUCTION

1.1 Motivation of Study

We are interested in the inference in multivariate regression models with the design matrix $X \in \mathbb{R}^{n\times p}$ and the corresponding response matrix $Y \in \mathbb{R}^{n\times m}$ as follows.

$$Y = XA^* + E \tag{1.1}$$

where E is a independent Gaussian noise with zero mean and variance σ^2, and $A^* \in \mathbb{R}^{p\times m}$ is the target regression matrix. Since there are multiple response variables and the corresponding multiple coefficients, selecting the relevant predictors to the estimation is equivalent to select the non-zero rows of A^*. The multivariate model is applied to wide range of studies such as multi-task learning, longitudinal data etc [24]. Specifically, we are interested in the inference on a high dimensional setup such that p is greater or even greater order than the sample size n, with the purpose of consistent selection and accurate prediction. Since X^TX is singular when $p > n$, the ordinary least square estimate is not well defined and variance is large. To reduce the variance, we can apply shrinkage methods which give more stable solutions by trading off the variance and bias. According to James-Stein phenomenon (1961), when $p \geq 3$, OLS is not admissible any more and the James-Stein estimator, which is one of shringkage estimators, gives smaller MSE. As one of shrinkage methods, we focus on the penalized least square estimation method of which solutions are defined by minimizing the prediction loss term plus penalty term as follows.

$$\hat{A} = \underset{A\in\mathbb{R}^{p\times m}}{\arg\min} \left\{ \frac{1}{2}\|Y - XA\|_F^2 + p_g(A;\tau) \right\} \tag{1.2}$$

where $p_g(\cdot;\tau)$ is group penalty function for τ is regularization parameter. Group Lasso [39, 24] which uses group L_1 norm ($p_g(A;\tau) = \tau\sum_{j\in[p]}\|A^j\|_2$, where A^j is jth row of A) reduces

dimensionality by selecting a sparse solution. However, it has some drawbacks such that the correlation between predictors should be low enough to predict accurately. For an example of the restriction, a condition called Restricted Eigenvalue (RE) condition [3], is required for controlling prediction error and estimation error at the optimal rate, which is of course much less restricted condition than the positive definiteness of X^TX, but still empirically stringent. Under the condition, the correlation of covariates should be controlled small enough, so the method may perform poorly in most cases. In this dissertation, we propose Group l_0 regularization, which uses $p_g(A;\tau) = \frac{\tau}{2}\sum_{j\in[p]} \mathbf{1}_{\{\|A\|_2\neq 0\}}$ as its penalty function instead of l_1 penalty. It does not only enforce more sparsity so to allow more parsimonious model by using nonconvex penalty, but also involves no bias on the selected subset. Our non-asymptotic studies will show that accurate prediction or consistent selection could be achieved under much relaxed condition via Group L_0 regularization than via Group L_1 (group Lasso).

Now we show literature review related to our problem on a simpler model. To avoid ambiguity, we use the following univariate response linear model and then extend its techniques to the original model (1.1). Let $Z \in \mathbb{R}^{N\times K}$ be a design matrix and $y \in \mathbb{R}^N$, the corresponding response vector satisfying

$$y = Z\gamma^* + \epsilon \tag{1.3}$$

where $\gamma^* \in \mathbb{R}^K$ be a sparse regression coefficient vector, $\varepsilon \sim \mathbf{N}(0, \sigma^2\mathbb{I}_N)$, and $N < K$ or possibly $N \ll K$. On this problem, various methods have been studied, some based on l_1-regularization, including Lasso [33] and Dantzig seletor [8], and some based on l_2-regularization refered as Ridge regression [19]. Their error metrics including estimation and prediction loss have been obtained, especially for l_1 regularization there have been several nonasymptotic studies [33, 8, 25, 42, 5, 46, 36]. General definition of the penalized least square estimator is given by

$$\hat{\gamma} = \arg\min_{\gamma\in\mathbb{R}^K} \left\{ \frac{1}{2}\|y - Z\gamma\|_2^2 + p(\gamma;\tau) \right\}. \tag{1.4}$$

The l_1 penalty, $p(\gamma;\tau) = \tau\|\gamma\|_1$, leads to L_1 regularization (Lasso), l_2 penalty, $p(\gamma;\tau) = \tau^2/2\|\gamma\|_2$, leads to Ridge estimate, and l_q leads to Bridge regression [16]. All techniques can be extended to group variable selections, for example Group Ridge regression can be defined with the penalty $p_g(A;\tau) = \frac{\tau}{2}\|A\|_F^2 = \sum_{j\in[p]}\|A^j\|_2^2$ in (1.2). Note that for this univariate regression model, the loss is defined in l_2 norm instead of Frobenius norm.

From Figure 1.1 we can see the difference between some popular penalties by naive comparison. On both graphs, we assume the design matrix is orthogonal or identity, and

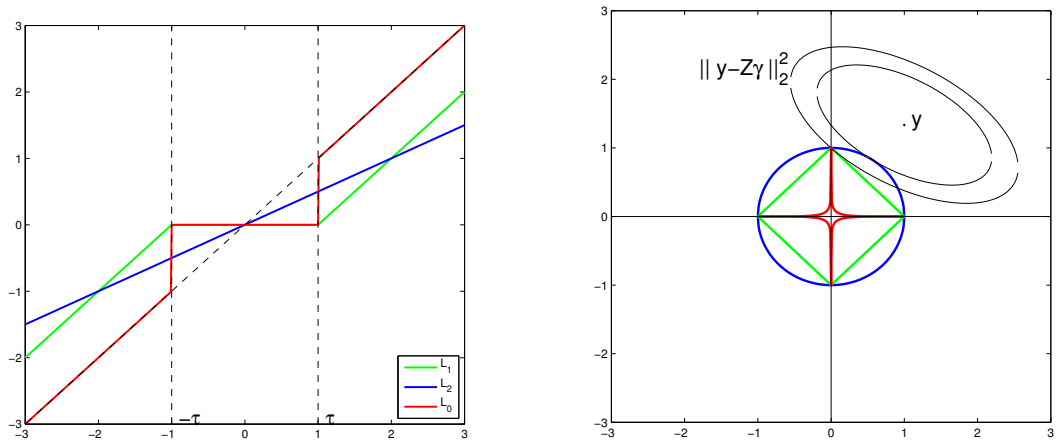

Figure 1.1: (a) Solution via L_1, L_2, and :$_0$ penalties, (b) For $K = 2$, comparison among L_1, L_2, and L_q for $q = 1/5$

show the solution paths via L_1, L_2 and L_q for $q < 1$ (or L_0) penalties. When $K = 1$(the plot on the left side), outside of threshold τ, both l_1 and l_2 regularization shrink large values so cause bias, but l_0 regularization does not cause any bias (l_0 gives the same solution path as OLS (dotted line) on the selected subset so it seems like restricted OLS). On the other hand, inside of the threshold both L_1 and L_0 set the values to zero so they give sparse solution, but :$_2$ does not. The similar features are shown when $K = 2$ (the plot on the right hand side). The ellipsoid denotes RSS=$\|y - Z\gamma\|_2^2$ and the estimates are defined as the minimizers of the error on each restricted region. Both L_1 and L_q regularization take solutions on the y-axis so to give sparse solutions, and L_q has even more chance to have a sparse solution than L_1 because it is nonconvex function. On the other hand, L_2 does not select sparse solution because of its convexity, thus in order to obtain more sparsity, one must come to nonconvex penalty. On the other hand, Ridge estimate (L_2 regularization) can perform better than the others especially when the predictors are highly correlated because it shrinks values proportionally so smaller variance of solution is obtained when correlation is high so it would give better prediction accuracy when the true model is not sparse (therefore the contribution of unselected subset via the other penalties must be small). For Lasso, despite of the more possibility of sparsity recovery, very restricted condition should be satisfied to ensure the errors are controlled and usually accurate prediction cannot be obtained. Alternative methods have been studied as the remedies to drawbacks of Lasso. One of them

is Adaptive Lasso [47] which gives weight of l_1 type penalty in data dependent way such that $w_i = 1/\hat{\gamma}_{OLS,i}$, but it is hard to be defined on high dimensional setup. Another method called elastic net [48], which uses the combined penalty of l_1 and l_2 norm, performs well on highly correlated covariates as l_2, and achieves sparsity like l_1 regularization. However, it has a drawback, that both of l_1 and l_2 shrink the selected subset so causes larger bias, which is refered as double-shrinkage (the problem is solved via corrected version of elastic net in that paper). All of these method are based on l_1 penalty for sparsity, so they do not better job for reducing demensionality. Accordingly, we consider nonconvex penalties which enforces more sparsity to select more spare subset. One concern is how to achieve our final goal (both prediction accuracy and selection consistency) in addition to the sparsity.

We first study two aspects of inference, prediction accuracy and consistent selection, via L_1 regularization, as a limit standard to achieve. On the prediction accuracy, the oracle inequalities [3] have been studied and the optimal error rates are achieved in the minimax sense [11, 38]. Under some restrictions on design matrix, its l_2 prediction loss and l_q estimation loss have the optimal rate as follow [27, 38].

$$\begin{aligned} &\|Z(\hat{\gamma} - \gamma^*)\|_2^2 = O_p(J^* \sigma^2 \log K) \\ &\|\hat{\gamma} - \gamma^*\|_q^q = O_p(J^* (\sigma^2 \log K/N)^{q/2}), \ \forall q \in [1, 2] \\ &\|\hat{\gamma}\|_0 = O_p(J^*) \end{aligned} \tag{1.5}$$

where J^* is the cardinality of γ^*. In comparison to OLS, there are inflation factor $\log K$ which is the cost for not knowing the true support set. Since each row of $\hat{\gamma} - \gamma^*$ is contaminated by Gaussian noise $(Z^T Z)^- Z\varepsilon$ and the true support set is unknown, in order to upper bound each row of absolute value of $\hat{\gamma} - \gamma^*$, the maximum of absolute value of the Gaussian noise of K dimension should be upper bounded and $\lim_{K\to\infty} \max_{i\in[K]} |z_i| / \sqrt{2 \log K/N} = 1$ [37] (since Z is normalized such that each column's l_2 norm is N). Therefore, for J^* of true cardinality the estimation error in l_q norm is upper bounded by the rate (1.5). The factor is not removable, and thus to achieve estimation consistency $\log K \lesssim o(N)$. For Lasso, this condition is not consistent with the other condition required for the oracle property : $\hat{\gamma}_{\mathcal{J}^{*c}} = 0$ and $(\hat{\gamma} - \hat{\gamma}^0)_{\mathcal{J}^*}$ is approximately normal with covariance $(1/n) I^{-1}_{\mathcal{J}^*}(\hat{\gamma}^0_{\mathcal{J}^*}, 0)$ where $\hat{\gamma}^0$ is restricted OLS, $\mathcal{J}^*$ is the true support set and $I_{\mathcal{J}^*}$ is Fisher information [14]. Therefore asyptotically speaking, Lasso does not work as well as OLS even for a large sample which is referred as Lasso bias [43]. On the other hand, to obtain the optimal rates of errors via Lasso, the restrictions on the design matrix and sparsity condition are required. For instance, restricted eigen value condition (RE) controls a kind of singularity of design matrix restricted to a cone defined based on the true sparsity, therefore the collinearity of design

matrix and sparsity should affect on how well the errors are able to be controlled. The correlation among covariates should be low enough and the sparsity of true model must be restricted as well. There are simialr conditions required on the design matrix such as RIP, Coherence condition among others, and reserchers have compared the relationships [35, 38]. In contrast, nonconvex regularization requires less restricted condition such as SCIF and RIF [38, 43] as well as achieves the oracle property [20].

For an accurate feature selection, first, correlation between relevant predictors and irrelevant ones should be relatively small so to readily classify the relevant predictors, and secondly, signal strength compared to noise level (or SNR) which can be defined as $\min|\gamma^*|/\sigma^2$ should be greater than a certain threshold. There have been great deal of study focusing on the selection consistency and the required conditions via Lasso [46, 36, 40, 12, 23, 4]. For instance, under the irrepresentable condition [46, 36], solutions have the same sign as the true model with high probability if in addition $\min|\gamma^*_{\mathcal{J}^*}| \gtrsim O(\sqrt{|\mathcal{J}^*|\log K/N})$ is satisfied. Since two conditions are related to each other, so more strict incoherence condition is required to achieve weaker condition of SNR. Only under the restricted condition called mutual incoherence condition, the minimum signal strength is $O(\sqrt{\log K/N})$ which does not include unknown parameter $|\mathcal{J}^*|$ [40]. In contrast, nonconvex penalty obtains selection consistency under much relaxed conditions. Ye and Zhang [38] showed selection consistency under sign restircted cone invertibility factor(SCIF) condition which is relaxed than the mutual incoherence condition up to constants and SNR is $O(\sqrt{\log K/N})$. MCP proposed by Zhang [41] which is concavity-restricted penalized estimator, achieves selection consistency under much relaxed one than irrepresentable condition and the optimal rate of SNR. Zhang [45] suggested another nonconvex regularization through multi-stage convex relaxation procedure, which reduce Lasso bias in every iterations by updating threshold based on solutions from the previous step. Selection consistency was shown under RIP condition-slightly restricted than SRC but only up to constant-and under SNR greater than $O(\sqrt{\log K/N})$.

In conclusion, Lasso requires strong incoherence condition for selection consistency and SNR usually inflated by $|\mathcal{J}^*|$. It also requires restricted condition for controlling errors at a certain level. One remedy is to relax convexity, or to penalize via nonconvex function, such as l_q-norm with $q < 1$. It was shown that even under mild l_2 regulaity conditions such as SRC, RIP, SCIF, or RIF selection consistency is achieved via nonconvex penalties, provided the lower rate of SNR [43]. L_0-regularization as one of the nonconvex penalty solves the issues mentioned above as well as allows more parsimonious model. Since the optimization problem involves computational difficulties from its discontinuity and non-differentiability,

we apply TISP [29] as a computational approach. we can obtain local solutions to the optimization problem via Hard-TISP under a condition. We show some non-asymptotic studies on L_0 regularization.

Although l_0-regularization works better than Lasso in most cases, there is a issue caused by the regularization parameter which is usually tuned by prediction accuracy. For example, it may perform poorly when signals is too low, and even worse than Lasso for selection aspect. Since there is only one regularization parameter, it cannot be chosen optimally for selection accuracy. We study $L_0 + L_2$ to solve the problem since two regularization parameters work independently for selection and prediction respectively. Also, as mentioned before, OLS is not admissible when the dimension is high. Since $L_0 + L_2$ shrinks the values after selection, it may give smaller MSE than the restricted OLS. These facts can be seen in some simulations [33], the best subset selection which is somewhat similar to l_0 regularization gives more sparse solution than Lasso and most of cases the subset selection and Lasso do work better than others, but l_2 regularization outperforms in some cases such as large number of small effects or highly correlated covariates, even though it does not select sparse solution. Therefore, we study $l_0 + l_2$ regularization, which can perform better than l_1 or l_0 in some challenging situations. Since l_0 selects sparse subset at the same time l_2 shrinks large values, and thereby $l_0 + l_2$ gives the sparse solution and achieves prediction accuracy as well. These facts are shown in our theories on the method later. Note that unlike elastic net which has double-shrinkage by both penalties, it does not involve any large bias since l_0 does not shrink. The difference between the elastic net and $l_0 + l_2$ regularization is the former reduce variance by deducting the same amount from the large values while the latter shrinks proportionally. Therefore, the elastic net preserves the pairwise correlation among covariates while $l_0 + l_2$ de-correlate at each iteration. It affects to the selection pattern, one selects the whole pairwise highly correlated group but the other selects only one covariate among the group. We will study when $l_0 + l_2$ regularization can do better job than the others, through theoritical research and simulation. We will show how to choose the optimal coefficient associated with l_2 penalty, and from the empirical Bayesian point of view we use the James-Stein estimator as the approximate coefficient. These ideal choice of weight show the potential of $L_0 + L_2$ which improves the estimation accuracy. For example, by adopting block thresholding [6, 7] we can assign different weight to each block to lower the estimation error.

1.2 Summary of Chapters

This dissertation is organized as follow. In Chapter 2 we define the L_0 regularization on general univariate models and introduce Hard-TISP the corresponding numerical algorithm. In Chapter 3 we study L_0 regularization in non-asymptotic way. We will compare the required conditions for controlling the errors with those of Lasso. Later we study the probability of sign agreed selection and discuss the restrictions required for Lasso to achieve sign consistency in comparison with that of l_0-regularization. $L_0 + L_2$-regularization is studied on the Chapter 4. Similarly to the previous chapter, the oracle inequalities and selection consistency will be discussed. Also, we study the optimal choice of l_2 penalty by adopting the empirical Bayesian idea. In Chapter 5, we introduce multivariate model and extend our results from the previous chapters into the multivariate model. We also show the minimax rate of errors. In Chapter 6.1, we show a comparison among L_0 and L_1, $L_0 + L_2$ on some specific cases, and in Chapter 7, we discuss about our results in comparison to Lasso. Our future works will be listed in Chapter 8.

CHAPTER 2

BASICS

In this section, we introduce notations and study the optimization problem and the numeric algorithm as an efficient approach to the original problem.

Again, note that we first consider univariate response model: $y = Z\gamma^* + \varepsilon$, where $Z \in \mathbb{R}^{N\times K}$ is design matrix, $y \in \mathbb{R}^N$ is response vector, $\gamma^* \in \mathbb{R}^K$ is target vector, and $\varepsilon \sim \mathbf{N}(0, \sigma^2\mathbb{I}_N)$. We assume that Z is column normalized design matrix to be $\|Z_j\|_2 = \sqrt{N}$ for all Z_j, which is jth column of Z. We let $\Sigma = Z^TZ$ as gram matrix and $\Sigma_{\mathcal{I}} = Z_{\mathcal{I}}^T Z_{\mathcal{I}}$, $\Sigma_{\mathcal{I},\mathcal{I}'} = Z_{\mathcal{I}}^T Z_{\mathcal{I}'}$ for $\mathcal{I} \cup \mathcal{I}' = \emptyset$. Also, we define projection matrix of Z, $P := Z(Z^TZ)^-Z^T$, where $(Z^TZ)^-$ is Moore-Penrose inverse. Throughout the paper, we assume Z and γ^* as deterministic if there is no further assumption.

We first introduce some notations for seeking convenience. From now on we define $[l] := \{1, 2, \cdots, l\}$ for every positive integer l, and Z_j is jth column vector of Z, $\forall j \in [K]$. We define the support set of $\gamma \in \mathbb{R}^K$ as $\mathcal{J}(\gamma)$, such that

$$\mathcal{J}(\gamma) = \{j : \gamma_j \neq 0,\ \forall j \in [K]\}$$

For notation simplicity, we let $\mathcal{J} := \mathcal{J}(\gamma)$, $\hat{\mathcal{J}} := \mathcal{J}(\hat{\gamma})$, and $\mathcal{J}^* := \mathcal{J}(\gamma^*)$, where $\hat{\gamma}$ is an estimator of the γ^*. We define $\tilde{\mathcal{J}} := \hat{\mathcal{J}} \cup \mathcal{J}$, which denotes index set of all non-zero elements included in either of $\hat{\gamma}$ or γ^*. For all index set $\mathcal{I} \subset [K]$, the corresponding cardinality is denoted by $|\mathcal{I}|$. The cardinalities, $|\mathcal{J}|$, $|\hat{\mathcal{J}}|$, $|\mathcal{J}^*|$, and $|\tilde{\mathcal{J}}|$ can be briefly written as J, $\hat{J}$, J^*, and $\tilde{J}$ respectively. The $\gamma_{\mathcal{I}}$ is subset of γ containing all corresponding elements to a index set $\mathcal{I} \subset [K]$. Similarly, sub-matrix of Z, formed by the columns corresponding to the index set $\mathcal{I}$, is defined as $Z_{\mathcal{I}} \subset \mathbb{R}^{N\times|\mathcal{I}|}$. For instance, $Z_{\mathcal{J}}$ is sub-matrix of Z containing columns which match to the support set of γ. Throughout this paper, for all $v \in \mathbb{R}^K$, $\|v\|_q = \{\sum_{j\in[K]}|v|^q\}^{1/q}$ and for $A \in \mathbb{R}^{m\times n}$, $\|A\|_F$, $\|A\|_2$, $d_i(A)$ denote Frobenius norm, the

largest singular value, and the ith singular value of A, respectively. We write for brevity $\|v\|_2 = \|v\|$ for vector l_2-norm where it causes no ambiguity.

In section 5, we will consider grouping on the univariate model and extend our result from univariate model to multi-response model. Let $Y \in \mathbb{R}^{n\times m}$ be a multivariate response matrix and $X \in \mathbb{R}^{n\times p}$ a design matrix.

$$Y = XA^* + E$$

where $A^* \in \mathbb{R}^{p\times m}$ is the target matrix, and $E \in \mathbb{R}^{n\times m}$ is zero mean Gaussian noise. By vectorizing, the model is transformed to univariate response one as follow:

$$y = \mathrm{vec}(Y) = \begin{bmatrix} Y_1 \\ \vdots \\ Y_m \end{bmatrix}, \quad \gamma^* = \mathrm{vec}(A^*), \quad \varepsilon = \mathrm{vec}(E)$$

$$x = I \otimes X = \begin{bmatrix} X & & 0 \\ & \ddots & \\ 0 & & X \end{bmatrix}$$

where m is the number of groups, so $K = pm$ and $N = nm$.

Now we go back to the simple linear model and study l_0-norm regularization. The optima for the following optimiztion problem are our solutions.

$$F(\gamma) = \frac{1}{2}\|Z\gamma - y\|^2 + p\left(\gamma;\lambda\right) \ , \quad \forall\gamma \in \mathbb{R}^K \tag{2.1}$$

$$\hat{\gamma} = \underset{\gamma\in\mathbb{R}^K}{\arg\min}\, F(\gamma)$$

where $p(\gamma;\lambda) = \lambda^2/2 \cdot \|\gamma\|_0$ for l_0 regularization. Instead of optimizing the original problem, which involves numerical difficulties because the penalty function is discontinuous and non-differentiable at zero, we turn to a efficient computational approach called Hard Thresholding-based Iterative Selection Procedure(hard-TISP) [29]:

$$\hat{\gamma}^{(j+1)} = \Theta_H((I - \Sigma)\hat{\gamma}^{(j)} + Z^T y \ ; \ \lambda) \tag{2.2}$$

where $\Theta_H(\alpha;\lambda)$ is hard thresholding equation, with $\hat{\gamma}^{(j)}$ as updated vector at jth iteration step. The TISP can be viewed as e-m algorithm by the following two steps. First for $\gamma, \gamma\prime \in \mathbb{R}^K$ let's define,

$$g(\gamma', \gamma) = \frac{1}{2}\|y - Z\gamma\prime\|^2 + \frac{\lambda^2}{2}\|\gamma\prime\|_0 + \frac{1}{2} < (I - \Sigma)(\gamma\prime - \gamma), \gamma\prime - \gamma > .$$

For fixed γ, the first step is to minimize g over $\gamma\prime$ which is equivalent to

$$\arg\min_{\gamma\prime}[\frac{1}{2}\|\gamma\prime - \{(I-\Sigma)\gamma + Z^T y\}\|^2 + \frac{\lambda^2}{2}\|\gamma\prime\|_0,],$$

and the second step is minimizing g over γ for given $\gamma\prime$ such that

$$\arg\min_{\gamma} \frac{1}{2} < (I-\Sigma)\gamma, \gamma - 2\gamma\prime > .$$

Note that the first step is optimizing orthogonal design which is computationally easier to handle, and the second step is a convex optimizaiton and its optima is $\gamma = \gamma\prime$ if $\|\Sigma\|_2 < 1$. Therefore, minimizing g is equivalent to minimizing F and it is more easier to conduct through hard-TISP. That is, a fixed points of (2.2) is a solution to (2.1). Theorem 2.1 in [30] guarantees that if $\rho = \|Z^T Z\|_2 \leq 1$ and for any penalty function p satisfying

$$P(t;\lambda) - P(0;\lambda) = \int_0^{|t|} (\sup\{s : \Theta_H(s,\lambda) \leq u\} - u)du + q(t;\lambda) \tag{2.3}$$

with $q(t;\lambda)$ nonnegative and $q(\Theta_H(t;\lambda);\lambda) = 0, \forall t \in RR$, the value of the function F in (2.1) decreases at each iteration such that

$$F(\gamma^{(j)}) - F(\gamma^{(j+1)}) \geq C \cdot \|\gamma^{(j)} - \gamma^{(j+1)}\|_2^2$$

where $C = 1 - \rho$. If further $\rho < 1$ then any limit poit of $\gamma^{(j)}$ should be a fixed point of the hard-TISP or a stationary point of (2.1). If q-function is equal to zero, then we obtain the hard penalty function [30] given by

$$p(t;\lambda) = \begin{cases} -t^2/2 + \lambda|t|, & \text{if} |t| < \lambda \\ \lambda^2/2, & \text{otherwise} \end{cases} \tag{2.4}$$

The entropy penalty $p(t;\lambda) = \lambda^2/2 I(|t| \neq 0)$ is given if q-function is

$$q(t;\lambda) = \begin{cases} (\lambda - |t|)^2/2 & \text{if } 0 < |t| < \lambda \\ 0, & t = 0 \text{ or } |t| \geq \lambda \end{cases}$$

which is more strict than (2.4). Both penalties result in the same thresholding rule, but we cannot guarantee the hard thresholding estimates correspond to the global solution to those penalty functions. Note that the decreasing property is guaranteed only when the largest singular value of $Z^T Z$ is smaller than 1. Therefore first we need a proper preliminary scaling by a constant k_0 greater than $\|Z^T Z\|_2$ keeping the same optimization problem such that

$$F(\gamma) = \frac{1}{2k_0^2}\|y - Z\gamma\|^2 + p(\gamma; \frac{\lambda}{k_0}) , \quad \forall \gamma \in \mathbb{R}^K, \tag{2.5}$$

where $p(\gamma; \frac{\lambda}{k_0}) = \frac{(\lambda/k_0)^2}{2}\|\gamma\|_0$ for l_0 regularization. Therefore, the previous threshold λ in (2.2) is changed to λ/k_0, then we have a new Θ_H-estimator given by

$$\hat{\gamma} = \vec{\Theta}_H\left(\left(I - \frac{\Sigma}{k_0^2}\right)\hat{\gamma} + \frac{Z^T y}{k_0^2}\ ;\ \frac{\lambda}{k_0}\right). \tag{2.6}$$

That is for every $j \in [K]$,

$$\hat{\gamma}_j = \begin{cases} \hat{\gamma}_j + \frac{1}{k_0^2} Z_j^T\,(y - Z\hat{\gamma})\ , & \text{if } \left|\hat{\gamma}_j + \frac{1}{k_0^2} Z_j^T\,(y - Z\hat{\gamma})\right| > \frac{\lambda}{k_0} \\ 0\ , & \text{otherwise} \end{cases} \tag{2.7}$$

thus, after sufficiently large iterations we have $|\hat{\gamma}_j| \geq \frac{\lambda}{k_0}$, for all $\hat{\gamma}_j \neq 0$.

Remark 2.1 *Generalized sign $\hat{S}$ is defined in [29] as*

$$\hat{S} = \hat{S}(\hat{\gamma}; \lambda) = \{\hat{s} \in \mathbb{R}^K : \Theta(\hat{\gamma} + \lambda\hat{s}; \lambda) = \hat{\gamma}\} \quad \textit{if } \hat{\gamma} \in \mathbb{R}^K \tag{2.8}$$

and $\hat{S}_j = 0$ otherwise. Thus, $\hat{s} \in \hat{S}(\hat{\gamma}; \lambda/k_0)$ satisfies $\hat{\gamma}_j + \frac{\lambda}{k_0}\hat{s}_j = \hat{\gamma}_j + \frac{1}{k_0^2} Z_j^T\,(y - Z\hat{\gamma})$, and therefore from (2.7), $\hat{s}_j = 0$ if $j \in \hat{\mathcal{J}}$, or $\hat{s}_j \in [0, 1]$ otherwise. Later, we will use the following extended version of Karush-Kuhn-Tucker (KKT) conditions to prove some theorems.

$$Z_j^T Z(\hat{\gamma} - \gamma^*)/k_0^2 - Z_j^T \varepsilon/k_0^2 + \lambda/k_0 \hat{s}_j = 0, \forall j \in [K], \tag{2.9}$$

where $\hat{s}_j \in [0, 1]$, $\forall j \notin \hat{\mathcal{J}}$, or $\hat{s}_j = 0$, otherwise.

Since all of the fixed points of TISP algorithm are not the global minimizer of (2.5), we cannot assure to get a global solutions through the iterative algorithm. Hence, on the study of errors of global solutions, possibly not unique, $\hat{\gamma}$ we use (2.5) only. Later for the study on feature selection we show the probability of selecting correct subset via a local solution by using the above extended KKT conditions.

CHAPTER 3

THEORIES ON THE L_0 REGULARIZATION

In this chapter we study performance of l_0-norm regularization focusing on the prediction and selection. We consider a sparse coefficient vector which contains many zeros or many of relatively small values close to zero. The corresponding conditions of sparsity are hard sparsity or soft sparsity : the former assumption means that the number of exact non-zeros are constrainted and the latter means many of the covariates only make a small overall contribution to the model although the coefficients may be non-zero [27]. It is worth pointing out that the sparsity assumption is not required for l_0 regularization, in contrast that it is necessarily involved in the restrictions on design matrix such as RIP, RE conditions for Lasso.

During the study on l_0 regularization we found out the similar study had done in [43]. In his study, 'Null consistency' is used for deriving the upper bound, which is similar to our way to cover $< \varepsilon, (\hat{\gamma} - \gamma^*) >$. We will discuss about the 'Null consistency' on $l_0 + l_2$ study, but in this chapter we state our result on l_0 without the property. Also, we borrow the idea of bounding estimation error in the paper, so we show the proof on our setup independently and discuss about the meaning.

3.1 Oracle Bounds

Theorem 3.1 *Consider the model* (1.3) *and a global solution to* (2.5)*,* $\hat{\gamma}$*. Let* $K > 1$*. For any* $c > 2$ *and* $\alpha > 0$*, we choose* λ *as*

$$\lambda = \sigma\sqrt{c((\alpha+1)\log K + 2)}. \tag{3.1}$$

Then with probability greater than $1-\frac{K^{-\alpha}}{1-K^{-\alpha}}$ *for any solution to* (2.1) *and for any* γ *and its cardinarity* J *we have that*

$$\|Z(\hat{\gamma}-\gamma^*)\|^2 \leq \frac{c+2}{c-2}\|Z(\gamma-\gamma^*)\|^2+\frac{2c\lambda^2}{c-2}J. \tag{3.2}$$

In addition, if we choose $\lambda=\sigma\sqrt{2((\alpha+1)\log K+2)}$, *then for any* $J\in[K]$ *and for any* $\xi>0$, *we have the following inequality with the probability greater than* $1-K^{-\alpha}$,

$$|\hat{\mathcal{J}}\backslash\mathcal{J}^*| \leq \frac{(\xi+1)}{\lambda^2}\|Z(\hat{\gamma}-\gamma^*)\|^2+\frac{(\xi+1)}{\lambda^2}\frac{\xi+2}{\xi}\|Z(\gamma-\gamma^*)\|^2+\frac{\xi+2}{\xi}J. \tag{3.3}$$

Corollary 3.1 *Suppose that conditions in Theorem 3.1 is satisfied with the* λ *as in* (3.1) *for a given* $c>2$ *and* $\alpha>0$. *Then with the probability greater than* $1-\frac{K^{-\alpha}}{1-K^{-\alpha}}$, *we have*

$$\|Z(\hat{\gamma}-\gamma^*)\|^2 \leq \frac{2\lambda^2 c}{c-2}J^*.$$

If we take $\lambda=\sigma\sqrt{2c((\alpha+1)\log K+2)}$ *for a given* $c>1$, *then with the probability greater than* $1-e^{-\alpha}$,

$$|\hat{\mathcal{J}}\backslash\mathcal{J}^*| \leq \frac{c+1}{c-1}J^*. \tag{3.4}$$

Definition 3.1 *Zhang and Zhang [43]. Restricted invertibility factors(RIF)*
For $\delta\geq 0$ *and* $\mathcal{J}\subset[K]$,

$$RIF_q(\delta,\mathcal{J})=\inf\left\{\frac{|\mathcal{J}|^{1/q}\|\Sigma u\|_\infty}{\sqrt{N}\|u\|_q} : \|u_{\mathcal{J}^c}\|_0 < \delta\|u_{\mathcal{J}}\|_0\right\}. \tag{3.5}$$

Theorem 3.2 *Zhang and Zhang [43] Let* $\hat{\gamma}$ *is a global solution to* (2.1) *and* $K>1$, $1\leq q\leq\infty$. *The* $\lambda=2\sigma\sqrt{c((\alpha+1)\log K+2)}$ *for a given* $c>1$, $\alpha>0$ *and then we have the following inequality with the probability greater than* $1-K^{-\alpha}$,

$$\|\hat{\gamma}-\gamma^*\|_q \leq \frac{1+\sqrt{c}}{\sqrt{c}}\cdot\frac{|\mathcal{J}^*|^{1/q}\lambda/\sqrt{N}}{RIF_q(\delta,\mathcal{J}^*)}., \tag{3.6}$$

where $\delta=(c+1)/(c-1)$.

Corollary 3.2 *Suppose that all conditions in Theorem 3.2 are satisfied then* $\hat{\gamma}$ *satisfies*

$$\|\hat{\gamma}-\gamma^*\|_\infty \leq \frac{1+\sqrt{c}}{\sqrt{c}}\cdot\frac{\lambda/\sqrt{N}}{RIF_\infty(\delta,\mathcal{J}^*)}. \tag{3.7}$$

Remark 3.1

(1) The RIF condition is invented to achieve the optimal error bound for estimation. Since $\|\Sigma u\|_\infty$ *does not involve any extra* J^* *term, in contrast, using* $\|Zu\|_2$ *instead of the infinity norm gives the looser upper bound. See below. If RIF has positive value then,*

$$\frac{J^{*1/q}\|\Sigma u\|_\infty}{N\|u\|_q} \le \frac{J^{*1/q}\max_{i\in[N]}\|Z_i\|_2\|Zu\|_2}{N\|u\|_q} = \frac{J^{*1/q}\|Zu\|_2}{\sqrt{N}\|u\|_q} := C_q$$

therefore, $\|u\|_q \lesssim \frac{J^{*1/q}\|Zu\|_2}{\sqrt{N}C_q} = O(\frac{J^{*1/q+1/2}\lambda}{\sqrt{N}C_q})$*. Since infinity norm involves linear form of* u *and* l_2 *norm imposes quadratic form of* u*, bounding by using* $\|Zu\|_2$ *will give sub-optimal rate. Note that in general restrictions for Lasso, the numerator is* l_2 *norm therefore the estimation error cannot be derived for all* $q \in [1,\infty]$ *and cannot achieve the optimal rate. SCIF condition for Lasso is suggested in [38] to obtain the optimal rate.*

(2) The probabilities are close to one when the dimension K *is large, and the regularization parameter* λ *is decided only by* K *and* α*.* K *is in the logarithm form since each row of* $\hat{\gamma}-\gamma^*$ *is contaminated by* K *dimensional Gaussian noise* $(Z^TZ)^-Z^T\varepsilon$ *and the maximum of its absolute value is at the rate of* $O(\sigma\sqrt{\log K/N})$ *since* $\|Z_j\|_2 = \sqrt{N}$*.*

(3) The rate of prediction l_2 *loss, estimation* l_q *loss, and selection error can be described as*

$$\frac{\|Z(\hat{\gamma}-\gamma^*)\|_2^2}{\log K} + \frac{\|\hat{\gamma}-\gamma^*\|_q^q}{(\log K/N)^{q/2}} + |\hat{\mathcal{J}}\backslash\mathcal{J}^*| = O_p(J^*).$$

Note that the error rates via l_0 *regularization achieve the same rate via Lasso as in* (1.5)*, without any restrictions for prediction error. The RIF condition is not quiet comparable to the conditions on Lasso, since the restricted vector space defined in each conditions are not the same, but empirically RIF is more relaxed than RE condition. We will compare two conditions in the next section.*

(4) The inequalities (3.2) *and* (3.3) *are hold for any* $\gamma \in \mathbb{R}^K$ *and we did not make any assumption on* γ^**. Therefore if we choose* γ^* *is approximate sparse* γ *such that a certain component of* γ^* *is nonzero but close to zero. Then* $\|Z(\gamma-\gamma^*)\|_2^2$ *is governed by* J *which is strong sparsity.*

Proof. *Lemma 4.1*

From Lemma A.2, for given $\alpha > 0$ and $\lambda = \sigma\sqrt{\alpha+5+\log K}/\eta$, we have

$$\|\varepsilon^T P_{\mathcal{J}}\|_2^2 \le \lambda^2\eta^2 J\ ,\quad \forall \mathcal{J}\subset[K]$$

with the probability greater than $1-\frac{e^{-\alpha}}{1-e^{-\alpha}}$. Then we have,

$$\varepsilon^T Z\gamma = \varepsilon^T P_{\mathcal{J}} Z_{\mathcal{J}}\gamma_{\mathcal{J}} \le \|\varepsilon^T P_{\mathcal{J}}\|_2\|Z\gamma\|_2 \le \lambda\eta\sqrt{J}\|Z\gamma\|_2 = \lambda\eta\sqrt{\|\gamma\|_0}\|Z\gamma\|_2.$$

Therefore

$$\begin{aligned}\frac{1}{2k_0^2}\|\frac{\varepsilon}{\eta}-Z\gamma\|_2^2-\frac{1}{2k_0^2}\|\frac{\varepsilon}{\eta}\|_2^2+\frac{\lambda^2}{2k_0^2}\|\gamma\|_0 &= -\frac{1}{\eta k_0^2}\varepsilon^T Z\gamma+\frac{1}{2k_0^2}\|Z\gamma\|_2^2+\frac{\lambda^2}{2k_0^2}\|\gamma\|_0\\ &\geq \frac{1}{2k_0^2}\left(\|Z\gamma\|_2-\lambda\sqrt{\|\gamma\|_0}\right)^2\\ &\geq 0\ ,\quad \forall\gamma\in\mathbb{R}^K\end{aligned}$$

■

Proof. *Theorem 3.1*

First we prove the inequality (3.2)

From (2.5), for all $\gamma\in\mathbb{R}^K$

$$\frac{1}{2k_0^2}\|Z\hat{\gamma}-y\|^2+\frac{\lambda^2}{2k_0^2}\hat{J}\ \leq\ \frac{1}{2k_0^2}\|Z\gamma-y\|^2+\frac{\lambda^2}{2k_0^2}J$$

then

$$\frac{1}{2}\|Z(\hat{\gamma}-\gamma^*)\|^2\leq\frac{1}{2}\|Z(\gamma-\gamma^*)\|^2+\varepsilon^T Z(\hat{\gamma}-\gamma)+\frac{\lambda^2}{2}(J-\hat{J}) \tag{3.8}$$

Then we have,

$$\begin{aligned}\varepsilon^T Z(\hat{\gamma}-\gamma)&=\varepsilon^T Z_{\tilde{\mathcal{J}}}(\hat{\gamma}-\gamma)_{\tilde{\mathcal{J}}}=\varepsilon^T P_{\tilde{\mathcal{J}}}Z_{\tilde{\mathcal{J}}}(\hat{\gamma}-\gamma)_{\tilde{\mathcal{J}}}\\ &\leq\|\varepsilon^T P_{\tilde{\mathcal{J}}}\|\ \|Z_{\tilde{\mathcal{J}}}(\hat{\gamma}-\gamma)_{\tilde{\mathcal{J}}}\|\\ &=\|\varepsilon^T P_{\tilde{\mathcal{J}}}\|\ \|Z(\hat{\gamma}-\gamma)\|\\ &\leq\|\varepsilon^T P_{\tilde{\mathcal{J}}}\|\ \ (\ 2\|Z(\hat{\gamma}-\gamma^*)\|^2+\ 2\|Z(\gamma-\gamma^*)\|^2)^{1/2}\\ &\leq\frac{c}{2}\|\varepsilon^T P_{\tilde{\mathcal{J}}}\|^2\ +\frac{1}{c}\|Z(\hat{\gamma}-\gamma^*)\|^2+\frac{1}{c}\|Z(\gamma-\gamma^*)\|^2\end{aligned} \tag{3.9}$$

for any $c>0$. The last inequality is derived by using $2xy\leq 1/cx^2+cy^2$.

Then from (3.8), on the event $\mathcal{A}_{\lambda/\sqrt{c}}$ as defined in Lemma A.2 we have

$$\begin{aligned}&\|Z(\hat{\gamma}-\gamma^*)\|^2\\ &\leq\frac{c+2}{c-2}\|Z(\gamma-\gamma^*)\|^2+\frac{c^2}{c-2}\|\varepsilon^T P_{\tilde{\mathcal{J}}}\|_2^2+\frac{\lambda^2 c}{c-2}(J-\hat{J})\\ &\leq\frac{c+2}{c-2}\|Z(\gamma-\gamma^*)\|^2+\frac{c}{c-2}\left(c\|\varepsilon^T P_{\tilde{\mathcal{J}}}\|_2^2-\lambda^2(J+\hat{J})+2\lambda^2 J\right)\\ &\leq\frac{c+2}{c-2}\|Z(\gamma-\gamma^*)\|^2+\frac{c}{c-2}\max_{\mathcal{I}\subset[K]}\left\{c\|\varepsilon^T P_{\mathcal{I}}\|_2^2-\lambda^2|\mathcal{I}|\right\}\\ &+\frac{2c\lambda^2}{c-2}J\end{aligned} \tag{3.10}$$

for any $c > 2$. From (3.10), Lemma A.2, and with a given $\alpha > 0$, $c > 2$, and λ as,

$$\lambda = \sigma\sqrt{c(\alpha + 2 + \log K)},$$

we have the following inequality with the probability greater than $1 - \frac{e^{-\alpha}}{1-e^{-\alpha}}$.

$$\|Z(\hat{\gamma} - \gamma^*)\|^2 \leq \frac{c+2}{c-2}\|Z(\gamma - \gamma^*)\|^2 + \frac{2c\lambda^2}{c-2}J.$$

Secondly, we will show (3.3).

From (3.8) and (3.9), by substituting $c = 2/t$,

$$\begin{aligned}
&(1-t)/2\|Z(\hat{\gamma} - \gamma^*)\|_2^2 \\
&\leq (1+t)/2\|Z(\gamma - \gamma^*)\|_2^2 + 1/t\|P_{\tilde{\mathcal{J}}}\varepsilon\|_2^2 + \lambda^2/2(J - \hat{J}) \\
&\leq (1+t)/2\|Z(\gamma - \gamma^*)\|_2^2 + 1/t \max_{\mathcal{I}\subset[K]}\{\|P_{\mathcal{I}}\varepsilon\|_2^2 - b\lambda^2/2\} + \lambda^2/2(b/t|\hat{\mathcal{J}} \cup \mathcal{J}| + J - \hat{J}) \\
&\leq (1+t)/2\|Z(\gamma - \gamma^*)\|_2^2 + \lambda^2/2\{(b/t - 1)|\hat{\mathcal{J}}\backslash\mathcal{J}| + (b/t + 1)|\mathcal{J}\backslash\hat{\mathcal{J}}| + b/t|\hat{\mathcal{J}} \cap \mathcal{J}|\}
\end{aligned}$$

where the last inequality hold on the event $\mathcal{A}_{\sqrt{b}\lambda/\sqrt{2}}$ for arbitrary $b < t$. Therefore for $\lambda = \sigma\sqrt{2(\alpha + 2 + \log K)/b}$, the following inequality holds with the probability greater than $1 - e^{-\alpha}$. Since $t > 1$ and $b < t$,

$$\begin{aligned}
&|\hat{\mathcal{J}}\backslash\mathcal{J}| \\
&\leq \left(1 - \frac{b}{t}\right)^{-1}\left\{\frac{t-1}{\lambda^2}\|Z(\hat{\gamma} - \gamma^*)\|^2 + \frac{t+1}{\lambda^2}\|Z(\gamma - \gamma^*)\|^2 + \left(\frac{b}{t} + 1\right)|\mathcal{J}\backslash\hat{\mathcal{J}}| + \frac{b}{t}|\mathcal{J} \cap \hat{\mathcal{J}}|\right\} \\
&\leq \left(1 - \frac{b}{t}\right)^{-1}\left\{\frac{t-1}{\lambda^2}\|Z(\hat{\gamma} - \gamma^*)\|^2 + \frac{t+1}{\lambda^2}\|Z(\gamma - \gamma^*)\|^2 + \left(\frac{b}{t} + 1\right)J\right\} \qquad (3.11)
\end{aligned}$$

Taking $b = 1$ and substituting $t = 1 + \xi$ for an arbitrary $\xi > 0$, then since $(1 - 1/t)^{-1} = (\xi + 1)/\xi$ and $1/t + 1 = (\xi + 2)/(\xi + 1)$, we have

$$\begin{aligned}
|\hat{\mathcal{J}}\backslash\mathcal{J}^*| &\leq \frac{\xi+1}{\xi}\left\{\frac{\xi}{\lambda^2}\|Z(\hat{\gamma} - \gamma^*)\|^2 + \frac{\xi+2}{\lambda^2}\|Z(\gamma - \gamma^*)\|^2 + \frac{\xi+2}{\xi+1}J^*\right\} \\
&\leq \frac{(\xi+1)}{\lambda^2}\|Z(\hat{\gamma} - \gamma^*)\|^2 + \frac{(\xi+1)}{\lambda^2}\frac{\xi+2}{\xi}\|Z(\gamma - \gamma^*)\|^2 + \frac{\xi+2}{\xi}J^*.
\end{aligned}$$

Substituting $\gamma = \hat{\gamma}$ gives (3.3).

For the proof of Corollary 3.1, the first inequality is simply given by taking $\gamma = \gamma^*$ in (3.2). The second inequality is derived if we take $t = 1$ in (3.11). ■

Proof. *Theorem 3.2* For any $t > 0$,

$$\begin{aligned}
&\|y - Z\hat{\gamma}\|^2/(2k_0^2) + \lambda^2/(2k_0^2)\|\hat{\gamma}\|_0 \\
&\leq \|y - Z\hat{\gamma} - tZ_j\|^2/(2k_0^2) + \lambda^2/(2k_0^2)(\|\hat{\gamma}_{-j}\|_0 + \mathbf{I}_{\{\hat{\gamma}_j + t \neq 0\}}).
\end{aligned}$$

Then

$$\begin{aligned}
Z_j^T(y - Z\hat{\gamma}) &\leq t\|Z_j\|^2/2 + \lambda^2/(2t)(\mathbf{I}_{\{\hat{\gamma}_j + t \neq 0\}} - \mathbf{I}_{\{\hat{\gamma}_j \neq 0\}}) \\
&\leq tN/2 + \lambda^2/(2t).
\end{aligned}$$

Also, on the event $A := \cap_{j \in [K]}\{\|Z_j^T\varepsilon\|_2^2/N \leq \lambda^2\eta^2\}$,

$$\begin{aligned}
Z_j^T\varepsilon/\eta = Z_j^T \cdot \frac{Z_j Z_j^T}{\|Z_j\|_2^2}\frac{\epsilon}{\eta} &\leq \|Z_j\|_2 \cdot \frac{\|Z_j Z_j^T\varepsilon\|_2}{N\eta} \\
&\leq t\|Z_j\|_2^2/2 + N\|Z_j^T\varepsilon\|_2^2/(N^2\eta^2 2t) \\
&\leq tN/2 + \lambda^2/2t.
\end{aligned}$$

Note that $\|Z_j^T\varepsilon\|_2^2/\|Z_j\|_2^2 \sim \sigma^2 \cdot \chi_1^2$, and from Lemma A.1 we have

$$\begin{aligned}
P(A^c) &\leq \sum_{\jmath \in [K]} \{\|Z_j^T\varepsilon\|_2^2/\|Z_j\|_2^2 > \lambda^2\eta^2\} \\
&\leq K \exp\left[-\frac{(\lambda^2\eta^2/\sigma^2 - 1)^2}{4\lambda^2\eta^2/\sigma^2}\right] \\
&= \exp\left[-\frac{(\lambda^2\eta^2/\sigma^2 - 1)^2}{4\lambda^2\eta^2/\sigma^2} + \log K\right] \\
&\leq e^{-\alpha},
\end{aligned}$$

for some $\alpha > 0$. If $\lambda \geq 2\sigma\sqrt{\alpha + 1 + \log K}/\eta$, then the event A holds with the probability greater than $1 - e^{-\alpha}$.

From those results

$$\begin{aligned}
\|Z^T Z(\gamma^* - \hat{\gamma})\|_\infty = \|Z^T(y - \varepsilon - Z\hat{\gamma})\|_\infty &\leq \|Z^T(y - Z\hat{\gamma})\|_\infty + \|Z^T\varepsilon\|_\infty \\
&\leq (1 + \eta)(tN/2 + \lambda^2/(2t)),
\end{aligned}$$

and since $t = \lambda/\sqrt{N}$ is the minimizer of the upper bound so we have

$$\|Z^T Z(\gamma^* - \hat{\gamma})\|_\infty \leq (1 + \eta)\lambda\sqrt{N}.$$

Then we will derive upper bound of $\|\gamma^* - \hat{\gamma}\|_q$ for any $q \geq 1$. In corollary 3.1 we already have $\|(\gamma^* - \hat{\gamma})_{\hat{\mathcal{J}} \setminus \mathcal{J}^*}\|_0 \leq \delta \|(\gamma^* - \hat{\gamma})_{\mathcal{J}^*}\|_0$ with $\delta = (1+\eta^2)/(1-\eta^2)$ for a given $\eta < 1$. Therefore,

$$\|\gamma^* - \hat{\gamma}\|_q \leq \frac{|\mathcal{J}^*|^{1/q} \|Z^T Z(\gamma^* - \hat{\gamma})\|_\infty}{RIF_q(\delta, \mathcal{J}^*)} \leq \frac{|\mathcal{J}^*|^{1/q}(1+\eta)\lambda/\sqrt{N}}{RIF_q(\delta, \mathcal{J}^*)}.$$

Substituting $c = 1/\sqrt{\eta}$ then we have the inequlity. ■

3.2 Asymptotic Studies

The optimal rate of estimation error in l_2 norm is shown as $O(J^* \log K/N)$, which is intuitively understandable since the prediction error is $O(J^* \log K)$ and $\|Z_j\|_2 = \sqrt{N}$. Our study is based on nonasymptotic analysis, but it is worth to compare with other estimator based on the key aspects of good estimator to evaluate on general criterion. For instance, the OLS is evaluated as good estimator, which is asymptotically Normal distributed and consistently estimate the true parameter. The OLS estimator, $\hat{\gamma}^0$ is approximately,

$$\sqrt{N}(\hat{\gamma}^0 - \gamma^*) \sim \mathcal{N}(0, I^{-1}),$$

therefore $\|\hat{\gamma}^0 - \gamma^*\|_2^2 \approx O(1/N)$ and $P(\|\hat{\gamma}^0 - \gamma^*\|_2^2 \geq \varepsilon) \geq 1/(N\varepsilon) \to 0$, which shows the OLS is consistent estimator. On the other hand, for l_0 regularization, if $K, N \to \infty$, J^* is fixed, and $\log K = o(N)$ that is $\exp(N)$ increases faster than K, then $\hat{\gamma}$ consistently estimates γ^*. If $K \leq N$ where K is fixed wile N increases, it works as good as OLS on consistent estimation. Furthermore, if the true cardinality increases as K increases, we need $J^* \log K = o(N)$ to achieve consistancy.

Note that we consider the specific univariate model which is the vectorized version of multivariate model as described in Chapter 2, and the design matrix Z is a block diagonal matrix which consists of $X \in \mathbb{R}^{n \times p}$. Then the design matrix should be normalized such that $\|Z_j\|_2 = \sqrt{n}$ and accordingly, the scaling numerater of RIF should be n instead of N. In this setup, the estimation loss in l_2 norm is upper bounded at the order of $O(J^* \log K/n)$ and the restricted OLS estimator should satisfy

$$\sqrt{n}(\hat{\gamma}^0 - \gamma^*) \sim \mathcal{N}(0, I^{-1}).$$

Therefore, to achieve the root-n consistency, $\log K = \log m + \log p = o(n)$. The group size and the dimensionality should increase slower than $exp(n)$.

Lasso has the same error bound so it achieves the estimation consistency with the same rate of K and N, but it does not work as good as OLS, which is explained in the following subsection.

3.2.1 Root-n Consistency and Oracle Property

The asymptotic property of estimators are studied more in detail in [22] on nonsigular design and [15] on orthogonal design setup. A penalized likelihood estimator which minimizes (2.1) converges at the rate which is

$$O(n^{-1/2} + \max\{p'(|\gamma|) : \gamma \neq 0\}/N) \tag{3.12}$$

since roughly speaking, (and under the assumption that the gram matrix is nonsingular) $\hat{\gamma} = \quad (Z^T Z)^{-1}\ (Z^T y - p'(|\gamma|))$. An estimator with the convergent rate as $1/\sqrt{N}$ (root-n consistency) achieves another desired aspect, the oracle property, which is given by

$$\hat{\gamma}_{\mathcal{J}^{*c}} = 0, \quad \sqrt{N}(\hat{\gamma}_{\mathcal{J}^*} - \hat{\gamma}^0_{\mathcal{J}^*}) \sim AN(b, I^{-1}_{\mathcal{J}^*})$$

where $\hat{\gamma}^0_{\mathcal{J}^*}$ is restricted OLS on the true support set and b is a function of γ^*, σ^2 which is small. $I_{\mathcal{J}^*}$ denotes the Fisher information of $\hat{\gamma}^0$ restricted to the support set. The property means that the estimator works as well as OLS with knowing support set. To achieve the root-n consistency condition, the second term of the above rate should be $O(n^{-1})$. For Lasso, the rate is $O(n^{-1/2} + \sqrt{\log K}/N)$ and if $\log K = O(N)$ then it achieves root-n consistency so the oracle property as well. For hard penalty $p(\gamma; \lambda) = \lambda^2 - (|\gamma| - \lambda)^2 I(|\gamma| < \lambda)$ which is another resulting function of hard thresholding, the rate is $O(n^{-1/2}(1 + \lim_{\gamma \to 0}(-2|\theta|/\sqrt{N} + \sqrt{\lambda}/\sqrt{N}\})$, therefore if $\sqrt{\log K/N} \to 0$ then it has root-n consistency. On the other hand, for the aspect of consistent estimating, the rate of $\|\hat{\gamma} - \gamma^*\|^2$ should tend to zero, which is asymptotically $\sigma^2/N - 2\varepsilon^T p'(|\gamma|)/(N\sqrt{N}) + (p'(|\gamma|)/N)^2$. Note that since $|\varepsilon_i| \approx O(\sqrt{\log N}) \lesssim O(\sqrt{N})$, if $p'(|\gamma|)_i/\sqrt{N} = o(1)$ then the estimator estimate consistently. For both of Lasso and hard penalty, $\sqrt{\log K/N} = o(1)$ is required to obtaining the condition. This condition is easily derived when considering the optimal rate of estimation error bound in l_2 norm is $O(\sqrt{\log K/N})$ for both Lasso and l_0 regularization. Note that for Lasso both conditions for root-n consistency and consistent estimation are not satisfied, in contrast that for hard penalty two conditions are the same. Therefore, $\log K \approx o(N)$ rate guarantees the two desired aspects of good estimator.

3.3 The Conditions Required for Controlling Errors

As shown in the previous section, the prediction error bound is upper bounded at the optimal rate under no constraint. For l_1-regularization, to control the prediction error in l_2 norm requires some assumptions that connect $\|Z\Delta\|_2^2$ with $\|\Delta\|_1^2$ or $\|\Delta\|_2^2$ for Δ included

in $\mathcal{C}(\delta, \mathcal{J}) := \{\Delta \in \mathbb{R}^K : \|\Delta_{\mathcal{J}^c}\|_1 \leq \delta \|\Delta_{\mathcal{J}}\|_1\}$, which is restricted vector space naturally satisfied by l_1 penalized estimator. For $\hat{\gamma}^L$ which denotes the solution via Lasso, from the optimization problem (1.4) with l_1 penalty,

$$\begin{aligned}
\|Z(\hat{\gamma}^L - \gamma^*)\|_2^2/2 &\leq \lambda(\|\gamma^*\|_1 - \|\hat{\gamma}^L\|_1)/2 + \varepsilon^T Z(\gamma^* - \hat{\gamma}^L) \\
&\leq \lambda(\|\gamma^* - \hat{\gamma}^L\|_1)/2 + \varepsilon^T Z(\gamma^* - \hat{\gamma}^L) \\
&\leq \lambda(\|\gamma^* - \hat{\gamma}^L\|_1)/2 + \|Z^T\varepsilon\|_\infty \cdot \|\gamma^* - \hat{\gamma}^L\|_1 \\
&\leq \lambda(1/2 + c_0)(1+\delta)\|(\hat{\gamma}^L - \gamma^*)_{\mathcal{J}^*}\|_1 \\
&\leq \lambda(1/2 + c_0)(1+\delta)\sqrt{J^*}\|(\hat{\gamma}^L - \gamma^*)_{\mathcal{J}^*}\|_2
\end{aligned}$$

where $\|Z^T\varepsilon\|_\infty \leq c_0$ and $\lambda \propto \sqrt{\log K}$. Note that we use the fact that $\max_{i\in[K]} Z_i\varepsilon/\sqrt{2\log K} \to 1$ *a.s.* and $\hat{\gamma}^L - \gamma^* \in \mathcal{C}(\delta, \mathcal{J}^*)$ for some constant δ (which is known as 3 for Lasso), and Caucy-Schwarz at the last inequality. Under the assumption called RE condition [2, 38]

$$RE_2(\delta, \mathcal{J}^*) := \inf\{\|Z\Delta\|_2/(\|\Delta_{\mathcal{J}^*}\|_2\sqrt{N}) : \Delta \in \mathcal{C}(\delta, \mathcal{J}^*)\} > 0,$$

the prediction error is upper bounded as

$$\|Z(\hat{\gamma}^L - \gamma^*)\|_2 \leq \lambda/\sqrt{N}(1 + 2c_0)\delta\sqrt{J^*}/RE_2(\delta, \mathcal{J}^*).$$

Similarly the estimation error in l_2 norm can be bounded by the optimal rate under the condition by using the prediction error bound. The cone of vector space $\mathcal{C}(\delta, \mathcal{J})$ is defined naturally from the optimization problem itself. For instance, in [38] the cone is defined in the following way. Since KKT condition for Lasso is $Z_j^T Z(\hat{\gamma}^L - \gamma^*) = Z_j^T\varepsilon - \lambda\hat{s}_j$ for $\hat{s}_j = sign(\hat{\gamma}_i^L), \forall j \in \hat{\mathcal{J}}$ or $\hat{s}_j \in [-1, 1]$, otherwise. Therefore,

$$\|Z(\hat{\gamma}^L - \gamma^*)\|_2^2 = (\hat{\gamma}^L - \gamma^*)_{\hat{\mathcal{J}}\setminus\mathcal{J}^*}(Z^T_{\hat{\mathcal{J}}\setminus\mathcal{J}^*}\varepsilon - \lambda) + (\hat{\gamma}^L - \gamma^*)_{\mathcal{J}^*}(Z^T_{\hat{\mathcal{J}}\setminus\mathcal{J}^*}\varepsilon - \lambda\hat{s}_{\mathcal{J}^*})$$

and

$$\|(\hat{\gamma}^L - \gamma^*)_{\hat{\mathcal{J}}\setminus\mathcal{J}^*}\|_1 \leq \frac{\|Z^T\varepsilon\|_\infty + \lambda}{\|Z^T\varepsilon\|_\infty - \lambda}\|(\hat{\gamma}^L - \gamma^*)_{\mathcal{J}^*}\|_1.$$

of which the coefficient can be bounded with high probability ($\|Z^T\varepsilon\|_\infty$ is function of $\log K$). On the other hand, for l_0 regularization, we cannot construct the sub-vector space based on l_1 norm, instead from the result of oracle bound we can constrain the vector space to satisfy (3.4) such that

$$\mathcal{C}^0_{(\delta,\mathcal{J})} := \{\Delta \in \mathbb{R}^K : \|\Delta_{\mathcal{J}^c}\|_0 \leq \delta\|\Delta_{\mathcal{J}}\|_0\}.$$

Since the eigen value condition-the smallest eigen value of gram matrix corresponding to some specific vectors should be positive- on a restricted vector space is much easier to be satisfied than on the unrestricted space $\mathbb{R}^K$. Since

$$\|\Delta\|_1/\max|\Delta| \leq \|\Delta\|_0 \leq \|\Delta\|_1/\min|\Delta|$$

$\mathcal{C}^0(\delta, \mathcal{J}^*)$ implies $\mathcal{C}(\delta, \mathcal{J}^*)$, if $\max|\Delta_{\mathcal{J}^{*c}}| \leq \min|\Delta_{\mathcal{J}^*}|$ for the same δ. Note that under RIF condtion the infinity norm of estimation error is bounded at $O(\lambda)$ and therefore if $\min|\Delta_{\mathcal{J}^*}| = \min|(\hat{\gamma} - \gamma^*)_{\mathcal{J}^*}| \geq O(\lambda)$ then the previous condition is satisfied. That is, for $q = 2$, if $\min|\gamma^*| \geq O(\lambda)$ and RIF condition is satisfied then $\Delta \in \mathcal{C}(\delta, \mathcal{J}^*)$ and since

$$0 < \frac{|\mathcal{J}^*|^{1/2}\|\Sigma\Delta\|_\infty}{N\|\Delta\|_2} \leq |\mathcal{J}^*|^{1/2}\frac{\max\|Z_i\|_2\|Z\Delta\|_2}{N\|\Delta\|_2} \leq |\mathcal{J}^*|^{1/2}\frac{\|Z\Delta\|_2}{\sqrt{N}\|\Delta_{\mathcal{J}^*}\|_2},$$

the RE condition is satisfied. However, we cannot say one condition implies the other in general setup, even though practically RE condition is known as harder one to be satisfied than CIF which is a version of RIF defined on different vector space [43]. The different version of restrictions on design matrix required for upper bounding the errors are discussed in [35].

As we discussed in Remark 3.1, the RIF condition using $\|Z^T Z(\hat{\gamma} - \gamma^*)\|_\infty$ as the numerator gives the optimal error bound. Indeed it is invented to make the optimal result came out, therefore we can say the condition may be more tight than the others.

Also note that the denominator of RE constant is restricted to the true support set $\mathcal{J}^*$ for bounding estimation error. Therefore if the true cardinality is close to K then the value is closer to zero. Note that the denominator of RIF is not restricted to the true support set. As we mentioned , there is no relationship such that one implies the other, but empirically RIF is more relaxed than RE. We calculate the values on a setup (actually on a multivariate setup, where dimensionality is 1000, sample size is 100, and correlation is high such that $corr(X_i, Xj) = 0.9^{|i-j|}$), and if the true cardinaltiy is 5 then RE = 0.23 and RIF = 1.31. However the true cardinality is 1000 then RE=0.77 and RIF=80.01. Since RE in both cases is close to zero and it is included in the upper bound of estimation error as a denominator, so the smaller constant makes the upper bound tends to infinity. Note that RIF has larger value especially when the true cardinality is high, so L_0 can promise a better estimation accuracy.

3.4 Selection

The main purpose of high dimensional analysis is to select the sparse true relevant subset consistently. The definition of selection consistency is $P(\mathcal{J}^* = \hat{\mathcal{J}}) \to 0$ and sign consistency is described as $P(sign(\hat{\gamma}) = sign(\gamma^*)) \to 0$. For Lasso, the sign consistency is achieved under some stringent conditions about the incoherence of design matrix. The conditions are relaxed for l_0 regularization still achieving the consistent selection. The selecting behavior of a fixed point of hard thresholding is shown from the extended of KKT condition (which is derived from the definition of the general sign of Θ_H estimator. She [29] showed the probability of selecting with sign agreement. Later in [43], the sign consistency of a global solution is shown by comparing with the restricted LSE. It achieves sign consistency when the strength of restricted LSE is bounded below at the rate of λ. Here we study the selection performance of local solutions in the similar way as done in [29] and compare the rate of convergence and conditions with those of Lasso.

Lemma 3.1 *[29] Let* $\tau := d_{min}(\Sigma_{\mathcal{J}^*})$ *and* $\kappa_1 := \|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty/\sqrt{J^*}$. *Suppose that* $\min_{j\in[\mathcal{J}^*]}|\gamma_j^*| > \lambda(1+\eta)k_0$, *where* $\eta \le (\sqrt{J^*}\kappa_1)^{-1}$. *There exist a solution* $\hat{\gamma}$ *satisfying the following probability,*

$$P[sgn(\gamma^*) = sgn(\hat{\gamma})]$$
$$\ge \left[1 - 2\Phi\left(\frac{\eta\lambda k_0\sqrt{\tau}}{\sigma}, +\infty\right)\right]^{J^*} \cdot \left[1 - 2\Phi\left(\frac{\lambda k_0(1-\eta\kappa_1\sqrt{J^*})}{\sigma\sqrt{N}}, +\infty\right)\right]^{K-J^*}.$$

Remark 3.2
Suppose $\lambda = \sigma\sqrt{\log K}$, *since* λ *is chosen to minimize the model error. Let* $\eta = \eta_2/(\kappa_1\sqrt{J^*})$ *for any* $\eta_2 \in (0,1)$. *If* $\min_{j\in[\mathcal{J}^*]}|\gamma_j^*| > \lambda k_0(1+\eta)$ *then*

$$P[sgn(\gamma^*) = sgn(\hat{\gamma})]$$
$$\ge \left[1 - 2\Phi(\frac{\eta_2\sqrt{\tau}}{\sqrt{J^*}}\frac{\sqrt{\log K}}{k_0}, +\infty)\right]^{J^*} \cdot \left[1 - 2\Phi((1-\eta_2)\sqrt{\log K}, +\infty)\right]^{K-J^*} \quad (3.13)$$

We used that $k_0^2 = \|\Sigma\|_2 \ge \|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_2 \ge \|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty/\sqrt{J^*} \ge \kappa_1$ *on the first probability bound in the previous lemma. Since* $N = \|Z\|_2^2 = \|UDV_i^T\|_2^2 = \sum D_i^2 V_i V_i^T \le k_0^2$ *the second term has the lower bound. Note that to achieve the estimation consistency,* $\log K = o(N)$ *and therefore* $\log K = o(k_0^2)$ *as well. Both terms are non-decreasing along with* K, *since the slope of the first part is* $J^* exp(-\log K/k_0^2) g_1(K)^{J^*-1}$ *where* $g_1(K)$ *tends to zero, and for the second part,* $(1 - J^*/K) g_2(K)^{K-J^*-1}$ *where* $g_2(K)$ *tends to one. Therefore, if*

$\log K = O(N)$ *as* N *tends to infity then the sign agreement is achieved with the probability tends to one. Note that we assumed* $\kappa_1\sqrt{J^*} = \|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty$ *is fixed as a constant. If* $\|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty = O(\tau/\sqrt{J^*})$ *then the first term is lower bounded by* $\left[1 - 2\Phi(\eta_2\sqrt{\log K}/\sqrt{\tau})\right]^{J^*}$ *and the second term is* $[1 - 2\Phi((1-\eta c\tau)\sqrt{K}, +\infty)]^{K-J^*}$*. Therefore, if* $\mathcal{J}^*$ *is getting larger and therefore the smallest eigen value of* $Z_{\mathcal{J}^*}$ *is getting smaller, the both terms are tending to one so the probability tends to one faster. The rate of* κ_1 *implies* $\kappa := \max_{j\in[\mathcal{J}^{*c}]}\|\Sigma_{j,\mathcal{J}^*}\|_2 \gtrsim O(\tau/J^*)$ *which is the opimal rate for obtaining selection cosistency via Lasso. In this setup, SNR is at the rate of* $O(k_0 \log K/\tau)$ *where* $1/\tau$ *increase as* J^* *increases, but not linearly related. Note that this condition is not requied for* l_0 *penalty, although it makes the covergence faster. When* $\log K = o(N)$ *and* $\|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\| = O(1)$*, it achieve the sign consistency. In this case the signal-to-noise ratio is provided at* $O(k_0 \log K)$ *which does not involve* J^**.*

Remark 3.3 *Note that both (3.14) and (3.15) do not require stringent incoherence condition. Since the probability of* $\|Z_{\mathcal{J}^{*c}}{}^T(Z_{\mathcal{J}^*}\Sigma_{\mathcal{J}^*}^{-1}Z_{\mathcal{J}^*}{}^T - I)\varepsilon\|_\infty$ *is smaller than the probability of* $\|Z_{\mathcal{J}^{*c}}^T\varepsilon\|_\infty$*, neither of two conditions do not involve any restriction on the incoherence, say* $\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}$*. Therefore,* l_0 *regularization requires the restriction on the overall correlation between covariates only, but not on the correlation between the irrelevant group and relevant group. We will show that for Lasso, the sign consistency is obtained under a small incoherence condition, that is when the relevant variables can be easilly separated from the rest.*

Proof. *Theorem 3.1* It is clear that a Θ_H estimator $\hat{\gamma}$ achieves sign consistency iff $\mathcal{J}^* = \hat{\mathcal{J}}$ and $|(\hat{\gamma} - \gamma^*)_j| < |\gamma_j^*|$, $\forall j \in \mathcal{J}^*$. From the extended KKT conditions,

$$\Sigma_{\mathcal{J}^*}(\hat{\gamma} - \gamma^*)_{\mathcal{J}^*} + \Sigma_{\mathcal{J}^*,\mathcal{J}^{*c}}\hat{\gamma}_{\mathcal{J}^{*c}} - Z_{\mathcal{J}^*}^T\epsilon - \lambda k_0\hat{S}_{\mathcal{J}^*} = 0$$
$$\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}(\hat{\gamma} - \gamma^*)_{\mathcal{J}^*} + \Sigma_{\mathcal{J}^{*c}}\hat{\gamma}_{\mathcal{J}^{*c}} - Z_{\mathcal{J}^{*c}}^T\epsilon - \lambda k_0\hat{S}_{\mathcal{J}^{*c}} = 0.$$

It is clear that $\mathcal{J}^* = \hat{\mathcal{J}}$ if $\hat{S}_{\mathcal{J}^{*c}} \in [-1, 1]$, $\hat{S}_{\mathcal{J}^*} = 0$ and $\hat{\gamma}_{\mathcal{J}^{*c}} = 0$, $\min_{\mathrm{j}\in\mathcal{J}^*}|\hat{\gamma}_j| > \lambda k_0$ from (2.7). For sign agreement, $|\gamma_j^*| > |(\hat{\gamma} - \gamma^*)_j|$ is required. We assume that $\Sigma_{\mathcal{J}^*}$ in nonsingular. If $\hat{\gamma}_{\mathcal{J}^{*c}} = 0$ and $\hat{S}_{\mathcal{J}^*} = 0$ then

$$(\hat{\gamma} - \gamma^*)_{\mathcal{J}^*} = \Sigma_{\mathcal{J}^*}^{-1}Z_{\mathcal{J}^*}^T\epsilon$$
$$\{\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\Sigma_{\mathcal{J}^*}^{-1}Z_{\mathcal{J}^*}^T - Z_{\mathcal{J}^{*c}}^T\}\epsilon = \lambda k_0\hat{S}_{\mathcal{J}^{*c}}.$$

Thus, if we have

$$|\gamma_j^*| > |((\Sigma_{\mathcal{J}^*})^{-1}Z_{\mathcal{J}^*}^T)_j\epsilon| + \lambda k_0, \ \forall j \in \mathcal{J}^*$$

then $|\hat{\gamma}_j| > \lambda/k_0$ and $|\gamma^*_j| > |(\hat{\gamma} - \gamma^*)_j|$ for all $j \in \hat{\mathcal{J}}$. Therfore, if $\hat{\gamma}$ satisfies

$$\min_{j \in \mathcal{J}^*} |\gamma^*_j| > \|\Sigma^{-1}_{\mathcal{J}^*} Z^T_{\mathcal{J}^*} \epsilon\|_\infty + \lambda k_0 \tag{3.14}$$

$$\|\{\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*} \Sigma^{-1}_{\mathcal{J}^*} Z^T_{\mathcal{J}^*} - Z^T_{\mathcal{J}^{*c}}\}\epsilon\|_\infty \leq \lambda k_0 \tag{3.15}$$

then $\hat{\gamma}$ selects the true subset and achieve sign agreement as well. Since

$$\begin{aligned}\{\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*} \Sigma^{-1}_{\mathcal{J}^*} Z^T_{\mathcal{J}^*} - Z^T_{\mathcal{J}^{*c}}\}\epsilon\|_\infty &\leq \|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*} \Sigma^{-1}_{\mathcal{J}^*} Z^T_{\mathcal{J}^*} \varepsilon\|_\infty + \|Z^T_{\mathcal{J}^{*c}} \epsilon\|_\infty \\ &\leq \|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty \cdot \|\Sigma^{-1}_{\mathcal{J}^*} Z^T_{\mathcal{J}^*} \varepsilon\|_\infty + \|Z^T_{\mathcal{J}^{*c}} \epsilon\|_\infty\end{aligned}$$

if $\|\Sigma^{-1}_{\mathcal{J}^*} Z^T_{\mathcal{J}^*} \varepsilon\|_\infty \leq O(\lambda)$, $\|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty \leq O(1)$, and $\|Z^T_{\mathcal{J}^{*c}} \varepsilon\|_\infty \leq O(\lambda)$ then SNR has the optimal rate $O(\lambda)$ and the second condition (3.15) holds. Let UDU^T be SVD of $\Sigma_{\mathcal{J}^*}$, then ith diagonal element of $\Sigma^{-1}_{\mathcal{J}^*}$ is $\sum_{j \in [J^*]} D_j^{-1} u_{ij}^2 \leq 1/\tau$, where τ is the smallest eigenvalue of $\Sigma_{\mathcal{J}^*}$. If $\|\Sigma^{-1}_{\mathcal{J}^*} Z^T_{\mathcal{J}^*} \varepsilon\|_\infty \leq \lambda \eta k_0$ and $\|Z^T_{\mathcal{J}^{*c}} \epsilon\|_\infty \leq \lambda k_0 (1 - \eta \kappa_1 \sqrt{J^*})$ where $\kappa_1 := \|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty / \sqrt{J^*} = \max_{j \in \mathcal{J}^{*c}} \|\Sigma_{j,\mathcal{J}^*}\|_1 / \sqrt{J^*}$, then (3.15) is satisfied and SNR has the optimal rate as $O(\lambda)$. Since $Z^T_{\mathcal{J}^{*c}} \varepsilon \sim N(0, \sigma^2 \Sigma_{\mathcal{J}^{*c}})$, $\Sigma^{-1}_{\mathcal{J}^*} Z^T_{\mathcal{J}^{*c}} \varepsilon \sim N(0, \sigma^2 \Sigma^{-1}_{\mathcal{J}^*})$, and $diag_i(\Sigma_{\mathcal{J}^{*c}}) = \|Z_i\|_2^2 = N$ the following probabilities are derived. We use Lemma A.3 in [32] as well.

$$P\left[\|\Sigma^{-1}_{\mathcal{J}^*} Z^T_{\mathcal{J}^*} \varepsilon\|_\infty \leq \eta \lambda k_0\right] \geq \left[1 - 2\Phi\left(\frac{\eta \lambda k_0 \sqrt{\tau}}{\sigma}, +\infty\right)\right]^{J^*}$$

$$P\left[\|Z^T_{\mathcal{J}^{*c}}\|_\infty \leq \lambda k_0 (1 - \eta \kappa_1 \sqrt{J^*})\right] \geq \left[1 - 2\Phi\left(\frac{\lambda k_0 (1 - \eta \kappa_1 \sqrt{J^*})}{\sigma \sqrt{N}}, +\infty\right)\right]^{K - J^*}.$$

Then since $\|\Sigma^{-1}_{\mathcal{J}^*} Z^T_{\mathcal{J}^*} \varepsilon\|_\infty \leq \eta \lambda k_0$ with high probability, the SNR condition, $\min_{j \in \mathcal{J}^*} |\gamma^*_j| > \lambda k_0 (1 + \eta)$, for η chosen small enough that $\eta \leq (\sqrt{J^*} \kappa_1)^{-1} = \|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty$. ■

3.5 Comparison

The incoherence constraints are required for achieving selection consistency, more specifically speaking the correlation between the relevant subset and the irrelevant subset should be small enough to get separated easily. Also, the signal should be strong enough compared to noise. Since these two condition work not independently, when one is weak then the other condition gets stringent. The representative restrictions on design matrix are the irrepresentable condition [46, 36] and the mutual coherence condition [12, 23, 4]. First, we will show how they affect on SNR to see the relationship between them. As briefly mentioned in Introduction, convex penalties require less restricted assumptions for achieving the optimal SNR. For now consider l_1-norm penalty $p(t; \lambda) = \lambda |t|_1$ in (2.1). For avoiding ambiguity, we

let $\hat{\beta}$ and β^* be a solution to the problem and the true coefficient vector. For the proof, we study KKT condition (2.9) for Lasso which holds for a global solution $\hat{\beta}$ when $\hat{S}$ is its subgradient vector [36]. In addition, if the subgradient holds $\hat{s}_j < 1\ \forall j \in \mathcal{J}^*$ then $\hat{\beta}_{\mathcal{J}^*} = 0$. The sign agreement holds, i.e. $sign(\hat{\beta}_{\mathcal{J}^*}) = sign(\beta^*_{\mathcal{J}^*})$ and $\hat{\beta}_{\mathcal{J}^{*c}} = 0$, if and only if the following conditions hold elementwise.

$$\begin{aligned}
&\Sigma_{\mathcal{J}^*}(\hat{\beta} - \beta^*)_{\mathcal{J}^*} - X^T_{\mathcal{J}^*}\varepsilon = -\lambda sign(\beta^*_{\mathcal{J}^*}) \\
&|\hat{\beta}_{\mathcal{J}^*} - \beta^*_{\mathcal{J}^*}| < |\beta^*_{\mathcal{J}^*}| \\
&|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}(\hat{\beta} - \beta^*)_{\mathcal{J}^*} - X^T_{\mathcal{J}^{*c}}\varepsilon| < \lambda
\end{aligned}$$

since $\hat{s}_j \in (-1, 1)$, $\forall j \in \mathcal{J}^{*c}$, or $\hat{s}_j = sign(\beta^*_j)$, otherwise for Lasso. That is, if

$$|(\Sigma_{\mathcal{J}^*})^{-1}X^T_{\mathcal{J}^*}\varepsilon| < |\beta^*_{\mathcal{J}^*}| - \lambda|(\Sigma_{\mathcal{J}^*})^{-1}sign(\beta^*_{\mathcal{J}^*})| \tag{3.16}$$

$$\begin{aligned}
&|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}(\Sigma_{\mathcal{J}^*})^{-1}X^T_{\mathcal{J}^*}\varepsilon - X^T_{\mathcal{J}^{*c}}\varepsilon| \\
&\qquad < \lambda(1 - |\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}(\Sigma_{\mathcal{J}^*})^{-1}sign(\beta^*_{\mathcal{J}^*})|\ )
\end{aligned} \tag{3.17}$$

then sign-consistency is achieved. Therefore, it is required to satisfy

$$(i)\ \min_{j\in\mathcal{J}^*}|\beta^*_j| > \lambda\|(\Sigma_{\mathcal{J}^*})^{-1}sign(\beta^*_{\mathcal{J}^*})\|_\infty, \tag{3.18}$$

$$(ii)\|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}(\Sigma_{\mathcal{J}^*})^{-1}sign(\beta^*_{\mathcal{J}^*})\|_\infty < 1, \tag{3.19}$$

and the probability of sign agreement is product of the probability of (3.16) and probablity of (3.17). Note that if $\|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty \cdot \|(\Sigma_{\mathcal{J}^*})^{-1}sign(\beta^*_{\mathcal{J}^*})\|_\infty < 1$ is satisfied with $\|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty = O(1)$ then SNR, the right hand side of (3.18) has the optimal rate as $O(\lambda)$.

Zhao and Yu [46] proposed the irrepresentable condition, which is equavalent to (3.19), and showed sign consistency under the irrepresentable condition and $\|\Sigma_{\mathcal{J}^*}\|_2 \geq M_2$ for some positive constant M_2. Sign consistency is achieved with SNR provided at $O(\sqrt{J^*}\lambda)$. Since under those conditions

$$\begin{aligned}
\lambda\|(\Sigma_{\mathcal{J}^*})^{-1}sign(\beta^*_{\mathcal{J}^*})\|_\infty &\leq \lambda \max_{j\in\mathcal{J}^*}\|(UD^{-1})^j\|_2 \cdot \|U^T sign(\beta^*_{\mathcal{J}^*})\|_2 \\
&\leq \lambda/M_2 \cdot \sqrt{J^*},
\end{aligned}$$

where $UD^{-1}U^T$ is SVD of $\Sigma^{-1}_{\mathcal{J}^*}$. Note that under the irrepresentable condition, $\|\Sigma^{-1}_{\mathcal{J}^*}\|_\infty \lesssim O(1)$ is required for having the optimal rate of SNR, since

$$\|(\Sigma_{\mathcal{J}^*})^{-1}sign(\beta^*_{\mathcal{J}^*})\|_\infty \leq \quad \|(\Sigma_{\mathcal{J}^*})^{-1}\|_\infty \|sign(\beta^*_{\mathcal{J}^*})\|_\infty \leq O(1) \cdot 1.$$

Later the mutual coherence condition is suggested as follows.

$$\mu := \max_{1\le i<j\le K} |\Sigma_{ij}/N| \le \frac{1}{7\alpha s}, \quad \text{for some } \alpha > 0$$

where $J^* < s$ and $7\alpha > 1$. Therefore the right hand side of (3.18) is upper bounded as follows.

$$\begin{aligned}
&\|(\Sigma_{\mathcal{J}^*})^{-1} sign(\beta^*_{\mathcal{J}^*})\|_\infty \\
&\le \|\Sigma_{\mathcal{J}^*}/N \cdot (\Sigma_{\mathcal{J}^*})^{-1} sign(\beta^*_{\mathcal{J}^*})\|_\infty + \|(I - \Sigma_{\mathcal{J}^*}/N)(\Sigma_{\mathcal{J}^*})^{-1} sign(\beta^*_{\mathcal{J}^*})\|_\infty \\
&\le \|sign(\beta^*_{\mathcal{J}^*})\|_\infty/N + \|I - \Sigma_{\mathcal{J}^*}/N\|_\infty \|(\Sigma_{\mathcal{J}^*})^{-1} sign(\beta^*_{\mathcal{J}^*})\|_\infty \\
&\le 1/N + \mu(J^* - 1) \cdot \|(\Sigma_{\mathcal{J}^*})^{-1} sign(\beta^*_{\mathcal{J}^*})\|_\infty \\
&\le 1/N + 1/(7\alpha) |(\Sigma_{\mathcal{J}^*})^{-1} sign(\beta^*_{\mathcal{J}^*})\|_\infty
\end{aligned}$$

since $1 - \Sigma_{j,j}/N = 0$, when $\|Z_i\|_2 = \sqrt{N}$. Then $\|(\Sigma_{\mathcal{J}^*})^{-1} sign(\beta^*_{\mathcal{J}^*})\|_\infty \le c/N$ for some constant $c > 0$. Therefore $\min|\hat{\beta}_{\mathcal{J}^*}| > c/N \cdot \lambda$, which is the optimal rate. In addition,

$$\begin{aligned}
\|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}(\Sigma_{\mathcal{J}^*})^{-1} sign(\beta^*_{\mathcal{J}^*})\|_\infty &\le \|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty \cdot \|(\Sigma_{\mathcal{J}^*})^{-1} sign(\beta^*_{\mathcal{J}^*})\|_\infty \\
&\le \mu J^* \cdot 7\alpha/(N(7\alpha - 1)) \le 1/(N(7\alpha - 1)) < 1,
\end{aligned}$$

so the second condition (3.19) is satisfied. Note that in this case $\|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^c}\|_\infty \le \mu J^* = O(1)$. Bunea [4] proposed the similar conditions as the mutual incoherence condition and showed sign consistency without post-thresholding. The need of post-thresholding which was proposed by [23] is to guarantee the uniqueness of such solution, so we will do post-thresholding by l_∞-norm estimation loss (3.7). From these fact we can see the condition of $\|\Sigma^{-1}_{\mathcal{J}^*}\|_\infty$ and $\|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty$ decide the rate of SNR required for sign agreed selection. Since $\|\Sigma^{-1}_{\mathcal{J}^*}\|_\infty$ should be bounded at the rate of $1/\|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty$, and if both are constant then SNR has the optimal rate.

Other assumptions regarded as less restricted than the irrepresentable condition and the conditions including RIP, SRC give SNR at the rate of $O(\sqrt{J^*}\lambda)$. They control spectrum of eigenvalues under sparsity restriction or rank restriction. Zhang [40] proposed Sparse Riesz Condition (SRC), under which Lasso achieves the selection consistency on the l_r sparsity restriction for $r > 0$, which is absolute sum of most effective coefficeints is smaller than a certain threshold. In that paper, the estimator selecting subset with the same rate as the true dimensionality would be uique under the SRC, but still the SNR is not the optimal rate. Both of strong irrepresentable condition and SRC are regarded weaker than the Mutual Incoherence condition, and SRC is even weaker than the irrepresentable condition up to constant, but there is no relationship among the above three conditions such that one

condition implies the others. It is said that the irrepresentable condition is almost necessary and sufficient condition for Lasso, and under more stringent condition (mutual incoherent condition) the SNR has the improved rate [44] .

SCIF is improved condition than mutural incoherence by up to constant. Nonconvex penalty possibly achives the optimal rate of SNR under less restricted conditions. Zhang [41] suggested MCP, concavity-restricted penalized estimator, and showed that under more relaxed condition than irrepresentable condition, indeed similar to SRC, selection consistency was achieved, provided SNR given by $O(\sqrt{\log K})$ via general nonconvex penalty satisfying concavity such that $1/\delta := \sup_{0<t_1<t_2}\{\dot{\rho}(t_1;\lambda)-\dot{\rho}(t_2;\lambda)\}/(t_2-t_1) < \min_{|A|\leq J^*} d_{min}(Z_A^T Z_A)$. Note that the concavity is defined as regarding to the smallest eigenvalu of $\Sigma_{\mathcal{J}^*}$. As shown in the previous section, SNR for l_0 regularization is $O(\lambda/\tau)$ which is smaller rate than $\lambda\delta$. Intuitively, since the minimum strength should be greater than $|(\hat{\gamma}-\gamma^*)_{\mathcal{J}^*}|$ and for nonconvex penalty the maximum of $|(\hat{\gamma}-\gamma^*)_{\mathcal{J}^*}|$ is decided when the slope of solution path is getting smaller, therefore it depends inversely on the concavity $1/\delta$. Zhang [45] suggested another nonconvex regulaization through multi-stage convex relaxation procedure, which reduce Lasso bias in each iteration by refining threshold based on solutions from the previous step. Selection consistency was shown under RIP condition-slightly restricted than SRC but only up to constant-and under SNR greater than $O(\sqrt{(1/N)\log K})$. The improvement compared to MCP is computational efficiency, such that only finite number of iteration for convex optimization are required. Based on these works, results on general nonconvex regularizations were shown by [43]. Under less restricted condition RIF, l_q loss has the optimal rate of $O(|\mathcal{J}^*|^{1/q}\tau)$ and selection consistency is achieved via the nonconvex regularizations.

In conclusion, for nonconvex penalty function achieves selection consistency under less restricted condition than Lasso, even with the optimal rate of SNR. Comparing l_0 as a nonconvex regularization to Lasso, we can see the aspect clearly. For τ is the smallest eigen value of $\Sigma_{\mathcal{J}^*}$, if $\|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty = O(\tau/\sqrt{J^*})$, (since $\|\Sigma^{-1}_{\mathcal{J}^*}\|_\infty \leq \|\Sigma^{-1}_{\mathcal{J}^*}\|_2\sqrt{J^*} = \sqrt{J^*}/\tau$), then SNR is provided as $O(\sqrt{J^*}\lambda/\tau)$. Else if $\|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty \leq O(1)$, then SNR is provided as $O(\lambda)$ which is the optimal rate. Note that as mentioned in the previous section, for l_0 regularization, if $\|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}\|_\infty \leq O(1)$ then it achieve the sign consistency and the convergence rate is fast with the optimal rate of SNR. However, if that is $O(\tau/\sqrt{J^*})$ then the SNR is $O(\lambda/\tau)$ which is smaller rate than $O(\lambda\sqrt{J^*})$ and the sign consistency is achieved. However, in l_0 case, there is no condition which should be satisfied as in Lasso, other than SNR condition.

CHAPTER 4

THEORIES ON THE $L_0 + L_2$ REGULARIZATION

In this chapter we study $l_0 + l_2$ regularization [29]. Our theories on L_0 regularization show that it always does a better job than L_1 regularization but there are issues caused by computation. Note that the regularization parameter λ of both methods are tuned in terms of prediction accuracy and therefore we cannot choose the optimal λ based on the selection accuracy. For example, if the signal strength is low but the correlation is high, then λ for L_0 regularization is chosen too large because it does not shrink large values. Therefore L_0 may under-select so it causes missed selection which is much serious problem than false alarm (which is error of over selection). There is only one regularization parameter that works both for selection and prediction so it cannot give the optimal result. This is an issue not only for L_0 but also for L_1 either, so for better results, two regularizations are needed to be separated.

The $l_0 + l_2$ regulariztion uses the combined penalty of l_0 and l_2, and there are two regularization parameters which control selection and prediction independently (in the computation). Also, according to James-Stein phenomenon (1961) shrinkage method gives smaller MSE and $L_0 + L_2$ which first selects a subset then shrinks the values, would give better solution (L_0 does not shrink after select subset). Indeed adding l_2 penalty increases stability of solution as its effect on the elastic-net [48], and therefore it works better than l_1 penalty on highly correlated covariates. Figure 1.1 shows if two covariates are almost identical (the angle of ellipses is $\pi/4$) then the solution is not unique (which is along with the edge of the square). In contrast, l_2 penalty has unique solution even when the covariates are highly correlated, although the solution is not sparse. Indeed, in use of elastic net, consistency in selection and prediction is guaranteed under a condition which is much more relaxed than that of Lasso, since conditions of restricted eigenvalue condition of the augmented design matrix, $Z^T Z + wI$, should be much relaxed than that of $Z^T Z$. Elastic Irrepresentable con-

dition (EIC) is compared with Irrepresentable condition in [21]. We expect the same effect of adding l_2 penalty, still keeping the advantage of l_0 which is sparse recovery (even better than L_1 on this aspect). Also, it will probably improve the rate of error in prediction or estimation as shown in the elastic net [18].

We consider the univariate model (1.3) here. The penalty function is defined as follows. For a diagonal matrix, $W \in \mathbb{R}^{K\times K}$ where $w_i > 0$, $\forall i \in [K]$,

$$\hat{\gamma} = \arg\min_{\gamma\in\mathbb{R}^K} F(\gamma), \qquad \text{where } F(\gamma) = \frac{1}{2}\|y - Z\gamma\|_2^2 + p(\gamma; W, \lambda) \tag{4.1}$$

where $p(\gamma; W, \lambda) = \frac{1}{2}\sum_{i=1}^K w_i\gamma_i^2 + \frac{\lambda^2}{2}\|\gamma\|_0$. When every coefficients of l_2 are the same, $p(\gamma; wI_K, \lambda) = \frac{1}{2}w\|\gamma\|_2^2 + \frac{\lambda^2}{2}\|\gamma\|_0$. The $l_0 + l_2$ penalty given in [30] is $p(\gamma; wI_K, \frac{\lambda}{\sqrt{1+w}})$ where l_0 coefficient is dependent on w. See if $\mu = 0$ then it is l_0 regularization.

4.1 Method

The $l_0 + l_2$ regularization still involves the numerical difficalties cause by l_0 penalty and the solution is not continuous as well, although the combined penalty funciton is convex lower semi continuous function. It is worth pointing out the property that $\liminf_{t\to 0} p(t; w, \lambda) \geq p(0; w, \lambda)$ and $\dot{p}(t; w, \lambda) = 0$ for $t = 0+$, or wt for $t > 0$. By use of this property, a local solution may be found as an approximate global solution under some conditions, which is our future plan. The numerical approach given in [30] for $l_0 + l_2$ penalty called Hybrid-TISP is applied. The hybrid hard-ridge-thresholding rule is defined by

$$\Theta(t; \lambda, \eta) = \begin{cases} 0, & \text{if} |t| < \lambda \\ \frac{t}{1+w}, & \text{if} |t| \geq \lambda \end{cases},$$

and if q-function in (2.3) is $q(\theta; \lambda, w) = (1 + w)/2(|\theta| - \lambda)^2 1_{0<|\theta|<\lambda}$, then the $l_0 + l_2$ penalty $p(\gamma; wI_K, \frac{\lambda}{\sqrt{1+w}})$ is obtained. We consider the following optimization problem.

$$F(\gamma) = \frac{1}{2}\|y - Z\gamma\|^2 + \frac{w}{2}\|\gamma\|^2 + \frac{\lambda^2}{2(1+w)}\|\gamma\|_0 \tag{4.2}$$

For the understanding of Hybrid thresholding rule, we apply the optimization problem to hard thresholding. Define $g(\gamma, \gamma')$ for any $\gamma, \gamma' \in \mathbb{R}^K$,

$$g(\gamma, \gamma') = \frac{1}{2}\|y - Z\gamma\|_2^2 + \frac{w}{2}\|\gamma\|_2^2 + \frac{\lambda^2}{2(1+w)}\|\gamma\|_0 + \frac{1}{2}(\gamma - \gamma')^T((1-w)I - Z^TZ)(\gamma - \gamma').$$

Then for given γ', minimizing g over γ is finding minimizer of g_1 where

$$g_1(\gamma) = \frac{1}{2}\|\gamma - \{Z^Ty + ((1-w)I - Z^TZ)\gamma'\}\|_2^2 + \frac{\lambda^2}{2(1+w)}\|\gamma\|_0.$$

Similarly, minimizing g over γ' for given γ is equivalent to solving $g_2(\gamma') =< ((1-w)I - Z^TZ)\gamma', \gamma' - 2\gamma >$. The solution to the second step is $\gamma' = \gamma$, when $\|Z^TZ\|_2 < 1$. The hard thresholding gives the solution such that $\hat{\gamma} = \Theta_H((1-w)I\hat{\gamma} - \Sigma\hat{\gamma} + Z^Ty : \lambda/\sqrt{1+w})$. Since $\arg\min F(\gamma)$ is equavalent to $\arg\min \frac{1}{1+w}F(\gamma)$, the hard thresholding of scaled problem is

$$\hat{\gamma} = \frac{1}{1+w}\Theta_H(\hat{\gamma} - \Sigma\hat{\gamma} + Z^Ty : \lambda) \tag{4.3}$$

$$= \Theta(\hat{\gamma} - \Sigma\hat{\gamma} + Z^Ty : \lambda, w). \tag{4.4}$$

From the definition of generalized sign, $\hat{\gamma}_i - Z_i^TZ\hat{\gamma} + Z_i^Ty = (1+w)\hat{\gamma}_i + \lambda\hat{s}_i$, where $\hat{s}_i = 0$ for $i \in \hat{\mathcal{J}}$ or $\hat{s}_i \in [-1,1]$ otherwise. We define a new general sign $\hat{S}_H$ corresponding to Hybrid thresholding as (2.8) then

$$\lambda\hat{s}_{H,i} = -Z_i^TZ\hat{\gamma} + Z_i^Ty = w\hat{\gamma}_i + \lambda\hat{s}_i.$$

Since if $|\hat{\gamma}_j^{(l-1)} - Z_j^TZ\hat{\gamma}^{(l-1)} + Z_j^Ty| < \lambda$ then $\hat{\gamma}_j^{(l)} = 0$ and $\hat{s}_j \in [-1,1]$, so $\hat{s}_{H,i} \in [-1,1]$ as well. Otherwise, $\hat{\gamma}^l = w^{-1}(-Z_j^TZ\hat{\gamma}^{(l-1)} + Z_j^Ty)$ and $\hat{s}_j = 0$, but if $|\hat{\gamma}_j^{(l)} - Z_j^TZ\hat{\gamma}^{(l-1)} + Z_j^Ty| = (1+w)|\hat{\gamma}^{(l)}| < \lambda$, then $\hat{\gamma}^{(l+1)} = 0$ and therefore $\hat{s}_{H,i} = 0$. Otherwise $\hat{s}_{H,i} = w/\lambda\hat{\gamma}_i^{(l+1)}$ which is the general sign of Ridge thresholding. Therefore,

$$\hat{s}_{H,i} = \begin{cases} \in [-1,1], & \text{if } \hat{\gamma}_j = 0 \\ 0, & \text{if } |\hat{\gamma}_j| \in (0, \frac{\lambda}{1+w}) \\ \frac{w}{\lambda}\hat{\gamma}_j, & \text{if } |\hat{\gamma}_j| \geq \frac{\lambda}{1+w} \end{cases}.$$

For the limit point of $\hat{\gamma}^l$ to be a fixed point of the procedure, the largest eigen value of Z^TZ smaller than 1. Therefore we consider another optimization problem with penalty coefficients replace by w/k_0^2 and λk_0^2 respectively to match the scaling factor. Since

$$\begin{aligned} \hat{\gamma} &= \arg\min_{\gamma\in\mathbb{R}^K} \left\{ \frac{1}{2k_0^2}\|y - Z^T\gamma\|_2^2 + \frac{w}{2k_0^2}\|\gamma\|_2^2 + \frac{\lambda^2/k_0^4}{2(1+w/k_0^2)}\|\gamma\|_0 \right\} \\ &= \arg\min_{\gamma\in\mathbb{R}^K} \left\{ \frac{1}{2(k_0^2+w)}\|y - Z^T\gamma\|_2^2 + \frac{w}{2(k_0^2+w)}\|\gamma\|_2^2 + \frac{\lambda^2}{2(k_0^2+w)^2}\|\gamma\|_0 \right\} \end{aligned} \tag{4.5}$$

the corresponding thresholding estimate is

$$\begin{aligned} \hat{\gamma} &= \Theta_H((1 - \frac{w}{k_0^2+w})\hat{\gamma} - \frac{1}{k_0^2+w}\Sigma\hat{\gamma} + \frac{1}{k_0^2+w}Z^Ty; \lambda\frac{\lambda}{k_0^2+w}) \\ &= \frac{1}{1+w/k_0^2}\Theta_H(\hat{\gamma} - \frac{\Sigma\hat{\gamma}}{k_0^2} + \frac{Z^Ty}{k_0^2}; \frac{\lambda}{k_0^2}) \\ &= \Theta_H(\hat{\gamma} - \frac{\Sigma\hat{\gamma}}{k_0^2} + \frac{Z^TZy}{k_0^2}; \frac{\lambda}{k_0^2}, \frac{w}{k_0^2}). \end{aligned}$$

Note that in each iteration of Hybrid-thresholding, the l_0 coefficient λ works for selection only but not for shrinkage while l_2 penalty coefficient w works for proportional shrinkage only. Since two factors perform independently for different purposes, each iteration can be separated into two step, the first one is to select non-zero subset and the second one is to shrink proportionally, which is the same performance of the Ridge-thresholding. Therefore, we study Ridge estimator first on the case where the target vector is not sparse, focusing on finding the optimal coefficient of l_2-penalty. After that by combining the selection part, we can get a insight into the $l_0 + l_2$ regularization.

4.2 Empirical Weight of L_2-penalty

In this section, we study the following l_2 regularization on the linear model (1.3). We assume the l_2 regularization on the selected subset, therefore the true vector is not sparse and the gram matrix is nonsingular so that the OLS can be defined.

$$G(\gamma) = \arg\min_{\gamma}\{\|y - Z\gamma\|^2/2 + w/2\|\gamma\|^2\}$$

The solution to the above optimization is $\hat{\gamma}_{Ridge} = (Z^TZ + wI)^{-1}Z^Ty$. The goal is to select the optimal w, which minimize the MSE :

$$\begin{aligned}
\mathbf{E}\left[\|Z\gamma^* - Z\hat{\gamma}\|_2^2\right] &= \mathbf{E}\left[\|Z(Z^TZ + wI)^{-1}Z^TZ\gamma^* - Z\gamma^* + Z(Z^TZ + wI)^{-1}Z^T\varepsilon\|_2^2\right] \\
&= \mathbf{E}\left[\|\{UD(D^TD + wI)^{-1}D^TU^T - I\}Z\gamma^* + UD(D^TD + wI)^{-1}D^TU^T\varepsilon\|_2^2\right] \\
&= \sum_{i\in[N]}\left(\frac{wD_i}{D_i^2 + w}\right)^2 (V_i^T\gamma)^2 + \sigma^2\sum_{i\in[N]}\left(\frac{D_i^2}{D_i^2 + w}\right)^2
\end{aligned}$$

where UDV^T is the spectral decomposition of Z. The first derivative respect to w gives

$$\frac{\partial\mathbf{E}\left[\|Z\gamma^* - Z\hat{\gamma}\|_2^2\right]}{\partial w} = \sum_{i\in[N]}(w(V_i^T\gamma^*)^2 - \sigma^2)\frac{D_i^4}{(D_i^2 + w)^3},$$

which is hard to solve and the solution involves the unknown parameter γ^*. Note that if we consider the different weigths and w_i is the ith coefficient, then the optimal weight should be $w_{i,opt} = \sigma^2/(V_i^T\gamma^*)^2$, which is approximately a noise to signal ratio.

Since the Ridge estimator is an ideal shrinkage estimator with non orthogonal design, so we borrow the idea of empirical Bayes to choose the unknown part $(Z^TZ + wI)^{-1}$ in the similar way which James-stein estimator(JSE) is given. For the understanding the

estimator, we discuss about the result in [9, 13] briefly. Consider the following simple linear model with identity design.

$$y = \mu + \epsilon$$

where $y \in \mathbb{R}^N$ and $\varepsilon \sim N(0, \sigma^2 I)$ when σ^2 is known. The MLE of μ is $\hat{\mu}_{MLE} = y$, and the corresponding MSE is $E[\|\mu - \hat{\mu}_{MLE}\|_2^2] = N\sigma^2$. Unlike the MLE, the shrinkage estimator, $\hat{\mu}_c = c \cdot y$, minimize MSE by trading off variance and bias. Since MSE of the $\hat{\mu}_c$ is $E[\|\mu - cy\|_2^2] = (1-c)^2\|\mu\|_2^2 + c^2\sigma^2 N$, the optimal choice of c to minimize the MSE is given by $c^* = \|\mu\|^2/(\|\mu\|^2 + N\sigma^2)$ and $MSE(\hat{\mu}_{c^*} = \frac{N\sigma^2\|\mu\|^2}{\|\mu\|^2 + N\sigma^2}$ which is smaller than that of MLE and the gap between both MSEs are greater when signal-to-noise ratio is small. It shows the shrinkage estimator performs well especially when the signal is weak. However $\hat{\mu}_{c^*}$ is ideal estimator which includes the unknown parameter. The JSE is data driven estimator applying empirical Bayes idea to the ideal shrinkage estimator. The $\hat{\mu}_{JSE}$ is defined as

$$\mu_{JSE} = \left(1 - \frac{(N-2)\sigma^2}{\|y\|_2^2}\right)_+ \cdot y$$

It is clear that $\inf_c E[\|\hat{\mu}_c - \mu\|^2] \leq E[\|\hat{\mu}_{JSE} - \mu\|^2$. However, it is shown that JSE works almost as good as the ideal shrinkage estimator such that

$$E[\|\hat{\mu}_{JSE} - \mu\|^2] \leq 4 + \inf_c E[\|\hat{\mu}_c - \mu\|^2],$$

and of course it has smaller MSE than MLE. Although it predicts poorly for the extreme individual values [13], JSE works better overall. The JSE is driven as the empirical Bayes estimator, when μ has its prior as Normal distribution.

$$y|\mu \sim N(\mu, \sigma^2 I), \quad \mu \sim N(0, AI)$$

and therefore, $\mu|y \sim N(A/(A+\sigma^2)y, A\sigma^2/(A+\sigma^2))$. Then the Bayes estimator would be $\hat{\mu}_{Bayes} = \left(1 - \frac{\sigma^2}{A+\sigma^2}\right) y$, which is one of the shrinkage estimator where $c = A/(A+\sigma^2)$ and its MSE as $(\sigma^2/(A+\sigma^2))^2\|\mu\|^2 + N(A\sigma)^2/(A+\sigma^2)^2$. We don't know the value of A, so the empirical Bayes ideas is applied. Since the marginal distribution of y is given by

$$y \sim N(0, (A+\sigma^2)I),$$

and therefore $\|y\|^2/(A+\sigma^2) \sim Gamma(N/2, 2)$ and $E[(N-2)/\|y\|^2] = 1/(A+\sigma^2)$. By substituting the unknown parameter by data-driven estimates $(N-2)/\|y\|^2$, we get the $\hat{\mu}_{JSE}$. Interestingly, the Bayes estimator is a Ridge estimator minimizing the following optimization problem for $w = \sigma^2/A$:

$$G(\mu) = \frac{1}{2}\|\mu - y\|_2^2 + \frac{w}{2}\|\mu\|_2^2$$

where $\hat{\mu}_{Ridge} = 1/(1+w)\cdot y$. From this we select $w = \left(1 - \frac{(N-2)\sigma^2}{\|y\|^2}\right)^{-1} - 1$, which includes the inverse of the coefficient in JSE. With the choice of w we get the MSE given by

$$E[\|\hat{\mu}_{JSE} - \mu\|^2] \leq N\sigma^2 - (N-2)^2\sigma^2/(\|\mu\|^2 + N\sigma^2) \leq E[\|\mu_{\hat{MLE}} - \mu\|^2].$$

We apply the ideas of JSE to our problem. First we consider non-sparse vector γ to our model where the dimension $K \leq N$, and the ridge estimator given by

$$\hat{\gamma}_{Ridge} := \arg\min_{\gamma} \left[\frac{1}{2}\|y - Z\gamma\|_2^2 + \frac{w}{2}\|\gamma\|_2^2\right]$$

then $\hat{\gamma}_{Ridge} = (Z^TZ + wI)^{-1}Z^Ty$. This estimator is the Bayes estimator when the prior of γ^* is given by

$$\gamma^* \sim N(0, \frac{\sigma^2}{w}I).$$

Since $y|\gamma^* \sim N(Z\gamma^*, \sigma^2 I)$ and

$$\begin{aligned} f(\gamma^*|y) &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\gamma^{*T}(Z^TZ + wI)\gamma^* - y^TZ\gamma^* + y^Ty\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\|(Z^TZ + wI)^{1/2}\{\gamma^* - (Z^TZ + wI)^{-1}Z^Ty\}\|^2\right\}, \end{aligned}$$

the posterior of γ^* is $N((Z^TZ + wI)^{-1}Z^Ty, \sigma^2(Z^TZ + wI)^{-1})$. Also, we have

$$\begin{aligned} f(y) \propto \int_{\mathbb{R}^K} &\exp\left\{-\frac{1}{2\sigma^2}\|(Z^TZ + \eta I)^{1/2}\{\gamma^* - (Z^TZ + wI)^{-1}Z^Ty\}\|^2\right\} \cdot \\ &\exp\left\{-\frac{1}{2\sigma^2}y^T(I - Z(Z^TZ + wI)^{-1}Z^T)y\right\} d\gamma^*, \end{aligned}$$

therefore the marginal distribution of y is given by

$$y \sim N(0, \sigma^2\{I_N - Z(Z^TZ + wI_K)^{-1}Z^T\}^{-1}).$$

Note that before applying the empirical Bayesian idea, we assume that σ is known therefore we need to estimate the unknown w only, by integrating the nuisance parameter γ^* out. Let UDV^T as the singular decomposition of Z, then

$$\begin{aligned} U_i^Ty &\sim N(0, \sigma^2(1 + \frac{D_i^2}{w}), \quad \forall i \in [K] \\ U_i^Ty &\sim N(0, \sigma^2), \quad \forall K < i \leq N \end{aligned}$$

therefore U^Ty follows N-dimensional independent Normal distribution. Since $E[\|U^Ty\|^2] = \sigma^2(N + \sum_{i\in[K]} tr(D^TD)/w)$, we approximate the Ridge regression as follows:

$$\hat{\gamma}_{JS} = (Z^TZ + \hat{w}I)^{-1}Z^Ty, \quad \text{where } \hat{w} = \left(\frac{tr(D^TD)}{\frac{1}{\sigma^2}\|y\|^2 - N}\right)_+. \tag{4.6}$$

where σ^2 is known. Since w works inversly in the estimator, we use the fact that $E[\frac{\|U^T y\|^2/\sigma^2 - N}{tr(D^T D)}] = 1/w$ to approximate $1/w$. Note that if $N \leq K$ then $U_i^T y \sim N(0, \sigma^2(1 + D_i^2/w))$, $\forall i \in [N]$, and w can be approximated by the same estimate since $E[(\|U^T y\|^2/\sigma^2 - N)/\sum_{i\in[N]} D_i^2] = 1/w$. The OLS is not defined in this case, and therefore we cannot compare its efficiency with that of OLS.

Theorem 4.1 *Consider the simple linear model (1.3) where $K \leq N$. For MLE of γ, $\hat{\gamma}_{MLE} = (Z^T Z)^{-1} Z^T y$, and James-Stein estimator (4.6), if*

$$\|y\|^2/\sigma^2 - N > 0 \tag{4.7}$$

and

$$\sum_{i\in[K]} \sum_{j\neq i} D_i^2 (U_j^T y)^2 \geq K(N - \sum_{i\in[K]} D_i^2)\sigma^2 \tag{4.8}$$

then

$$\begin{aligned} MSE[\hat{\gamma}_{JS}] &\leq \sigma^2 \left(K - \sum_{i\in[K]} \frac{tr(D^T D)}{D_i^2 \|Z\gamma^*\|^2/\sigma^2 + tr(D^T D)} \right) \\ &\leq MSE[\hat{\gamma}_{MLE}] \end{aligned}$$

Remark 4.1
(1) The approximate weight is $\hat{w} \approx \frac{\sigma^2}{\|Z\gamma^*\|^2/tr(D^T D)}$, which can be interpreted as noise/signal . If the signal is weak compared to the noise, the estimator is shrinken to zero by proportionally large amount so it cause the smaller variance. The empirical choice of w would be applied in the next chapter which minimize the oracle error bounds.
(2) The condition (4.8) does not give a tight upper bound of $MSE(\hat{\gamma}_{JS})$. (See the proof) That means we can make the condition more relaxed, even though it is not stringent condition already. Note that $\sum_{i\in[K]} \sum_{j\neq i} D_i^2 (U_j^T y) \sim \sum_{i\in[K]} T_i$, where $T_i = D_i^2 \sum_{j\neq i} \{(D_j V_j^T \gamma^*)^2 + \sigma^2\}$, which would be greater than $\sigma^2(N - \sum_{i\in[K]} D_i^2)$. Especially when the correlation is high so $\sum D_i^2$ is large, or K is close to N, the empirical estimator works much better than OLS.

Consider $l_0 + l_2$ regularization. At each iteration, a subset is selected and proportionally shrinken with the weight of $1/(1 + w)$. The empirical choice of w gives the insight into the regularization. If the underlying model is not sparse so the sparsity is closer to the sample size, or correlation among covariates is high, then the estimator would perform well. For every iterations, if the selected subset is changed, updating the corresponding w would

gives the smallest mse at the stage, therefore it might result in a faster convergence, than using a fixed w. For an iteration, if covariates corresponding to the selected subset by the l_0 threshold is highly correlated, $trace(D^TD)$ is large so l_2 penalty works more than l_0 at the iteration. At every iterations, the larger selected subset or the higher correlation among covariates result in the lower coefficient of l_0 regularization and the larger coefficient of l_2 penalty. Note that (4.1) involves σ, which is unknown. Therefore, by adjusting w and $\hat{\sigma}$ at each iteration, we can put more weight on one penalty than the other based on the currently selected covariates and the noise level. Note that if N is much greater than the dimension, (that is selected subset is sparse) then $\hat{w}$ is zero which means we do hard-thresholding at the iteration.

Since $\hat{w}$ is the same even when the cardinality is greater than the sample size, we can consider this estimate of w when the selected subset is greater than the sample size at each iteration. Algorithm is described here.

Remark 4.2 *(empirically adjusted hybrid thresholding.) For l th iteration,*
step (i) Select $\hat{\mathcal{J}}^{(l)}$ through hard thresholding for a given λ such that

$$\hat{\mathcal{J}}^{(l)} = \{j \in [K] : |\hat{\gamma}_{0,j}^{(l)}| > \lambda\}$$

where $\hat{\gamma}_{0,j}^{(l)} = \hat{\gamma}_j^{(l-1)} - \Sigma_{j,*}/k_0^2 \hat{\gamma}^{(l-1)} + Z_j^T y/k_0^2$.

step (ii) $\hat{\gamma}_{\hat{\mathcal{J}}^{(l)}}^{(l)} = \frac{1}{1+w^{(l)}} \hat{\gamma}_{0,\hat{\mathcal{J}}^{(l)}}^{(l)}$, and $\hat{\gamma}^{(l)} = 0$,
where $w^{(l)} = \{trace(D^TD)/(\|y\|^2/\hat{\sigma}_{(l-1)}^2 - N)\}_+$ for UDV^T is the singular value decomposition of $Z_{\hat{\mathcal{J}}^{(l)}}$.
step (iii) $\sigma_{(l)} = \|y - Z_{\hat{\mathcal{J}}^{(l)}} \hat{\gamma}_{\hat{\mathcal{J}}^{(l)}}^{(l)}\|^2/(N - |\hat{\mathcal{J}}^{(l)}|)$
Repeat step (i) to (iii) until convergence is met.

Proof. The MSE of $\hat{\gamma}_{JS}$ is derived in the similar way as in [9]. For U_i as ith column of U,

$$\begin{aligned}
&E[\|Z\hat{\gamma}_{JS} - Z\gamma^*\|^2] \\
&= E\left[\|UD(D^TD + \hat{w}I)^{-1}D^TU^Ty - Z\gamma^*\|^2\right] \\
&= E\left[\|\sum_{i\in[K]} U_i \frac{D_i^2}{D_i^2 + \hat{w}} U_i^T y - Z\gamma^*\|^2\right] \\
&= E\left[\|\sum_{i\in[N]} U_iU_i^Ty - Z\gamma^* - \sum_{i\in[K]} U_i \frac{\hat{w}}{D_i^2 + \hat{w}} U_i^T y - \sum_{i=K+1}^{N} U_iU_i^Ty\|^2\right] \\
&= E[\|y - Z\gamma^*\|^2] - 2E[(y - Z\gamma^*)^T g(y)] + E[\|g(y)\|^2] \qquad (4.9)
\end{aligned}$$

where $g(y) = \sum_{i\in[K]} U_i \frac{\hat{w}}{D_i^2+\hat{w}} U_i^T y + \sum_{i=K+1}^N U_i U_i^T y$. We use the following fact to calculate the second term of (4.9).

$$\begin{aligned}
&\int (y - Z\gamma^*)^T g(y) (2\pi\sigma^2)^{-1} \exp^{-\|y-Z\gamma^*\|^2/(2\sigma^2)} dy \\
&= \sigma^2 \int \nabla g(y) (2\pi\sigma^2)^{-1} \exp^{-\|y-Z\gamma^*\|^2/(2\sigma^2)} dy = \sigma^2 E[\nabla g(y)]
\end{aligned}$$

where $\nabla g(y) = \sum_{j\in[K]} \partial_j g_j(y)$. Therefore

$$\begin{aligned}
\nabla g(y) &= \nabla \left\{ \sum_{i\in[K]} U_i \frac{tr(D^T D)}{D_i^2(\frac{\|y\|^2}{\sigma^2} - N) + tr(D^T D)} U_i^T y + \sum_{i=K+1}^N U_i U_i^T y \right\} \\
&= \sum_{i\in[K]} \sum_{j\in[K]} U_{ij}^2 \frac{tr(D^T D)}{D_i^2(\frac{\|y\|^2}{\sigma^2} - N) + tr(D^T D)} \\
&- 2 \sum_{i\in[K]} \sum_{j\in[K]} U_{ij} \frac{tr(D^T D) D_i^2 y_j/\sigma^2}{(D_i^2(\frac{\|y\|^2}{\sigma^2} - N) + tr(D^T D))^2} + \sum_{i=K+1}^N U_{ij}^2 \\
&= \sum_{i\in[K]} \frac{tr(D^T D)}{D_i^2(\|y\|^2/\sigma^2 - N) + tr(D^T D)} - 2 \sum_{i\in[K]} \frac{tr(D^T D) D_i^2 (U_i^T y)^2/\sigma^2}{(D_i^2(\frac{\|y\|^2}{\sigma^2} - N) + tr(D^T D))^2} + (N-K).
\end{aligned}$$

Also the third term is

$$\begin{aligned}
&E[\|g(y)\|^2] \\
&= \sum_{i\in[K]} E\left[\left(\frac{tr(D^T D)}{D_i^2(\frac{\|y\|^2}{\sigma^2} - N) + tr(D^T D)} \right)^2 (U_i^T y)^2 \right] + \sum_{i=K+1}^N E(U_i^T y)^2
\end{aligned}$$

since $U_i^T U_j = 0$. Therefore from the above calculation of the second and third term, we have

$$\begin{aligned}
&(4.9) \\
&= N\sigma^2 \\
&- 2\sigma^2 \sum_{i\in[K]} E\left[\frac{tr(D^T D)}{D_i^2(\frac{\|y\|^2}{\sigma^2} - N) + tr(D^T D)} - 2\frac{tr(D^T D) D_i^2 (U_i^T y)^2/\sigma^2}{(D_i^2(\frac{\|y\|^2}{\sigma^2} - N) + tr(D^T D))^2} + (N-K) \right] \\
&+ \sum_{i\in[K]} E\left[\left(\frac{tr(D^T D)}{D_i^2(\frac{\|y\|^2}{\sigma^2} - N) + tr(D^T D)} \right)^2 (U_i^T y)^2 \right] + \sum_{i=K+1}^N E\left[(U_i^T y)^2 \right]
\end{aligned} \tag{4.10}$$

$$= K\sigma^2$$
$$- 2\sigma^2 E\left[\sum_{i\in[K]} \frac{tr(D^TD)}{D_i^2(\frac{\|y\|^2}{\sigma^2}-N)+tr(D^TD)} - 2\sum_{i\in[K]} \frac{tr(D^TD)D_i^2(U_i^Ty)^2/\sigma^2}{(D_i^2(\frac{\|y\|^2}{\sigma^2}-N)+tr(D^TD))^2}\right]$$
$$+ \sigma^2 E\left[\sum_{i\in[K]} \left(\frac{tr(D^TD)}{D_i^2(\frac{\|y\|^2}{\sigma^2}-N)+tr(D^TD)}\right)^2 - 4\sum_{i\in[K]} \frac{tr(D^TD)^2D_i^2(U_i^Ty)^2/\sigma^2}{(D_i^2(\frac{\|y\|^2}{\sigma^2}-N)+tr(D^TD))^3}\right]$$

$$\leq K\sigma^2$$
$$- \sigma^2 E\left[\sum_{i\in[K]} \frac{tr(D^TD)}{D_i^2(\frac{\|y\|^2}{\sigma^2}-N)+tr(D^TD)}\left\{1 - 4\frac{D_i^2(U_i^Ty)^2/\sigma^2\cdot D_i^2(\frac{\|y\|^2}{\sigma^2}-N)}{\left\{D_i^2(\frac{\|y\|^2}{\sigma^2}-N)+tr(D^TD)\right\}^2}\right\}\right]. \quad (4.11)$$

We use the fact that $\frac{tr(D^TD)}{D_i^2(\|y\|^2/\sigma^2-N)+tr(D^TD)} < 1$. If $4(U_i^Ty)^2 \leq \sum_{j\in[K]}(U_j^Ty)^2 - (N - tr(D^TD))\sigma^2/D_i^2$, $\forall i$ That is the condition (4.8) is satisfied, so (4.11) in the above inequality is negative. Therefore,

$$MSE(\hat{\gamma}_{JS}) \leq \sigma^2\left(K - \sum_{i\in[K]} \frac{tr(D^TD)}{D_i^2(E(\frac{\|y\|^2}{\sigma^2}-N)+tr(D^TD)}\right)$$
$$= \sigma^2\left(K - \sum_{i\in[K]} \frac{tr(D^TD)}{D_i^2\|Z\gamma^*\|^2/\sigma^2+tr(D^TD)}\right)$$

since $E[1/X] \geq 1/E[X]$. Note that MSE of $\hat{\gamma}_{JS}$ is smaller than $K\sigma^2$ which is MSE of MLE (since $MSE(\hat{\gamma}_{MLE}) = E[\|Z(Z^TZ)^{-1}Z^T\varepsilon\|^2]$). We believe $\hat{\gamma}_{JS}$ still achieve smaller MSE without the condition, but we don't show in more detail. ■

4.3 Error Bounds

In this chapter we show the oracle error bounds of l_0+l_2 regularization. First we consider a general case where penalty is $p(\gamma; W, \lambda)$ in (4.1). Define the augmented design matrix and the corresponding response vector, $\tilde{y}$, and noise $\tilde{\varepsilon}$ as follows.

$$\tilde{y} = \begin{bmatrix} y \\ 0 \end{bmatrix}, \quad \tilde{Z} = \begin{bmatrix} Z \\ W^{1/2} \end{bmatrix}, \quad \tilde{\varepsilon} = \begin{bmatrix} \varepsilon \\ -W^{1/2}\gamma^* \end{bmatrix}.$$

For the new linear model $\tilde{y} = \tilde{Z}\gamma + \tilde{\varepsilon}$, the l_0 regularization problem is

$$\hat{\gamma} = \arg\min_{\gamma\in\mathbb{R}^K} \frac{1}{2}\|\tilde{y} - \tilde{Z}\gamma\|_2^2 + \frac{\lambda^2}{2}\|\gamma\|_0. \quad (4.12)$$

which is equavalent to (4.1). Note that in this l_0 regularization $\tilde{\varepsilon}$ is not Gaussian noise anymore.

Theorem 4.2 *For any given $c > 1$ and $\alpha > 0$, λ is given by*

$$\lambda = \sigma\sqrt{c((\alpha+1)\log K + 2)}. \tag{4.13}$$

Then with the probability greater than $1 - \frac{K^{-\alpha}}{1-K^{-\alpha}}$,

$$\begin{aligned}&\|Z(\gamma^* - \hat{\gamma})\|_2^2 + \|W^{1/2}(\hat{\gamma} - \gamma)_{\mathcal{J}^{*c}}\|^2 \\ &\leq \frac{c+1}{c-1}\|Z(\gamma^* - \gamma)\|_2^2 + \frac{2c}{c-1}\|W^{1/2}(\gamma^* - \gamma)\|_2^2 + \frac{c^2}{(c-1)^2}\|W^{1/2}\gamma^*\|_2^2 + \frac{2c}{c-1}\lambda^2 J\end{aligned} \tag{4.14}$$

and

$$\|W^{1/2}(\hat{\gamma} - \gamma^*)\|_2^2 \leq \frac{4c^2}{(c-1)^2}\{\frac{\|W\gamma^*\|_2^2}{\phi} + 2(\lambda^2 J + \frac{c+1}{2c}\|\tilde{Z}(\gamma^* - \gamma)\|^2)\} \tag{4.15}$$

where $\phi := \inf\{\|\tilde{Z}\Delta\|^2/\|\Delta_{\mathcal{J}}\|^2 \colon \Delta \in \mathbb{R}^K\}$.

Corollary 4.1 *For the given λ as (4.13), with the probability greater than $1 - \frac{K^{-\alpha}}{1-K^{-\alpha}}$,*

$$\|Z(\gamma^* - \hat{\gamma})\|_2^2 + \|W^{1/2}_{\mathcal{J}^{*c}}(\hat{\gamma} - \gamma^*)_{\mathcal{J}^{*c}}\|^2 \leq \frac{c^2}{(c-1)^2}\|W^{1/2}\gamma^*\|_2^2 + \frac{2c}{c-1}\lambda^2 J^* \tag{4.16}$$

and

$$\|\gamma^* - \hat{\gamma}\|_2^2 \leq w_{min}^{-1}\left(\frac{2c}{c-1}\lambda^2 J^* + \frac{4c^2}{(c-1)^2}\frac{\|W\gamma^*\|_2}{\sqrt{\phi}}\sqrt{\frac{\|W\gamma^*\|_2^2}{\phi} + 2\lambda^2 J^*}\right) \tag{4.17}$$

where $w_{min} = \min_{i\in[K]} w_i$.

Remark 4.3

(1) Note that ϕ is greater than $(\min_{i\in\mathcal{J}^*} w_i)$ since $\|\tilde{Z}\Delta\|^2 = \|Z\Delta\|^2 + \|W^{1/2}\Delta\|^2 \geq \|W^{1/2}_{\mathcal{J}^*}\Delta_{\mathcal{J}^*}\|^2$. Therefore (4.17) is upper bounded by $O_p(\max\{w_{max}^2\|\gamma^*\|^2/w_{min}^2, \lambda^2 J^*/w_{min}\})$. Note that $\|\gamma^*\|^2 \leq \max\{\gamma_{min}^{*2}, \gamma_{max}^{*2}\}J^*$ and the optimal rate of $\max|\gamma^*|$ is known as $\sqrt{\log K}$ (since $\max|z|/\sqrt{2\log K} \to 1$ a.s. when $z \sim N_K(0,1)$). Therefore we can approximate $\|\gamma^*\|^2 \lesssim O(\lambda^2 J^*)$. Then (4.17) is at the rate of $O(\lambda^2 J^* \cdot \max\{w_{max}^2/w_{min}^2, 1/w_{min}\})$. Therefore, if the rate of w_is are the same the upper bound would be at the rate of $\lambda^2 J^*$. However, if we consider different rate of coefficient, the upper bound can be bounded at the smaller rate. For instance, let $w_{min} \approx O((\lambda^2 J^*)^q)$ and $w_{max} \approx O((\lambda^2 J^*)^{q/2})$ (since the magnitude can be larger even it has smaller order) then estimation has smaller order of $O((\lambda^2 J^*)^{1-q})$. Indeed

w_i works dependently with $\|\gamma^*\|$, we can classify γ^* and apply different weight for each block of γ^*.

(2) Unlike the estimation error, the prediction error (4.16) cannot lower bounded than $O(\lambda^2 J^*)$. To achieve the oracle rate, $\|W^{1/2}\gamma^*\|^2$ should be at most $O(\lambda^2 J^*)$.

(3) Consider $W = wI_K$. If $w \lesssim O(\lambda^2 J^*/\|\gamma^*\|^2)$ then the prediction error is upper bounded at the same optimal rate of $\lambda^2 J^*$. Otherwise, it has greater rate of error. The estimation error is at the rate of $w\|\gamma^*\|^2 + \frac{\lambda^2 J^*}{w}$ and if w is $O(\lambda^2 J^*/\|\gamma^*\|^2)$ then the bounds is $O(\|\gamma^*\|^2) \approx O(\lambda^2 J^*)$. That is, the rates of both prediction and estimation cannot be improved than $\lambda^2 J^*$ if weights for l_2 penalty are assigned as the same. Furthermore, if $w \gtrsim O(\lambda^2 J^*/\|\gamma^*\|^2)$ then prediction error has even greater rate.

Theorem 4.3 *For any given $c > 1$ and $\alpha > 0$, λ is chosen as (4.13). Then with the probability greater than $1 - \frac{K^{-\alpha}}{1-K^{-\alpha}}$, we have*

$$|\hat{\mathcal{J}} \backslash \mathcal{J}^*| \leq \frac{1}{\lambda^2} \cdot \frac{1+\varepsilon}{\varepsilon} \left[\varepsilon \frac{2c}{c-1} \lambda^2 J^* + 2 \frac{(\varepsilon+1)c-1}{c-1} \frac{\|W\gamma^*\|_2}{\sqrt{\phi}} \|\tilde{Z}(\gamma^* - \hat{\gamma})\|_2 \right] + \frac{2+\varepsilon}{\varepsilon} J^* \tag{4.18}$$

for any $\varepsilon > 0$.

Corollary 4.2 *Suppose the same setting as in Theorem 4.3. Then with the probability greater than $1 - \frac{K^{-\alpha}}{1-K^{-\alpha}}$,*

$$|\hat{\mathcal{J}} \backslash \mathcal{J}^*| \leq \frac{1}{\lambda^2} \cdot \frac{1+\varepsilon}{\varepsilon} \left[\varepsilon \frac{2c}{c-1} \lambda^2 J^* + 4 \frac{c((\varepsilon+1)c-1)}{(c-1)^2} \frac{\|W\gamma^*\|_2}{\sqrt{\phi}} \sqrt{\frac{\|W\gamma^*\|_2^2}{\phi} + 2\lambda^2 J^*} \right] + \frac{2+\varepsilon}{\varepsilon} J^* \tag{4.19}$$

for any $\varepsilon > 0$.

Remark 4.4

(1) Note that (4.19) is bounded at the rate of $O(\lambda^2 J^* + \|W\gamma^*\|^2/(\phi\lambda^2))$. As shown in Remark 4.3, $\|\gamma^*\|^2 \approx O(\lambda^2 J^*)$ and $\phi \geq w_{min}$, the rate of (4.19) is $O(\lambda^2 J^* + w_{max}^2 J^*/w_{min})$. To select with the same rate of true cardinality w_{max}^2/w_{min} should be at most $O(\lambda^2)$. As the example in Remark 4.3, if we select $w_{min} \approx O((\lambda^2 J^*)^q)$ and $w_{max} \approx O((\lambda^2 J^*)^{q/2})$ for $q < 1$, then the selected subset has the same order of J^*. Again, it shows that the weight should be assigned dependetly to each γ_i^*. Indeed the estimation error (4.17) is $O(\max\{\frac{\lambda^2 J^*}{w_{min}}, \frac{\|W\gamma^*\|^2}{w_{min}^2}\})$ and the selection error (4.19) is $O(\max\{J^*, \frac{\|W\gamma^*\|^2}{\lambda^2} w_{min}\})$. Therefore $\|W\gamma^*\|^2/w_{min}$ should be at most $O(\lambda^2 J^*)$. The optimal choice of w will be discussed later.

(2) Consider $W = wI_K$. Then the rate, $O(\lambda^2 J^* + wJ^*)$ is optimal when w is at most $O(\lambda^2)$. To achieve the optimal rate of prediction and estimation errors, $w \lesssim O(\lambda^2 J^*/\|\gamma^*\|^2) \approx O(1)$ so in this case $w \approx O(1)$. That is, when $p(\gamma; W, \lambda)$, if $\lambda = O(\sigma\sqrt{(\log K)})$ and $w_i = O(1)$ for all $i \in [K]$, then

$$\frac{\|Z(\gamma^* - \hat{\gamma})\|^2}{\lambda^2} + \frac{\|\gamma^* - \hat{\gamma}\|^2}{\lambda^2/N} + |\hat{\mathcal{J}} \backslash \mathcal{J}^*| = O(J^*) \quad w.h.p$$

Remark 4.5 *Suppose that there are L of blocks defined as*

$$B_l = \{i \in [K] : c(\log K)^{1/(2(l+1))} < |\gamma_i^*| \leq c(\log K)^{1/(2l)}\}, \quad \forall l = 1, \cdots, L$$

for an arbitrary constant $c > 0$. For the lth block, we assign the weight of l_2 penalty as $w_l = w \cdot (\sqrt{\log K})^{1-1/l}$ for a given constant $w = c_0 \cdot N$. Then

$$\|W\hat{\gamma}\|^2 \leq c^2 w^2 \sum_{l \in [L]} |B_l| \log K = (cw)^2 J^* \log K$$

and therefore $\|W\hat{\gamma}\|^2 / w_{min}^2 = O(J^* (\log K)^{1/L})$ since $w_{min} = w \cdot (\sqrt{\log K})^{1-1/L}$ for $K > 3$. Then

$$\|\gamma^* - \hat{\gamma}\|^2 = O(J^* \max\{(\log K)^{\frac{1}{2}+\frac{1}{2L}}/N, (\log K)^{1/L}\}) = O(J^* (\log K)^{\frac{1}{2}+\frac{1}{2L}}/N) \quad w.h.p$$

which is smaller rate than that of l_0 or Lasso, $O(J^* \log K/N)$. Thus, on this setup, we have

$$\frac{\|Z(\gamma^* - \hat{\gamma})\|^2}{\lambda^2} + \frac{\|\gamma^* - \hat{\gamma}\|^2}{\lambda^{1+1/L}/N} + |\hat{\mathcal{J}} \backslash \mathcal{J}^*| = O(J^*) \quad w.h.p$$

for $\lambda = O(\sqrt{\log K})$. Note that if $L = 1$, there is no improvement on the estimation bound. If the true vector is consist of few strong signals and many weak signals which contribute small amount to the estimation, then the block size L is large. In this case, the estimation bound is much improved. However, the blocking is based on the unknown parameter, in practice we need to make blocks based on their estimates. In each iteration, $\hat{\gamma}^{(j)}$ are ordered and assigned blocks, then different weights are defined to different blocks. We will try this with simulation later.

Remark 4.6 *Note that there is no condition required for obtaining the oracle bounds for the prediction, estimation and selection. Since the augmented design matrix $\tilde{Z}$ is positive definite so there is no need of restricted eigen value assumption. However, the numerater of the bounds contain ϕ which is the restricted smallest eigenvalue of $\tilde{Z}$. Since for a $u \in \mathbb{R}^K$, $\|(\tilde{Z} + W^{1/2})u\|_2^2 \geq \|(\tilde{Z} + W^{1/2})_{\mathcal{J}^*} u_{\mathcal{J}^*}\|_2^2 \geq \{d_{min}(\Sigma_{\mathcal{J}^*}) + \min_{i \in \mathcal{J}^*} w_i\} u_{\mathcal{J}^*}$, so if the*

correlation among the relevant covariates is low ($d_{min}(Z_{\mathcal{J}^*}^T Z_{\mathcal{J}^*})$ is large) then the estimator achieves the lower bound. In the previous section, the empirical choice of w is proportional to $\sum_{i\in\mathcal{J}^*} D_i^2$ which has the larger value when covariates corresponding the true support set are correlated the more. Therefore, the weight of l_2 penalty should be adjusted by the correlation level to achieve the smaller error bounds.

Definition 4.1 *(Null consistency [43]) Let $\eta \in (0, 1]$. We say that the regularization method (1.4) satisfies the η null consistency condition if the following equality holds:*

$$\min_{\gamma\in\mathbb{R}^K} \left\{ \frac{1}{2}\|\frac{\epsilon}{\eta} - Z\gamma\|_2^2 + p(\gamma;\tau) \right\} = \frac{1}{2}\|\frac{\varepsilon}{\eta}\|_2^2 \tag{4.20}$$

Lemma 4.1 *For a given $\alpha > 0$, suppose that λ is given by*

$$\lambda = \frac{\sigma}{\eta}\sqrt{2(\alpha + 1 + \log K)}. \tag{4.21}$$

Then, η-null consistency holds for both $l_0 + l_2$ regularization and l_0 regularization with high probability. That is, for the given λ, with the probability at least $1 - e^{-\alpha}$,

$$\min_{\gamma\in\mathbb{R}^K} \left\{ \frac{1}{2}\|\frac{\varepsilon}{\eta} - Z\gamma\|_2^2 + \frac{\lambda^2}{2}\|\gamma\|_0 \right\} = \frac{1}{2}\|\frac{\varepsilon}{\eta}\|_2^2$$

and

$$\min_{\gamma\in\mathbb{R}^K} \left\{ \frac{1}{2}\|\frac{\varepsilon}{\eta} - Z\gamma\|_2^2 + \frac{w}{2}\|\gamma\|_2^2 + \frac{\lambda^2}{2}\|\gamma\|_0 \right\} = \frac{1}{2}\|\frac{\varepsilon}{\eta}\|_2^2.$$

Remark 4.7 *As explained in [43], null consistency means for $\eta = 1$, if $\gamma^* = 0$ then we have the global solution $\hat{\gamma} = 0$ either. As a slightly stronger condition, for $\eta < 1$ the null condition means when ε increases proportionally by $1/\eta$, the global solution $\hat{\gamma} = 0$ is achieved when $\gamma^* = 0$. That means even when the noise increases proportionally true null set is detectable. This condition holds for both l_0 and $l_0 + l_2$ regularizations with high probability and the given λ as (4.21), which depends on the η. Note that λ and η are inversely related therefore if η is close to zero (noise level tends to infinity) then the large λ is chosen.*

This condition gives the easier way to derive the oracle bounds, since the condition shows $\varepsilon^T Z\gamma$ can be bounded by $[\|Z\gamma\|_2^2/2 + p(\gamma;\tau)] \cdot \eta$. Then for l_0 regularization, $\|Z(\hat{\gamma} - \gamma^*)\|^2 \lesssim \lambda^2 J^*$ which gives the optimal rate. However for $l_0 + l_2$ regularization, as shown in the previous work $\|Z(\hat{\gamma} - \gamma^*)\|^2 \lesssim \lambda^2 J^* + w/2\|\gamma^*\|^2$. The error bound can be sub-optimal since it depends on $w\|gams\|^2$ and $\lambda^2 J^*$. As pointed out before, since $\|gams\|^2 \leq J^* \max\{\gamma_{min}^*, \gamma_{max}^*\}^2$, the weight should be selected depending on the magnitude of signal in order to obtain the optimal rate of prediction error.

Proof. *Theorem 4.2.* The proof is similar to that of l_0 regularization except here we use augmented matrices. From the optimization problem (4.12), for any $\gamma \in \mathbb{R}^K$,

$$\frac{1}{2}\|\tilde{y} - \tilde{Z}\hat{\gamma}\|_2^2 + \frac{\lambda^2}{2}\hat{J} \leq \frac{1}{2}\|\tilde{y} - \tilde{Z}\gamma\|_2^2 + \frac{\lambda^2}{2}J$$

and thereby for any constant $c > 0$,

$$\begin{aligned}
&\frac{1}{2}\|\tilde{Z}(\gamma^* - \hat{\gamma})\|_2^2 \\
&\leq \frac{1}{2}\|\tilde{Z}(\gamma^* - \gamma)\|_2^2 + \frac{\lambda^2}{2}(J - \hat{J}) + \varepsilon^T P_{\tilde{\mathcal{J}}} Z(\hat{\gamma} - \gamma) - {\gamma^*_{\mathcal{J}^*}}^T W_{\mathcal{J}^*}(\hat{\gamma} - \gamma)_{\mathcal{J}^*} \\
&\leq \frac{1}{2}\|\tilde{Z}(\gamma^* - \gamma)\|_2^2 + \frac{\lambda^2}{2}(J - \hat{J}) + t\lambda\sqrt{|\mathcal{J} \cup \hat{\mathcal{J}}|}\|Z(\hat{\gamma} - \gamma)\|_2 + \|W^{1/2}\gamma^*\|_2 \|W^{1/2}_{\mathcal{J}^*}(\hat{\gamma} - \gamma)_{\mathcal{J}^*}\|_2 \\
&\leq \frac{1}{2}\|\tilde{Z}(\gamma^* - \gamma)\|_2^2 + \frac{\lambda^2}{2}(J - \hat{J}) + \frac{t}{2}\left(\sqrt{c}\lambda^2(J + \hat{J}) + \frac{1}{\sqrt{c}}\|Z(\hat{\gamma} - \gamma)\|_2^2\right) \\
&+ \|W^{1/2}\gamma^*\|_2 \|W^{1/2}_{\mathcal{J}^*}(\hat{\gamma} - \gamma)_{\mathcal{J}^*}\|_2
\end{aligned} \tag{4.22}$$

At the second inequality we assume the event $\mathcal{A}_{\{\lambda t\}} = \cap_{\mathcal{J} \subset [K] \cap 0}\{\|\varepsilon P_{\mathcal{J}}\|_2^2 \leq t^2\lambda^2 J\}$ holds, of which the probability is given in Lemma A.2. Let $t = 1/\sqrt{c}$, then on the event of $\mathcal{A}_{\{\lambda/\sqrt{c}\}}$,

$$\begin{aligned}
\frac{c-1}{2c}\|\tilde{Z}(\gamma^* - \hat{\gamma})\|_2^2 &= \frac{c-1}{2c}\|Z(\gamma^* - \hat{\gamma})\|_2^2 + \frac{c-1}{2c}\|W^{1/2}(\gamma^* - \hat{\gamma})\|_2^2 \\
&\leq \frac{c+1}{2c}\|\tilde{Z}(\gamma^* - \gamma)\|_2^2 + \lambda^2 J + \|W^{1/2}\gamma^*\|_2 \|W^{1/2}(\hat{\gamma} - \gamma)_{\mathcal{J}^*}\|_2,
\end{aligned} \tag{4.23}$$

and therefore

$$\begin{aligned}
&\frac{c-1}{2c}\|Z(\gamma^* - \hat{\gamma})\|_2^2 + \frac{c-1}{2c}\|W^{1/2}(\hat{\gamma} - \gamma)_{\mathcal{J}^{*c}}\|_2^2 \\
&\leq \frac{c+1}{2c}\|Z(\gamma^* - \gamma)\|_2^2 + \|W^{1/2}(\gamma^* - \gamma)\|_2^2 \\
&- \frac{c-1}{2c}\|W^{1/2}(\hat{\gamma} - \gamma)_{\mathcal{J}^*}\|_2^2 + \lambda^2 J + \|W^{1/2}\gamma^*\|_2 \|W^{1/2}(\hat{\gamma} - \gamma)_{\mathcal{J}^*}\|_2 \\
&\leq \frac{c+1}{2c}\|Z(\gamma^* - \gamma)\|_2^2 + \|W^{1/2}(\gamma^* - \gamma)\|_2^2 \\
&- \frac{c-1}{2c}\left(\|W^{1/2}(\hat{\gamma} - \gamma)_{\mathcal{J}^*}\|_2 - \frac{c}{c-1}\|W^{1/2}\gamma^*\|_2\right)^2 + \frac{c}{2(c-1)}\|W^{1/2}\gamma^*\|_2^2 + \lambda^2 J
\end{aligned} \tag{4.24}$$

Thus from (4.24)

$$\begin{aligned}
&\|Z(\gamma^* - \hat{\gamma})\|_2^2 + \|W^{1/2}(\hat{\gamma} - \gamma)_{\mathcal{J}^{*c}}\|^2 \\
&\leq \frac{c+1}{c-1}\|Z(\gamma^* - \gamma)\|_2^2 + \frac{2c}{c-1}\|W^{1/2}(\gamma^* - \gamma)\|_2^2 + \frac{c^2}{(c-1)^2}\|W^{1/2}\gamma^*\|_2^2 + \frac{2c}{c-1}\lambda^2 J
\end{aligned} \tag{4.25}$$

From (4.23) the estimation error in l_2 norm is simply derived. For $\phi := \inf\{\|\tilde{Z}\Delta\|^2/\|\Delta_{\mathcal{J}}\|^2: \Delta \in \mathbb{R}^K\}$,

$$\begin{aligned}(4.23) &= \frac{c+1}{2c}\|\tilde{Z}(\gamma^* - \gamma)\|_2^2 + \lambda^2 J + \|W\gamma^*\|_2\|(\hat{\gamma}-\gamma)_{\mathcal{J}^*}\|_2 \\ &\leq \frac{c+1}{2c}\|\tilde{Z}(\gamma^* - \gamma)\|_2^2 + \lambda^2 J + \frac{\|W\gamma^*\|_2}{\sqrt{\phi}}\|\tilde{Z}(\hat{\gamma}-\gamma)\|_2 \qquad (4.26)\end{aligned}$$

Therefore,

$$\begin{aligned}\|\tilde{Z}(\gamma^* - \hat{\gamma})\|_2 &\leq \frac{c}{c-1}\left[\|W\gamma^*\|_2/\sqrt{\phi} + \sqrt{\frac{\|W\gamma^*\|_2}{\sqrt{\phi}} + 4\frac{c-1}{c}(\lambda^2 J + \frac{c+1}{2c}\|\tilde{Z}(\gamma^* - \gamma)\|^2)}\right] \\ &\leq \frac{2c}{c-1}\sqrt{\frac{\|W\gamma^*\|_2^2}{\phi} + 2(\lambda^2 J + \frac{c+1}{2c}\|\tilde{Z}(\gamma^* - \gamma)\|^2)},\end{aligned}$$

from this the estimation error is bounded as (4.15). ■

Proof. *Theorem 4.3* From the last line of (4.22) we have

$$\begin{aligned}&\left(\frac{\sqrt{c}-t}{2\sqrt{c}}\right)\|\tilde{Z}(\gamma^* - \hat{\gamma})\|_2^2 \\ &\leq \frac{\lambda^2}{2}\left(1+t\sqrt{c}\right)|\mathcal{J}^*\backslash\hat{\mathcal{J}}| - \frac{\lambda^2}{2}\left(1-t\sqrt{c}\right)|\hat{\mathcal{J}}\backslash\mathcal{J}^*| + \frac{\lambda^2 t\sqrt{c}}{2}|\mathcal{J}^*\cap\hat{\mathcal{J}}| + \|W\gamma^*\|_2\|(\hat{\gamma}-\gamma^*)_{\mathcal{J}^*}\|_2\end{aligned}$$

where c and t are chosen such that $c < t^2$ and $t\sqrt{c} < 1$. We already assume the event $\mathcal{A}_{t\lambda}$ is held, so for $\lambda = \sigma\sqrt{\alpha + 2 + \log K}/t$, the event is held with the probability greater than $1 - \frac{e^{-\alpha}}{1-e^{-\alpha}}$. Let $\sqrt{c} = t/(1+\varepsilon)$ for $t < 1$ and $\varepsilon > 0$. Then

$$\begin{aligned}&|\hat{\mathcal{J}}\backslash\mathcal{J}^*| \\ &\leq \frac{2}{\lambda^2}\left(1-t\sqrt{c}\right)^{-1}\left[\left(\frac{t-\sqrt{c}}{2\sqrt{c}}\right)\|\tilde{Z}(\gamma^* - \hat{\gamma})\|_2^2 + \frac{\lambda^2}{2}\left(1+t\sqrt{c}\right)|\mathcal{J}^*\backslash\hat{\mathcal{J}}| + \frac{\lambda^2 t\sqrt{c}}{2}|\mathcal{J}^*\cap\hat{\mathcal{J}}|\right. \\ &\qquad\left. + \|W\gamma^*\|_2\|(\hat{\gamma}-\gamma^*)_{\mathcal{J}^*}\|_2\right] \\ &\leq \frac{1}{\lambda^2}\cdot\frac{1+\varepsilon}{1+\varepsilon-t^2}\left[\varepsilon\|\tilde{Z}(\gamma^* - \hat{\gamma})\|_2^2 + 2\|W\gamma^*\|_2\|(\hat{\gamma}-\gamma^*)_{\mathcal{J}^*}\|_2\right] + \frac{1+\varepsilon+t^2}{1+\varepsilon-t^2}J^* \\ &\leq \frac{1}{\lambda^2}\cdot\frac{1+\varepsilon}{1+\varepsilon-t^2}\left[\varepsilon\|\tilde{Z}(\gamma^* - \hat{\gamma})\|_2^2 + 2\frac{\|W\gamma^*\|_2}{\sqrt{\phi}}\|\tilde{Z}(\gamma^* - \hat{\gamma})\|_2\right] + \frac{1+\varepsilon+t^2}{1+\varepsilon-t^2}J^*\end{aligned}$$

Note that the condition $\mathcal{A}_{t\lambda}$ is equavalent to $\mathcal{A}_{\lambda/(\sqrt{c}(1+\varepsilon))}$. Therefore $\mathcal{A}_{\lambda/\sqrt{c}}$ is held as well therefore (4.26) can be applied. Since $t < 1$ and from (4.26), we have

$$\begin{aligned}&|\hat{\mathcal{J}}\backslash\mathcal{J}^*| \\ &\leq \frac{1}{\lambda^2}\cdot\frac{1+\varepsilon}{\varepsilon}\left[\varepsilon\frac{2c}{c-1}\lambda^2 J^* + 2\frac{(\varepsilon+1)c-1}{c-1}\frac{\|W\gamma^*\|_2}{\sqrt{\phi}}\|\tilde{Z}(\gamma^* - \hat{\gamma})\|_2\right] + \frac{2+\varepsilon}{\varepsilon}J^* \\ &\leq \frac{1}{\lambda^2}\cdot\frac{1+\varepsilon}{\varepsilon}\left[\varepsilon\frac{2c}{c-1}\lambda^2 J^* + 4\frac{c((\varepsilon+1)c-1)}{(c-1)^2}\frac{\|W\gamma^*\|_2}{\sqrt{\phi}}\sqrt{\frac{\|W\gamma^*\|^2}{\phi} + 2\lambda^2 J^*}\right] + \frac{2+\varepsilon}{\varepsilon}J^*\end{aligned}$$

for any $\varepsilon > 0$. We apply (4.26) then (4.18) and (4.19) are given. ■

Proof. *Lemma 4.1* Consider the simple case when all coefficients of l_2 penalty are the same as w. The null consistency is equivalent to

$$\begin{aligned} 1/\eta\varepsilon^T Z\gamma &\leq \|Z\gamma\|^2/2 + w\|\gamma\|^2/2 + \lambda^2\|\gamma\|_0/2 \\ &= \gamma_{\mathcal{J}}^T(Z_{\mathcal{J}}^T Z_{\mathcal{J}} + wI)\gamma_{\mathcal{J}}/2 + (\lambda\sqrt{J})^2/2. \end{aligned}$$

Since on the event $\mathcal{B}_{\mathcal{J}} = \{\|(Z_{\mathcal{J}}^T Z_{\mathcal{J}} + wI)^{-1/2} Z_{\mathcal{J}}^T \varepsilon\|_2 \leq \lambda\eta\sqrt{J}\}$,

$$\begin{aligned} &\gamma_{\mathcal{J}}^T(Z_{\mathcal{J}}^T Z_{\mathcal{J}} + wI)\gamma_{\mathcal{J}} - 2/\eta\varepsilon^T Z_{\mathcal{J}}\gamma_{\mathcal{J}} + (\lambda\sqrt{J})^2 \\ &\geq \|(Z_{\mathcal{J}}^T Z_{\mathcal{J}} + wI)^{1/2}\gamma_{\mathcal{J}}\|_2^2 - 2\|(Z_{\mathcal{J}}^T Z_{\mathcal{J}} + wI)^{-1/2} Z_{\mathcal{J}}^T\varepsilon/\eta\|_2 \cdot \|(Z_{\mathcal{J}}^T Z_{\mathcal{J}} + \eta I)^{1/2}\gamma_{\mathcal{J}}\|_2 + (\lambda\sqrt{J})^2 \\ &\geq (\|(Z^T Z + \eta I)^{1/2}\gamma\|_2 - \lambda\sqrt{J})^2 \\ &\geq 0. \end{aligned}$$

Therefore on the event $\mathcal{B}$ the null consistency holds. For UDV^T as the spectral decomposition of $Z_{\mathcal{J}}^T Z_{\mathcal{J}}$, $V(Z_{\mathcal{J}}^T Z_{\mathcal{J}} + wI)^{-1/2} Z_{\mathcal{J}}^T \varepsilon \sim N(0, \sigma^2(D^T D + wI)^{-1/2} D^T D (D^T D + wI)^{-1/2})$. Therefore from Lemma A.3 the probability of $\mathcal{B}^c$ is

$$\begin{aligned} P(\mathcal{B}_{\mathcal{J}}^c) &\leq \bigcup_{\mathcal{J}\subset[K]} \sum_{i\in\mathcal{J}} P\{\|V_i(Z_{\mathcal{J}}^T Z_{\mathcal{J}} + wI)^{-1/2} Z_{\mathcal{J}}^T\varepsilon\|_2^2 > \lambda^2\eta^2\} \\ &\leq \bigcup_{\mathcal{J}\subset[K]} \sum_{i\in\mathcal{J}} P\{\mathcal{W} \geq \frac{D_i^2 + w}{D_i^2}\frac{\lambda^2\eta^2}{\sigma^2}\} \\ &\leq \bigcup_{\mathcal{J}\subset[K]} \sum_{i\in\mathcal{J}} P\{\mathcal{W} \geq \frac{\lambda^2\eta^2}{\sigma^2}\} \\ &\leq \binom{K}{J} \sum_{i\in\mathcal{J}} exp[-\{\{\lambda\eta/\sigma - \sigma/(\lambda\eta)\}/2\}^2] \\ &\leq exp[-\{\{\lambda\eta/\sigma - \sigma/(\lambda\eta)\}/2\}^2 + \log K + \log\binom{K-1}{J-1}] \\ &\leq e^{-\alpha} \quad \text{for some } \alpha > 0, \end{aligned}$$

where $\mathcal{W} \sim \chi_1^2$. Therefore

$$(\lambda\eta/\sigma)^4 - 2(\log K + \alpha + 1)(\lambda\eta/\sigma)^2 + 1 \geq 0,$$

since $\binom{K-1}{J-1} \leq \left(\frac{K-1}{J-1}\right)^{(J-1)}$, for $J \geq 2$ or 1 for $J = 1$. Thus if λ is chosen as

$$\lambda \geq \frac{\sigma}{\eta}\sqrt{2(1 + \alpha + \log K)}$$

then the event $\mathcal{B}_{\mathcal{J}}$ holds for any $\mathcal{J} \subset [K]$ with the probability at least $1 - e^{-\alpha}$.
In the similar way, we can show the null consistency holds for l_0 regularizaiton. Since

$$\begin{aligned}
&\|Z_{\mathcal{J}}\gamma_{\mathcal{J}}\|_2^2 - 2\eta^T Z_{\mathcal{J}}\gamma_{\mathcal{J}}/\eta + \lambda^2\|\gamma\|_0 \\
&\geq (\|Z_{\mathcal{J}}\gamma_{\mathcal{J}}\|_2 - \lambda\sqrt{\|\gamma\|_0})^2 \\
&\geq 0
\end{aligned}$$

on the event $\mathcal{B}_{2,\mathcal{J}} := \{\|P_{\mathcal{J}}\eta\|_2 \leq \eta\lambda\sqrt{J}\}$. Note that $\mathcal{B}_{2,\mathcal{J}}$ implies $\mathcal{B}_{\mathcal{J}}$ and indeed during calculating $P[\mathcal{B}_{\mathcal{J}}^c]$ we upper bound it by $P(\mathcal{B}_{2,\mathcal{J}})$. Therefore with the same choice of λ and the probability l_0 regularization achieves the null consistency. ■

4.4 Convergence of Hellinger Distance

We study the convergence of the penalized estimate with Hellinger distance shown by [31]. It shows the similarity between two probability distribution defined by the Hellinger integral. To get the distance, it involves bracketing the restricted space upon the penalty functions. The Hellinger distance is less than the Kullback-Leiber distance which is used to define risk, so we expected the smaller order than that of risk.

Proposition 4.1 *Let* $\{Y_i\}_{i=1}^n \sim p_i(\theta_0, y)$ *and penalized criterion function is* $l(\theta, Y_i) - \lambda_n J(\theta)$ *where* $l(\theta, y)$ *and* $J(\theta)$ *are criterion function and penalty for* $\theta \in \Theta$ *with* Θ *as the parameter space.*

Suppose $f : \Theta \times \mathcal{Y}_i \to \mathbb{R}$ *with* $E_i f^2(\theta, Y_i) < +\infty$ *for all* $\theta \in \Theta$. *Let* $\mathcal{F} = \{f(\theta, \cdot) : \theta \in \Theta\}$ *and its u-bracketing is* $S(u, m) := \{f_1^l, f_1^u, \cdots, f_m^l, f_m^u\} \subset \mathcal{L}_2$ *satisfying* $\max_{i\in[m]}\|f_i^u - f_i^l\|_2 \leq u$. *If for any* $f \in \mathcal{F}$, *there exists a* j *such that* $f_j^l \leq f \leq f_j^u$, *a.e. in* P, *then Hellinger metric entropy with bracketing,* $H^B(v, \mathcal{F})$ *is equal to* $\log N(v, \mathcal{F}) = \log(\min\{m : S(v, m)\})$, *when* $f = p^{1/2}$ *for probability density function,* p.

Proposition 4.2 *Lemma 2.1 of [26] Suppose that for each* $\gamma \in \mathcal{C}_v$

$$\left(\mathbb{E}^* \sup_{u\in B_\delta(\gamma)} |p^{1/2}(y, \gamma) - p^{1/2}(y, u)|^2\right)^{1/2} \leq g(\delta)$$

for some strictly increasing continuous function $g : [0, \infty) \to [0, \infty)$. *(*$E^*$ *denotes upper expectation.) Then for* $u > 0$,

$$H^B(u, \mathcal{F}) \leq H(g^{-1}(u/2, \mathcal{F}))$$

Assumption 4.1 *([31]) There exist some constants $c_1, c_2 > 0$ such that for $\varepsilon > 0$,*

$$\sup_{\{k \geq 1\}} \psi(\varepsilon, k) \leq c_2 N^{1/2},$$

where $\psi(\varepsilon, k) = \int_L^{L^{1/2}} H^{1/2}(u, \mathcal{F}(k))du/L$ with $L = (c_1\varepsilon^2 + k - 1)$.

Lemma 4.2 *([31]) Suppose Assumption 4.1 holds. Then for the $\hat{\gamma}$ defined in (4.1) then*

$$P(h(\gamma^*, \hat{\gamma}) \geq \eta_N) \leq 7 \exp(-c_5 N \eta_N^2),$$

where $\eta_N = \max(\varepsilon_N, 1)$ with ε_N the smallest ε satisfying the Assumption 4.1.

Theorem 4.4 *For a global solution $\hat{\gamma}$ defined in (4.1) and for some constant $c_5 > 0$ and $c > 0$,*

$$P(h(\gamma^*, \hat{\gamma}) \geq \eta) \leq 7exp(-c_5 N \eta^2) \tag{4.27}$$

where $\eta = \min\left\{\sqrt{\frac{cK \log K}{wN}}, \sqrt{\frac{cK}{2N} \cdot \log\left(\frac{d_{max}}{N\sigma^2(\lambda^2 + w)}\right)}\right\}$

Remark 4.8 *The rate of convergence given in Theorem 4.4 is too large even greater than the optimal rate of estimation loss in l_2 norm. We will improve this rate by using more tight brackets. However, the rate is usable after selection. Let the cardinality of selected subset is s then $\eta = O(\min\{\sqrt{s \log s/N}, \sqrt{sd_{max}/N}\})$.*

Proof. For our case, $y_i \sim N(Z^i\gamma^*, \sigma^2)$ where Z^i is ith row of Z, and the Hellinger-distance is $h(\gamma^*, \gamma) = N^{-1}\sum_{i=1}^N h_i(\gamma^*, \gamma)$ where

$$\begin{aligned}
&h_i^2(\gamma^*, \gamma) \\
&= \int (p^{1/2}(y_i, \gamma) - p^{1/2}(y_i, \gamma^*))^2 dy_i \\
&= \frac{1}{2\pi\sigma^2} \int \left\{ exp\left(-\frac{(y_i - Z_i\gamma)^2}{2\sigma^2}\right) + exp\left(-\frac{(y_i - Z_i\gamma^*)^2}{2\sigma^2}\right) \right. \\
&\quad \left. -2exp\left(-\frac{(y_i - Z_i\gamma)^2 + (y_i - Z_i\gamma^*)^2}{4\sigma^2}\right) \right\} dy_i \\
&= 2 - \frac{2}{2\pi\sigma^2} \int exp\left(-\frac{(y_i - Z_i(\gamma + \gamma^*)/2)^2}{2\sigma^2}\right) \cdot exp\left(-\frac{(Z_i(\gamma - \gamma^*))^2}{8\sigma^2}\right) dy_i \\
&= 2 - 2exp\left(-\frac{(Z_i(\gamma - \gamma^*))^2}{8\sigma^2}\right).
\end{aligned}$$

For any $k \geq 1$, let $\mathcal{F} := \{p^{1/2}(\cdot, \gamma) : \gamma \in \mathbb{R}^K, \forall i \in [N]\}$ with $\mathcal{F}_{0+2}(k) := \{\gamma \in \mathbb{R}^K : w\|\gamma\|_2^2 + \lambda^2\|\gamma\|_0 \leq k\}$. We need to calculate $H^B(v, \mathcal{F}_{0+2})$ with bracketing of $\mathcal{F}_{0+2}$.

Notice that,

$$\mathcal{F}_{0+2}(k) \subset \mathcal{F}_2(k/(\lambda^2/l^2 + w)),$$

where $\mathcal{F}_2(u) = \{\gamma \in \mathcal{F}_\infty : \|\gamma\|_2^2 \leq u\}$ for any $u > 0$, and $\mathcal{F}_\infty := \{\gamma \in \mathbb{R}^K : \max_{i\in[K]}|\gamma_i| \leq l\}$ for $l \leq \sqrt{k/(\lambda^2 l^2 + w)}$.

Assume that $\|Z^i\|_2^2 < \infty$ for any $i \in [N]$. Let $S(v)$ is a v-bracketing of $\mathcal{F}_{0+2}$ as given by $S(v) := \{f^l(\gamma_v), f^u(\gamma_v) : \|f^l(\gamma_v) - f^u(\gamma_v)\|_2 \leq v\}$ and for any $p^{1/2}(\cdot, \gamma)$ there exists a pair of f^l, f^u such that $f^l(\gamma_v) \leq p^{1/2}(\cdot, \gamma) \leq f^u(\gamma_v)$. Note that f in $S(v)$ depends on γ_v and we define $\{B_\delta(\gamma_v) : f(\gamma_v) \in S(v)\}$ as a collection of δ-neighborhoods of members of $\mathcal{C}_v$, which is set of γ_vs.

$$B_\delta(\gamma) := \{\tilde{\gamma} \in \mathcal{F}_{0+2}(k) : \|\tilde{\gamma} - \gamma\|_2 \leq \delta\}.$$

Then we can apply the Proposition 4.2. Since for all $\gamma \in \mathbb{R}^K$,

$$\|Z(\gamma - \gamma^*)\|_2^2 = -8\sigma^2 \sum_{i=1}^N \log(1 - h_i^2(\gamma^*, \gamma)/2) \geq 4\sigma^2 \sum_{i=1}^N h_i^2(\gamma^*, \gamma)$$

$$\begin{aligned}
\mathbb{E}^*[\sup_{\tilde{\gamma}\in B_\delta(\gamma)} N^{-1} \sum_{i=1}^N (p^{1/2}(y_i, \gamma) - p^{1/2}(y_i, \tilde{\gamma}))^2] &= \sup_{\gamma\in B_\delta(s)} N^{-1} \sum_{i=1}^n \left\{ \int (p^{1/2}(y_i, \gamma) - p^{1/2}(y_i, \tilde{\gamma}))^2 dy_i \right\} \\
&\leq \frac{1}{4N\sigma^2} \sup_{\gamma\in B_\delta(s)} \|Z(\gamma - \tilde{\gamma})\|_2^2 \\
&\leq \frac{1}{4N\sigma^2} \sup_{\gamma\in B_\delta(s)} d_{max}^2 \|\gamma - \tilde{\gamma}\|_2^2 \leq \frac{d_{max}^2}{4N\sigma^2}\delta^2
\end{aligned}$$

where d_{max} is the largest eigen value of Z. Therefore, $\delta = u\sigma\sqrt{N}/d_{max}$ and

$$\begin{aligned}
H^B(u, \mathcal{F}_{0+2}(k)) &\leq H(\frac{u\sigma\sqrt{N}}{d_{max}}, \mathcal{F}_{0+2}(k)) \\
&\leq H(\frac{u\sigma\sqrt{N}}{d_{max}}, \mathcal{F}_2(\frac{k}{\lambda^2/l^2 + w})) \\
&\leq cK \log\left(\max\left(l \cdot \frac{d_{max}}{u\sigma\sqrt{N}}, 1\right)\right) \\
&\leq cK \log\left(\max\left(\sqrt{\frac{k}{\lambda^2/l^2 + w}} \cdot \frac{d_{max}}{u\sigma\sqrt{N}}, 1\right)\right)
\end{aligned}$$

for some constant $c > 0$.

Then from Assumption 4.1 and Lemma 4.2, the convergence rate can be calculated. For $L > 1$ Assumption 4.1 is satisfied for any ε since the left hand side is negative. We now consider $L \leq 1$,

$$\psi(\varepsilon, k) \leq \int_L^{L^{1/2}} \sqrt{cK \log(a\sqrt{k}/\sqrt{N})} du/L = \frac{1-\sqrt{L}}{\sqrt{L}} \sqrt{cK \log(a^2 k/N)/2} := \psi_1(\varepsilon, k)$$

where $a = d_{max}/(\sigma(\lambda^2/l^2 + \mu))$ and $L = c_1\varepsilon^2 + k - 1$. Since the upper bound $\psi_1(\varepsilon, k)$ is non-increasing with respect to k, therefore $\sup_{k\geq 1} \psi(\varepsilon, k) \leq \psi_1(\varepsilon, 1)$. If $1/\sqrt{L} \leq \frac{c_2\sqrt{N}}{\sqrt{cK \log(a/N)/2}}$ then the assumption is satisfied. That is,

$$\varepsilon \geq \sqrt{\frac{cK}{2c_1 c_2^2 N} \cdot \log\left(\frac{d_{max}}{N\sigma^2(\lambda^2 + \mu)}\right)}$$

From Lemma 4.2, we have the upper bound of Hellinger distance.

Note that $\mathcal{F}_{0+2}(t) \subset \{\gamma \in \mathcal{F}_\infty : \mu/K\|\gamma\|_1^2 + \lambda^2/l\|\gamma\|_1 \leq t\} \subset \{\gamma \in \mathcal{F}_\infty : \|\gamma\|_1 \leq \sqrt{tK/\mu}\}$. Therefore, from Lemma 3 in [27] we have

$$H^B(u, \mathcal{F}_{0+2}(t)) \leq \frac{ctK}{\mu}\left(\max\left(l\frac{d_{max}}{u\sigma\sqrt{N}}, 1\right)\right)^2 \log K.$$

for some constant $c > 0$. As the previous case, $\psi(\varepsilon, t) = \sqrt{\frac{ctK}{\mu}\log K}\left(\frac{\sqrt{L}-L}{L}\right)$ is nonincreasing function $\sup_{t\geq 1}\psi(\varepsilon, t) = \psi(\varepsilon, 1)$. Therefore, $\sqrt{L} \leq \sqrt{cK \log K/(c_2^2 \mu N)}$ satisfies Assumption 4.1. That is,

$$\varepsilon \geq \sqrt{\frac{cK \log K}{c_1 c_2^2 \mu N}}.$$

■

4.5 Selection

For the study of selection performance, we use the extended KKT condition as done for l_0 regularization. We study the probability of sign agreed selection achieved by a local solution which is a fixed point of the Hybrid estimator. The result will be compared to that of l_0 and Lasso.

Lemma 4.3 *Let $\lambda_2 = O(\sqrt{\log K})$. If w is assigned to satisfy $\min_{j\in\mathcal{J}^*}|\gamma_j^*| > \|w(\Sigma_{\mathcal{J}^*} + wI)^{-1}\gamma_{\mathcal{J}^*}^*\|_\infty$ and $w\|\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}(\Sigma_{\mathcal{J}^*} + wI)^{-1}\gamma_{\mathcal{J}^*}^*\|_\infty < \lambda_2$, then there exist a solution satisfying*

$$p[sign(\hat{\gamma}) = sign(\gamma^*)] =$$
$$p\left[\|(\Sigma_{\mathcal{J}^*} + wI)^{-1}Z_{\mathcal{J}^*}^T\varepsilon\|_\infty < \min_{j\in\mathcal{J}^*}|\gamma^*| - \|w(\Sigma_{\mathcal{J}^*} + wI)^{-1}\gamma_{\mathcal{J}^*}^*\|_\infty\right] \cdot$$
$$p\left[\|Z_{\mathcal{J}^{*c}}^T\varepsilon\|_\infty \leq \lambda_2 - \|w\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}(\Sigma_{\mathcal{J}^*} + wI)^{-1}\gamma_{\mathcal{J}^*}^*\|_\infty\right]$$

Remark 4.9

(1) If w is assigned to be comparably smaller than $\Sigma_{\mathcal{J}^*}$ or $\Sigma_{\mathcal{J}^*}$ is large (that is, correlation between relevant predictors is high), then $w(\Sigma_{\mathcal{J}^*} + wI)^{-1}$ is close to zero and the both of two preliminary conditions are satisfied. In this case, w is inversly related to $\Sigma_{\mathcal{J}^*}$ which is similar interpretation of the optimal choice of w given in the previous results.
(2) Also, in that case (w is relatively large), the probability of the first condition is even greater than that of L_0 regularization. Since diagonal entries of $(\Sigma_{\mathcal{J}^*} + wI)^{-1}\Sigma_{\mathcal{J}^*}(\Sigma_{\mathcal{J}^*} + wI)^{-1}$ is smaller than $\Sigma_{\mathcal{J}^*}^{-1}$, the required minimum strength can be lower than that of L_0. So, this can work better in lower signal strength when the predictors are highly correlated.

Proof. From the optimization problem (4.5) and the hybrid TISP,

$$Z^T Z\hat{\gamma} - Z^T y + w\hat{\gamma} + \lambda_2 \hat{S} = 0$$

where $\lambda_2 = \frac{\lambda}{k_0^2+w}$ and $\hat{s}_j \in [-1, 1]$ for $j \notin \hat{\mathcal{J}}$, or 0 otherwise. The sign agreed selection is achieved if $\hat{S}_{\mathcal{J}^{*c}} \in [-1, 1]$, $\hat{S}_{\mathcal{J}^*} = 0$ and $|\gamma_j^*| > |(\hat{\gamma} - \gamma^*)_j|$, $\forall j \in [K]$. From the hybrid TISP, a non-zero solution should greater than λ_2. If $\hat{\gamma}_{\mathcal{J}^{*c}} = 0$ and $\hat{S}_{\mathcal{J}^c} = 0$ then

$$\begin{aligned}
&(\hat{\gamma} - \gamma^*)_{\mathcal{J}^*} = (\Sigma_{\mathcal{J}^*} + wI)^{-1}(Z_{\mathcal{J}^*}^T \epsilon - w\gamma^*) \\
&\{\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}(\Sigma_{\mathcal{J}^*} + wI)^{-1} Z_{\mathcal{J}^*}^T - Z_{\mathcal{J}^{*c}}^T\}\epsilon - w\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}(\Sigma_{\mathcal{J}^*} + wI)^{-1}\gamma^* = -\lambda_2 \hat{S}_{\mathcal{J}^{*c}}.
\end{aligned}$$

Thus, if we have

$$\begin{aligned}
&\min_{j \in \mathcal{J}^*} |\gamma_j^*| > \|(\Sigma_{\mathcal{J}^*} + wI)^{-1}(Z_{\mathcal{J}^*}^T \epsilon - w\gamma^*)\|_\infty \\
&\|\{\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}(\Sigma_{\mathcal{J}^*} + wI)^{-1} Z_{\mathcal{J}^*}^T - Z_{\mathcal{J}^{*c}}^T\}\epsilon - w\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}(\Sigma_{\mathcal{J}^*} + wI)^{-1}\gamma^*\|_\infty \leq \lambda_2
\end{aligned}$$

then $\hat{\gamma}$ selects the true subset and achieves sign agreement as well. Therefore if we have

$$\|(\Sigma_{\mathcal{J}^*} + wI)^{-1} Z_{\mathcal{J}^*}^T \varepsilon\|_\infty < \min_{j \in \mathcal{J}^*} |\gamma^*| - \|w(\Sigma_{\mathcal{J}^*} + wI)^{-1}\gamma^*\|_\infty \tag{4.28}$$

$$\|Z_{\mathcal{J}^{*c}}^T (Z_{\mathcal{J}^*}(\Sigma_{\mathcal{J}^*} + wI)^{-1} Z_{\mathcal{J}^*}^T - I)\varepsilon\|_\infty \leq \lambda_2 - \|w\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}(\Sigma_{\mathcal{J}^*} + wI)^{-1}\gamma^*\|_\infty \tag{4.29}$$

then $sign(\hat{\gamma}) = sign(\gamma^*)$. Therefore the right hand sides of both (4.28) and (4.29) should be positive, and the probability of sign agreement is decided by two events (4.28) and (4.29). Note that all diagonal entries of $Z_{\mathcal{J}^{*c}}^T (Z_{\mathcal{J}^*}(\Sigma_{\mathcal{J}^*} + wI)^{-1} Z_{\mathcal{J}^*}^T - I)^2 Z_{\mathcal{J}^{*c}}$ are smaller than those of $Z_{\mathcal{J}^{*c}}^T Z_{\mathcal{J}^{*c}}$. Therefore, the probability of (4.29) is greater than $p[\|Z_{\mathcal{J}^{*c}}^T \varepsilon\|_\infty \leq \lambda_2 - \|w\Sigma_{\mathcal{J}^{*c},\mathcal{J}^*}(\Sigma_{\mathcal{J}^*} + wI)^{-1}\gamma^*\|_\infty\|]$ from Lemma A.3. ■

CHAPTER 5

THEORIES ON MULTIVARIATE RESPONSE MODEL

In this Section, we will extend the results on L_0-regularization and $L_0 + L_2$-regularization to the multivariate response model. We will study the oracle properties on the multi response model by extending those results from univariate case and after that we show the minimax rate for the prediction and the estimation loss with rank restriction.

$$Y = XA^* + E \tag{5.1}$$

where $Y \in \mathbb{R}^{n\times m}$, $X \in \mathbb{R}^{n\times p}$, $A^* \in \mathbb{R}^{p\times m}$, and $E \in \mathbb{R}^{n\times m}$ is zero mean random Gaussian noise. X is column normalized matrix such that $\|X_j\|_2 = \sqrt{n}$. Again, we let $\Sigma := X^T X$ where there is no abiguity to use the same notation as on the uni-model. The model (5.1) is equavalent to (1.3) by setting $y = \text{vec}$, $\gamma^* = \text{vec}(A^*)$, $\varepsilon = \text{vec}(E)$, and $Z = I_m \otimes X$. Since we consider grouping on (1.3) and the design matrix Z is block diagonal matrix, the results from the previous section would be modified a little. Let g as the number of groups, and G_i as index set of ith group. In our setup, the same group size is the same as, $g = m$ and each group is $G_i := \{p(i-1)+1, \cdots, pi\}$, and $A = [\gamma_{G_1}, \cdots, \gamma_{G_m}]$. The support set of A, A^*, and $\hat{A}$ are denoted by $\mathcal{M}$, $\mathcal{M}^*$, and $\hat{\mathcal{M}}$ respectively, where $\hat{A}$ is a solution to the following optimization problem.

$$\hat{A} = \underset{A\in\mathbb{R}^{p\times m}}{\arg\min}\, F(A), \quad \text{where } F(A) = \frac{1}{2k_0^2}\|Y - XA\|_F^2 + p(A; W, \lambda) \tag{5.2}$$

where $P(A;W,\lambda) = P(A;0,\lambda/k_0)$ for l_0 regularization and $p(A;1/k_0^2 W,\lambda/k_0^2)$ for $l_0 + l_2$ regularization, for

$$p(A;0,\lambda) = \frac{\lambda^2}{2k_0^2}\|A\|_{2,0} := \frac{\lambda^2}{2k_0^2}\sum_{j\in[p]} \mathbf{1}_{\{\|A^j\|\neq 0\}} = \frac{\lambda^2}{2k_0^2}|\mathcal{M}| \tag{5.3}$$

$$p(A;1/k_0^2 W,\lambda/k_0^2) = \frac{1}{2k_0^2}\|W^{1/2}A\|_F^2 + \frac{\lambda^2}{2k_0^2}\|A\|_{2,0} \tag{5.4}$$

where $\|\cdot\|_F$ is the Frobenius norm and A^j is jth row of A. Note that $\|\gamma\|_0 = m \cdot \|A\|_{2,0}$ and $\|\gamma\|_2^2 = \|A\|_F^2$, where $\gamma = \mathrm{vec}(A)$, since the group sizes are the same as m. Since the largest singular value, $\|X\|_2 = \|Z\|_2$, we have the same definition of k_0 as in the previous section such that $k_0 > \|X\|_2$. Then the corresponding Hard-TISP estimator is $\hat{A} = \vec{\Theta}_H\left(\left(I - \frac{\Sigma}{k_0^2}\right)\hat{A} + \frac{X^TY}{k_0^2}\,;\frac{\lambda}{k_0}\right)$. That is

$$\hat{A}^j = \begin{cases} \hat{A}^j + \frac{1}{k_0^2}X_j^T\left(Y - X\hat{A}\right), & \text{if } \left\|\hat{A}^j + \frac{1}{k_0^2}X_j^T\left(Y - X\hat{A}\right)\right\|_2 > \frac{\lambda}{k_0^2} \\ 0, & \text{otherwise} \end{cases}.$$

The corresponding Hard-Ridge TISP estimator to the penalty $p(\gamma; w/k_0^2 I, \lambda/k_0^2)$ is

$$\hat{A} = \vec{\Theta}\left(\left(I - \frac{\Sigma}{k_0^2}\right)\hat{A} + \frac{X^TY}{k_0^2}\,;\frac{\lambda}{k_0},\frac{1}{k_0^2}W\right).$$

and therefore

$$\hat{A}^j = \begin{cases} \frac{1}{1+w_j/k_0^2}\left\{\hat{A}^j + \frac{1}{k_0^2}X_j^T\left(Y - X\hat{A}\right)\right\}, & \text{if } \left\|\hat{A}^j + \frac{1}{k_0^2}X_j^T\left(Y - X\hat{A}\right)\right\|_2 > \frac{\lambda}{k_0^2} \\ 0, & \text{otherwise} \end{cases}.$$

Note that for Hard-Ridge TISP the selected subset is shrinken by w_j but for Hard TISP there is no shrinkage. Also for Hard-Ridge, λ works for selection and W works for shrinkage to enhance prediction accuracy, independently. As in the univariate model, each optimization problem is scaled down to satisfy the condition $\|\Sigma\|_2 < 1$ which guarantees a limit point of each procedure converges to a local solution to the original optimization problem.

5.1 Group L_0 Regularization

In this section we study the oracle inequalities on multi-response model for l_0 regularization.

Lemma 5.1 *Let $P = X(X^TX)^-X^T$, and $P_{\mathcal{M}} = X_{\mathcal{M}}(X_{\mathcal{M}}^TX_{\mathcal{M}})^-X_{\mathcal{M}}^T$, for any index set $\mathcal{M} \in [p]$. Define the event $\mathcal{B}_{\lambda_0}$ for given λ_0 such that*

$$\mathcal{B}_{\lambda_0} := \bigcap_{\{\mathcal{M}\subset[p]\}\cup\mathcal{M}=\emptyset} \left\{\|E^TP_{\mathcal{M}}\|_F^2 \leq \lambda_0^2|\mathcal{M}|\right\}.$$

where $E \in \mathbb{R}^{n\times m}$ is zero mean Gaussian noise with variance of σ^2. Then, for given $\alpha > 0$, if λ_0 is chosen as

$$\lambda_0 \geq \sigma\sqrt{(\alpha+1)\log p + 1 + m},$$

then the probability of $\mathcal{B}_{\lambda_0}$ is greater than $1 - \frac{p^{-\alpha}}{1-p^{-\alpha}}$.

Theorem 5.1 *Consider the model (5.1) and its global solution $\hat{A}$. Let $p > 1$. For any $c > 2$ and $\alpha > 0$, we choose λ as*

$$\lambda = \sigma\sqrt{c((\alpha+1)\log p + 1 + m)}. \tag{5.5}$$

Then with probability greater than $1 - \frac{p^{-\alpha}}{1-p^{-\alpha}}$ for any solution to (5.2) with the penalty (5.3) we have

$$\|X(\hat{A} - A^*)\|_F^2 \leq \inf_{A\in\mathbb{R}^{p\times m}} \left\{ \frac{c+2}{c-2}\|X(A-A^*)\|_F^2 + \frac{2c\lambda^2}{c-2}|\mathcal{M}| \right\}. \tag{5.6}$$

and

$$|\hat{\mathcal{M}}\backslash\mathcal{M}^*| \leq \inf_{A\in\mathbb{R}^{p\times m}} \left\{ \frac{\xi+1}{\lambda^2}\frac{c+2}{c-2}\|X(A-A^*)\|_F^2 + \frac{2(\xi+1)c}{c-2}|\mathcal{M}| \right\} + \frac{\xi+2}{\xi}|\mathcal{M}^*|. \text{ for any } \xi > 0. \tag{5.7}$$

Corollary 5.1 *Assume that all conditions in Theorem 5.1 are satisfied. Then with the same λ as defined in (5.5), we have*

$$\|X(\hat{A} - A^*)\|_F^2 \leq \frac{2\lambda^2 c}{c-2}|\mathcal{M}^*|, \tag{5.8}$$

and

$$|\hat{\mathcal{M}}\backslash\mathcal{M}^*| \leq \frac{c+1}{c-1}|\mathcal{M}^*|. \tag{5.9}$$

with the probability greater than $1 - \frac{p^{-\alpha}}{1-p^{-\alpha}}$.

Definition 5.1 *For $q \geq 1$, $\delta \geq 0$ and $\mathcal{M} \subset [p]$, define the restricted invertibility factors in multi-response version.*

$$RIFm_q(\delta, \mathcal{M}) = \inf\left\{ \frac{|\mathcal{M}|^{1/q}\|X^TXU\|_{max}}{\sqrt{n}\|U\|_q} : U \in \mathbb{R}^{p\times m}, \|U_{\mathcal{M}^c}\|_0 < \delta\|U_{\mathcal{M}}\|_0 \right\}, \tag{5.10}$$

where $\|V\|_{max} = \max_{i\in[n],j\in[m]}|V_{ij}|$, where $V \in \mathbb{R}^{n\times m}$.

Theorem 5.2 *Let $\hat{A}$ is a global solution to (5.2) and $p > 1$, $1 \leq q \leq \infty$. For given $c > 1$ and $\alpha > 0$, let $\lambda = 2\sigma\sqrt{c((\alpha+1)\log p + m + 1}$. If a positive constant $RIF_{m_q}(\delta, \mathcal{M}^*)$ exists for $\delta = (c+1)/(c-1)$, then we have the following inequality with the probability greater than $1 - p^{-\alpha}$.*

$$\|\hat{A} - A^*\|_q \leq \frac{|\mathcal{M}^*|^{1/q}(1 + 1/\sqrt{c})\lambda/\sqrt{n}}{RIFm_q(\delta, \mathcal{M}^*)}. \tag{5.11}$$

Corollary 5.2 *Assume that all conditions required in Theorem 5.2 are satisfied. Then with the same λ for a given $\alpha > 0$, and with the same probability we have*

$$\|\hat{A} - A^*\|_\infty \leq \frac{(1 + 1/\sqrt{c})\lambda/\sqrt{n}}{RIFm_\infty(\delta, \mathcal{M}^*)}, \tag{5.12}$$

Remark 5.1

(1) In Theorem 5.1, (5.6) and (5.7) hold for any $A \in \mathbb{R}^{p\times m}$ and that is why the inequalities are called the oracle inequalities. The prediction error is upper bounded by two terms, the prediction error of A and its cardinality, and A can be not sparse. Interestingly, the right hand sides can be smaller than the values when substituting $A = A^*$. That means a more sparse model A can give a smaller upper bound (Assume that the true model A^* is not sparse, then there can be a more sparse model which can gives a smaller value of upper bound.)

(2) Corollary 5.1 shows the rates of upper bound in terms of true model. On the multivariate model, the same optimal rates as Lasso are achieved as well.

$$\frac{\|X(\hat{A} - A^*)\|_F^2}{\log p + m} + \frac{\|\hat{A} - A^*\|_2^2}{(\log p + m)/n} + \hat{\mathcal{M}} = O(|\mathcal{M}^*|).$$

Note that the prediction error and the cardinarity of wrongly selected subset are controlled without any restriction (we did not make any assumption on the design matrix or the true signal.) Therefore, the optimal rate can be achieved even when the correlation between predictors is high or the true model is not sparse. For Lasso, the optimal rates are obtained only under a stringent restriction on Z and a sparsity restriction.

For the estimation error, it is bounded at the optimal rate as Lasso, but more relaxed condition is required. As we compared two conditions, RIF and RE of univariate versions, RIF is much more relaxed condition and its denumerator is bounded by $O(\lambda)$ so gives the optimal estimation error rate.

(3) In Theorem 5.1, the inequlities hold for any A. As we discussed in the univariate models, the rates of errors in Corollary 5.1 are true for approximately sparse A^* (there are a lot of predictors that contribute very small so the coefficients are close to zero). Consider

a certain row of A^* is close to zero and the row of A is exactly zero. Then in the right hand sides of the inequalities of Theorem 5.1, $X(\hat{A}-A^*)$ is very small and governed by $|\mathcal{M}|$ which is strong sparsity.
(4) λ is decided by p, n and σ only, and p is in log form and m is in plane form. Therefore for controlling errors, m, the number of response variables should not be too large either. The probabilities are close to one when the dimension is large
(5) On the asymptotic analysis, the estimation error bound can give the similar meaning as that in the univariate model. In the multivariate model the rate of l_2 norm is changed to $O(|mtms|(\log p + m)/n)$. When $|\mathcal{M}^*|$ is fixed but p, m, n increase, the estimation consistency can be achieved when $\log p + m \lesssim O(n)$. That means at least the group size (number of response variables) is not of much larger order than the dimensionality. On this case the relationship $p \lesssim O(exp(n))$ is meaningful. On the other hand, if m, p are fixed but $n \to \infty$ then it achieves the same rate of convergence as the restricted OLS (root-n-consistency).

Proof. *Lemma 5.1* We can prove this lemma easily by substituting Z in Lemma A.2 by $I \otimes X$. It is clear that $\|E^T P_{\mathcal{M}}\|_F^T \sim \sigma^2 \chi^2_{m \cdot q_{\mathcal{M}}}$, where $q_{\mathcal{M}} = \text{rank}(X_{\mathcal{M}}) \leq |\mathcal{M}| \wedge n$. Therefore, for fixed $\lambda_0 > 0$ and for $W_j \sim \chi^2_j$,

$$\begin{aligned}
P(\mathcal{B}^c_{\lambda_0}) &\leq \sum_{\mathcal{M} \subset [p]} P\left\{ \|E^T P_{\mathcal{M}}\|_F^2 > \frac{\lambda_0^2}{m} \cdot m|\mathcal{M}| \right\} + P\{\|E^T P_{\emptyset}\|_F^2 > 0\} \\
&\leq \sum_{i=1}^{p} \binom{p}{i} P\left[W_{i \cdot m} > (1 + \frac{\lambda_0^2}{m\sigma^2} - 1) i \cdot m \right] + 0 \\
&\leq \sum_{i=1}^{p} \binom{p}{i} \exp\left[-i \cdot m \frac{(\lambda_0^2/(m\sigma^2) - 1)^2}{4\lambda_0^2/(m\sigma^2)} \right] \\
&\leq \sum_{i=1}^{p} \exp\left[-i \left\{ m\rho^2 - \log \binom{p}{i} / i \right\} \right] \\
&\leq \sum_{i=1}^{p} \exp\left[-i \left\{ m\rho^2 - \log \left(\frac{pe}{i} \right) \right\} \right] \\
&\leq \frac{e^{-\alpha}}{1 - e^{-\alpha}} \quad \text{, for some } \alpha > 0
\end{aligned}$$

where $\rho = (\frac{\lambda_0}{\sqrt{m}\sigma} - \frac{\sigma\sqrt{m}}{\lambda_0})/2$. The last inequality implies that $\min_{i \subset [p]} [m(\frac{\lambda_0}{\sqrt{m}\sigma} - \frac{\sigma\sqrt{m}}{\lambda_0})^2/4 - \log \frac{pe}{i}] \geq \alpha$. Then,

$$\frac{\lambda_0}{\sqrt{m}\sigma} - \frac{\sigma\sqrt{m}}{\lambda_0} \geq \frac{2}{\sqrt{m}} (\alpha + \log p + 1)^{1/2},$$

and thereby

$$\lambda_0 \geq 2\sigma \frac{\sqrt{\alpha + \log p + 1} + \sqrt{\alpha + \log p + 1 + m}}{2}.$$

Therefore, for given $\alpha > 0$ if λ_0 is chosen by,

$$\lambda_0 \geq \sigma\sqrt{\alpha + \log p + m + 1},$$

then the probability of $\mathcal{B}_{\lambda_0}$ is greater than $1 - \frac{e^{-\alpha}}{1-e^{-\alpha}}$. ■

Proof. *Theorem 5.1* The proof of (5.6) is almost the same as the proof of Theorem (3.2). From (5.2)

$$\frac{1}{2k_0^2}\|X(\hat{A} - A^*)\|_F^2 \leq \frac{1}{2k_0^2}\|X(A - A^*)\|_F^2 + \frac{1}{k_0^2}\text{trace}(E^T X(\hat{A} - A)) + \frac{\lambda^2}{2k_0^2}(|\mathcal{M}| - |\hat{\mathcal{M}}|).$$

Let $\tilde{M} := \hat{\mathcal{M}} \cup \mathcal{M}$. We have

$$\begin{aligned}
\text{trace}(E^T X(\hat{A} - A)) &= \text{trace}(E^T P_{\tilde{M}} X_{\tilde{M}} (\hat{A} - A)_{\tilde{M}}) \\
&\leq \|E^T P_{\tilde{M}}\|_F \cdot \|X(\hat{A} - A)\|_F \\
&\leq \|E^T P_{\tilde{M}}\|_F (\|X(\hat{A} - A^*)\|_F^2 + \|X(A - A^*)_F^2)^{1/2} \\
&\leq \frac{c}{2}\|E^T P_{\tilde{M}}\|_F^2 + \frac{1}{c}\|X(\hat{A} - A^*)\|_F^2 + \frac{1}{c}\|X(A - A^*)\|_F^2. \qquad (5.13)
\end{aligned}$$

On the event of $\mathcal{B}_{\lambda/\sqrt{c}}$ as defined in Lemma 5.1,

$$\begin{aligned}
&\|X(\hat{A} - A^*)\|_F^2 \\
&\leq \frac{2c}{c-2}\left(\frac{c+2}{2c}\|X(A - A^*)\|_F^2 + \lambda^2|\mathcal{M}| + \frac{c}{2}\|E^T P_{\tilde{M}}\|_F^2 - \frac{\lambda^2}{2}(|\hat{\mathcal{M}}| + |\mathcal{M}|)\right) \\
&\leq \frac{c+2}{c-2}\|X(A - A^*)\|_F^2 + \frac{2c\lambda^2}{c-2}|\mathcal{M}| + \frac{c^2}{c-2}\max_{\mathcal{M}\subset[p]}\left\{\|E^T P_{\tilde{M}}\|_F^2 - \frac{\lambda^2}{c}|\tilde{M}|\right\} \\
&\leq \frac{c+2}{c-2}\|X(A - A^*)\|_F^2 + \frac{2c\lambda^2}{c-2}|\mathcal{M}|.
\end{aligned}$$

Now, we will show (5.7). For any $v > 0$, we have

$$\frac{1}{2k_0^2}\|Y - X\hat{A}\|_F^2 + \frac{\lambda^2}{2k_0^2}|\hat{\mathcal{M}}| \leq \frac{1}{2k_0^2}\|Y - XA^*\|_2^2 + \frac{\lambda^2}{2k_0^2}|\mathcal{M}^*| + v.$$

On the same event $\mathcal{B}_{\lambda/\sqrt{c}}$ as defined in Lemma 5.1, where $c > 1$ is a certain constant, $\|E^T P_{\mathcal{M}}\|_F^2 \leq \frac{\lambda^2}{c}|\mathcal{M}|$,

$$
\begin{aligned}
&\frac{1}{2}\|X(A^* - \hat{A})\|_F^2 \\
&\leq \mathrm{trace}(E^T X(A^* - \hat{A})) + \frac{\lambda^2}{2}(|\mathcal{M}^*| - |\hat{\mathcal{M}}|) + v \\
&\leq \|E^T P_{\tilde{\hat{M}}}\|_F \cdot \|X(A^* - \hat{A})\|_F + \frac{\lambda^2}{2}(|\mathcal{M}^*| - |\hat{\mathcal{M}}|) + v \\
&\leq \frac{t}{2\sqrt{c}}\|X(A^* - \hat{A})\|_F^2 + \frac{\sqrt{c}}{2t}\|E^T P_{\tilde{\hat{M}}}\|_F^2 + \frac{\lambda^2}{2}(|\mathcal{M}^*| - |\hat{\mathcal{M}}|) + v \\
&\leq \frac{t}{2\sqrt{c}}\|X(A^* - \hat{A})\|_F^2 + \frac{\lambda^2}{2}\left(\frac{1}{t\sqrt{c}}|\tilde{\hat{M}}| + |\mathcal{M}^*| - |\hat{\mathcal{M}}|\right) + v. \\
&= \frac{t}{2\sqrt{c}}\|X(A^* - \hat{A})\|_F^2 \\
&\quad + \frac{\lambda^2}{2}\left\{\left(\frac{1}{t\sqrt{c}} + 1\right)|\mathcal{M}^*| + \left(\frac{1}{t\sqrt{c}} - 1\right)|\hat{\mathcal{M}} \backslash \mathcal{M}^*| - |\mathcal{M}^* \cap \hat{\mathcal{M}}|\right\} + v, \qquad (5.14)
\end{aligned}
$$

where $\tilde{\hat{M}} := \mathcal{M}^* \cup \hat{\mathcal{M}}$. Substituting $t = (1+\xi)\sqrt{c}$ for an arbitrary $\xi > 0$ and $v = \lambda^2/2 \cdot |\mathcal{M}^* \cap \hat{\mathcal{J}}|$ then we have (5.7) after simple calcultation.

It is easy to show (5.9). Since t and v are arbitrary positive constants, if we substitute $t = \sqrt{c}$ and $v = \lambda^2/2 \cdot |\mathcal{M}^* \cap \hat{\mathcal{M}}|$ in (5.14), then we have (5.9). ■

Proof. *Theorem 5.2*

For any arbitarary constant $t > 0$ and $t \in [p], k \in [m]$, let $\hat{A}'_{jk} = t + \hat{A}_{jk}$, and $\hat{A}'_{i,i'} = \hat{A}_{i,i'}$ for all $i \neq j$, $i' \in [m]$, then

$$
\begin{aligned}
&\|Y - X\hat{A}\|_F^2/2 + \lambda^2/2\|\hat{A}\|_{2,0} \\
&\leq \|Y - X\hat{A}'\|_F^2/2 + \lambda/2\|\hat{A}'\|_{2,0} \\
&\leq \|Y - X\hat{A}\|_F^2/2 + t^2\|X_j\|_2^2/2 - tX_j^T(Y - X\hat{A})_k + \lambda^2/2\|\hat{A}^{-j}\|_{2,0} \\
&\quad + \mathbf{I}_{\{\|\hat{A}^j\|_2^2 + mt^2 + 2t\hat{A}_{jk} \neq 0\}},
\end{aligned}
$$

therefore

$$
\begin{aligned}
X_j^T(Y - X\hat{A})_k &\leq t\|X_j\|_2^2/2 + \lambda^2/(2t)(\mathbf{I}_{\{\|\hat{A}^j\|_2^2 + mt^2 + 2t\sum_{k\in[m]}\hat{A}_{jk} \neq 0\}} - \mathbf{I}_{\{\|\hat{A}^j\|_2 \neq 0\}}) \\
&\leq tn/2 + \lambda^2/(2t).
\end{aligned}
$$

Also, for a certain constant $c > 0$, and on the event

$\mathcal{B}_{1,\lambda/\sqrt{c}} := \bigcap_{k\in[p]}\bigcap_{j\in[m]}\{(X_j^TE_k)^2/\|X_j\|_2^2 \leq \lambda^2/c\}$,

$$\begin{aligned}\sqrt{c}X_j^TE_k = \sqrt{c}X_j^T\cdot\frac{X_jX_k^TE_k}{\|X_j\|_2^2} &\leq \|X_k^T\|_2\cdot\left\|\sqrt{c}\frac{X_k}{\|X_k\|_2^2}\cdot\right\|_2(X_k^TE_j)\\ &\leq t\|X_k\|_2^2/2 + c(X_k^TE_j)^2/(\|X_k\|_2^2 2t)\\ &\leq tn/2 + \lambda^2/(2t),\end{aligned}$$

Note that $(X_j^TE_k)^2/\|X_j\|_2^2 \sim \sigma^2\cdot\chi_1^2$. Let $W\sim\chi_1^2$ and then

$$\begin{aligned}P(\mathcal{B}^c_{1,\lambda/\sqrt{c}}) &\leq mp\cdot P\{\|W > \lambda^2/(c\sigma^2)\}\\ &\leq \exp\left[-\frac{(\lambda^2/(c\sigma^2)-1)^2}{4\lambda^2/(c\sigma^2)} + \log(mp)\right]\\ &\leq e^{-\alpha},\end{aligned}$$

for some $\alpha > 0$. If $\lambda \geq 2\sigma\sqrt{c(\alpha+1+\log(mp))}$, then the event $\mathcal{B}_{1,\lambda/\sqrt{c}}$ holds with the probability greater than $1-e^{-\alpha}$.

Since t is arbitrary constant and $\min_{t>0}\{tn/2+\lambda^2/(2t)\} = \lambda\sqrt{n}$,

$$\begin{aligned}\|X^TX(A^*-\hat{A})\|_{\max} &= \|X^T(Y-E-X\hat{A})\|_{\max}\\ &\leq \|X^T(Y-X\hat{A})\|_{\max} + \|X^TE\|_{\max}\\ &\leq (1+1/\sqrt{c})\lambda\sqrt{n}.\end{aligned}$$

From (5.9) and the definition of $RIFm_q(\delta,\mathcal{M}^*)$ with $\delta = (c+1)/(c-1)$ we have

$$\|\hat{A}-A^*\|_{\max} \leq \frac{|\mathcal{M}^*|^{1/q}\|X^TX(\hat{A}-A^*)\|_{\max}}{RIFm_q(\delta,\mathcal{M}^*)} \leq \frac{|\mathcal{M}^*|^{1/q}(1+1/\sqrt{c})/\sqrt{n}\lambda}{RIFm_q(\delta,\mathcal{M}^*)}$$

■

5.2 Group $L_0 + L_2$ Regularization

We study the oracle bounds for l_0+l_2 regularization as defined (5.2) with penalty First define the augmented matrix such that

$$\tilde{Y} = \begin{pmatrix} Y \\ 0 \end{pmatrix},\quad \tilde{X} = \begin{pmatrix} X \\ W^{1/2} \end{pmatrix},\quad \tilde{E} = \begin{pmatrix} E \\ -W^{1/2}A^* \end{pmatrix}$$

then (5.1) is equivalent to $\tilde{Y} = \tilde{X}A^* + \tilde{E}$ and the original optimization problem is described as l_0 optimization problem such that

$$\hat{A} = \arg\min_A \frac{1}{2}\|\tilde{Y}-\tilde{X}A\|_F^2 + \frac{\lambda^2}{2}\|A\|_{2,0} \tag{5.15}$$

We achieve the similar oracle inequalities as the univariate model, but here we put a special case where $A = A^*$ to compare the rates of error bounds only.

Theorem 5.3 *For any given $c > 1$ and $\alpha > 0$, λ is chosen as*

$$\lambda = \sigma\sqrt{c((1+\alpha)\log p + m + 1)}. \tag{5.16}$$

Then for any $\xi > 0$, with the probability at least $1 - \frac{p^{-\alpha}}{1-p^{-\alpha}}$, a global solution $\hat{A}$ to (5.2) with the penalty (5.4) satisfies,

$$\|X(A^* - \hat{A})\|_F^2 + \|W^{1/2}(\hat{A} - A^*)_{\mathcal{M}^{*c}}\|_F^2 \leq \frac{c^2}{(c-1)^2}\|W^{1/2}A^*\|_F^2 + \frac{2c}{c-1}\lambda^2|\mathcal{M}^*| \tag{5.17}$$

$$\|A^* - \hat{A}\|_F^2 \leq \frac{2c}{c-1}\frac{\lambda^2}{w_{min}}|\mathcal{M}^*| + \frac{4c^2}{(c-1)^2}\frac{\|WA^*\|_F}{(w_{min})^{3/2}}\sqrt{\frac{\|WA^*\|_F^2}{w_{min}} + 2\lambda^2|\mathcal{M}^*|} \tag{5.18}$$

$$|\hat{\mathcal{M}}\backslash\mathcal{M}^*| \leq \left((1+\xi)\frac{2c}{c-1} + \frac{2+\xi}{\xi}\right)|\mathcal{M}^*| + 4\frac{1+\xi}{\xi}\frac{(\xi+1)c^2}{(c-1)^2}\frac{\|WA^*\|_F^2}{\lambda^2\sqrt{w_{min}}}, \tag{5.19}$$

where $w_{min} = \min_{i\in[p]}\{w_i; w_i \neq 0\}$

The proofs are almost the same as the corresponding proof for the univariate model. The difference in the proof between multivariate and univariate models are shown in the previous section, so we skip the redundant work here. As on the univariate response model, there is no required condtion for obtaining the oracle bounds on multivariate regression model as well.

Remark 5.2

(1) The rates are given as

$$\begin{aligned}
&\|X(A^* - \hat{A})\|_F^2 \lesssim O_p(\|W^{1/2}A^*\|_F^2 + (\log p + m)|\mathcal{M}^*|)\\
&\|A^* - \hat{A}\|_F^2 \lesssim O_p(\|WA^*\|_F^2 w_{min}^{-2} + (\log p + m)|\mathcal{M}^*|w_{min}^{-1})\\
&|\hat{\mathcal{M}}\backslash\mathcal{M}^*| \lesssim O_p(\|WA^*\|_F^2(w_{min}(\log p + m))^{-1} + |\mathcal{M}^*|).
\end{aligned}$$

As we discussed in univariate models, there are additional term $\|W^{1/2}A^*\|_F^2$ on the right hand sides of upper bounds. Since W works associately with the true model A^*, the weight should be assigned related to A^*.

(2) The second term of left side of (5.17) is the estimation error on the wrongly selected subset. It can bounded by the same bound of the prediction error, therefore we can say that the estimation error on the wrongly selected subset can be small if the prediction is accurate.

Remark 5.3 *Choice of W*

(1) If we assume $A^*_{ij} \sim N_{pm}$ which is general prior given to A^*, then $\|A^{*j}\|_2^2 \sim \chi^2_m$ and $\|As\|_F^2 \lesssim \max\{\|\mathcal{A}^j\|_2\}|\mathcal{M}^*|$. If we assign constant level of W, then the same optimal rates as L_0 can be obtained.

(2) If $\|WA^*\|_F^2 w_{min}^{-1} \lesssim O(\lambda^2|\mathcal{M}^*|)$ then

$$\frac{\|X(A^* - \hat{A})\|_F^2}{\lambda^2} + \frac{\|A^* - \hat{A}\|_F^2}{\lambda^2 w_{min}^{-1}} + |\hat{\mathcal{M}}| = O(|\mathcal{M}^*|)$$

therefore the estimation error can have lower rate of upper bound. It depends on the smart choice of W, so we want to see the true potential of $L_0 + L_2$ by ideal choice of W as follow (Proofs are almost the same as in the univariate models so we briefly show the results only).

(i) If Σ, A^* are known, one of optimal choice of W is $w_j = O(\sigma^2/\|V_j^T A^*\|_2^2)$ for any $j \in [p]$ which minimize $E[\|X(\hat{A}_{R,W} - A^*)\|_F^2]$ where $\hat{A}_{R,W} = (X^TX + W)^{-1}X^TY$ and UDV^T is the singular value decomposition of X.

(ii) By applying block thresholding we can get a better estimation accuracy. For all $l = 1, ..., L$, define the lth block such that

$$B_l = \{i \in [p] : c(\log p)^{1/(2(l+1))} < |A_j^*| \leq c(\log p)^{1/(2l)}\},$$

and assign the weight to lth block as $w_l = w(\sqrt{\log p})^{1-1/l}$ for a positive constant $w = c_0 \cdot n$ then

$$\|A^* - \hat{A}\|_F^2 \lesssim O(|\mathcal{M}^*|\lambda^{1+1/L}/n)$$

which is smaller order than $O(|\mathcal{M}^*|\lambda^2/n)$ of L_0 regularization.

(iii) Assume that only σ is known. Let

$$\hat{A}_{JS} = (X^TX + w_{JS}I)^{-1}X^TY, \quad \text{where } w_{JS} = \left(\frac{tr(D^TD)}{\|UY\|_F^2/\sigma^2 - p}\right)_+$$

where UDV^T is the singular value decomposition. If $\sum_{i\in[p]}\sum_{j\neq i} D_i^2\|U_j^TY\|_2^2 \geq p(n - \sum_{i\in[p]} D_i^2)\sigma^2$ then

$$\begin{aligned} MSE(\hat{A}_{JS} &\leq \sigma^2\left(p - \sum_{i\in[p]} \frac{tr(D^TD)}{D_i^2\|XA^*\|_F^2/\sigma^2 + tr(D^TD)}\right) \\ &\leq MSE(\hat{A}^0) \end{aligned}$$

where $\hat{A}^0$ is restricted OLS. This result is an analogue of James-Stein (1961). It does not improve the error rate but it achieves smaller risk.

Note that all of the optimal weights of w can be interpreted as noise to signal ratio.

5.3 Selection

The selection problem is similarly solve as on the univariate model, but the difference arises in the different norm is used on multivariate setup. We study the selection performance both for L_0 and $L_0 + L_2$ regularization, but skip studying the specific probability of sign agreed selection, because they are almost the same. In this chapter, we focus on the differences from those of univariate case.

Lemma 5.2 *Let $\hat{A}$ is a local solution to (5.2) with L_0 penalty function (5.3). Let $\lambda = O(\sqrt{\log p + m})$. For a local solution $\hat{A}$, $P(sign(\hat{A}) = sign(A^*)) = P(\mathcal{B}'_1) \cdot P(\mathcal{B}'_2)$ where*

$$\mathcal{B}'_1 : \|\Sigma^{-1}_{\mathcal{M}^*} X^T_{\mathcal{M}^*} E\|_{max} \leq \min_{i \in \mathcal{M}^*, j \in [m]} |A^*_{ij}| \tag{5.20}$$

$$\mathcal{B}'_2 : \|X^T_{\mathcal{M}^{*c}} (I - X_{\mathcal{M}^*} \Sigma^{-1}_{\mathcal{M}^*} X^T_{\mathcal{M}^*}) E\|_{2,\infty} \leq \lambda k_0. \tag{5.21}$$

Remark 5.4 *Since diagonal entries of $X^T_{\mathcal{M}^{*c}}(I - X_{\mathcal{M}^*}\Sigma^{-1}_{\mathcal{M}^*}X_{\mathcal{M}^*})^2 X_{\mathcal{M}^{*c}}$ is less than those of $\Sigma_{\mathcal{M}^{*c}}$, the probability of (5.21) is less than $P(\|X^T_{\mathcal{M}^{*c}} E\|_2 \leq \lambda k_0)$ (see Lemma A.3). Therefore, this condition can be satisfied with large λ, without a restriction on the submatrix $\Sigma_{\mathcal{M}^*, \mathcal{M}^{*c}}$. Also, the first condition (5.20) does not involve the term either. From these fact, we can say that there is no stringent restriction on incoherence condition which controls the correlation between the relevant subset $X_{\mathcal{M}^*}$ and the irrelevant subset $X_{\mathcal{M}^{*c}}$. Roughly, if the overall correlation between covariates is low, then a local solution to L_0 regularization has the same sign as A^*.*

Remark 5.5 *Comparison with Lasso*
For Lasso, the solution $\hat{A}$ satisfies

$$\hat{A}^j = (\hat{A}^j + \lambda_2 \tilde{S}^j) \cdot \left(\frac{\|\hat{A}^j + \lambda_2 \tilde{S}^j\|_2 - \lambda_2}{\|\hat{A}^j + \lambda_2 \tilde{S}^j\|_2} \right)_+$$

where $\lambda_2 \tilde{S}^j = -\Sigma_{j,*}\hat{A} + X^T_j Y$. Then $\tilde{S}^j = \frac{\hat{A}^j}{\|\hat{A}^j + \lambda_2 \tilde{S}^j\|_2 - \lambda_2}$ if $j \in \hat{\mathcal{M}}$, or $\|\tilde{S}^j\|_2 \leq 1$ otherwise. Thus for $j \in \hat{\mathcal{M}}$, $\|\tilde{S}^j\|_2 = 1$ and on univariate model $\tilde{S}^j = sign(\hat{A})_j$. If $\mathcal{M}^* = \hat{\mathcal{M}}$ then we have

$$(\hat{A} - A^*)_{\mathcal{M}^*} = \Sigma^{-1}_{\mathcal{M}^*}(X^T_{\mathcal{M}^*} E - \lambda_2 \tilde{S}_{\mathcal{M}^*})$$
$$\|-\Sigma_{\mathcal{M}^{*c}, \mathcal{M}^*} \Sigma^{-1}_{\mathcal{M}^*}(X^T_{\mathcal{M}^*} E - \lambda_2 \tilde{S}_{\mathcal{M}^*}) + X^T_{\mathcal{M}^{*c}} E\|_2 \leq \lambda_2$$

and additionally, to achieve the sign agreement, $\|\hat{A} - A^*\|_{max} < \min |A^*_{ij}|$. Therefore if

$$\|\Sigma^{-1}_{\mathcal{M}^*} X^T_{\mathcal{M}^*} E\|_{max} < \min_{i \in [p], j \in [m]} |A^*_{ij}| - \lambda_2 \|\Sigma^{-1}_{\mathcal{M}^*} \tilde{S}_{\mathcal{M}^*}\|_{max} \tag{5.22}$$

$$\|X^T_{\mathcal{M}^{*c}}(I - X_{\mathcal{M}^*}\Sigma^{-1}_{\mathcal{M}^*} X^T_{\mathcal{M}^*}) E\|_2 \leq \lambda_2 - \lambda_2 \|\Sigma_{\mathcal{M}^{*c}, \mathcal{M}^*} \Sigma^{-1}_{\mathcal{M}^*} \tilde{S}_{\mathcal{M}^*}\|_2 \tag{5.23}$$

where $\tilde{S}^j$ *satisfies the condition* $\tilde{S}^j = \hat{A}^j/(\|\hat{A}^j + \lambda_2\tilde{S}^j\|_2 - \lambda_2)$ *for all* $j \in \mathcal{M}^*$. *Therefore,*

$$\min_{i\in[p],j\in[m]} |A^*_{ij}| > \lambda_2 \|\Sigma^{-1}_{\mathcal{M}^*}\tilde{S}_{\mathcal{M}^*}\|_{max}$$
$$\|\Sigma_{\mathcal{M}^{*c},\mathcal{M}^*}\Sigma^{-1}_{\mathcal{M}^*}\tilde{S}_{\mathcal{M}^*}\|_2 < 1$$

should be satisfied preliminarly and the probability of sign agreement, $p(sign(\hat{A}) = sign(A^*))$, *is a product of the probabilities of* (5.22) *and* (5.23). *Note that on an univariate model,* $\tilde{S}_{\mathcal{M}^*} = sign(\hat{\gamma}_{\mathcal{M}^*})$, *but on a multivariate model, it has complicated form. At least we can see,* (5.23) *involves a strict incoherence condition (restriction on* $\Sigma_{\mathcal{M}^{*c},\mathcal{M}^*}$ *since the right hand side should be positive.*

Comparing this result with L_0 *regularization, Lasso requires restriction on the correlation between the relevant subset and irrelevant subset of predictors. This incoherence condition is not the same as the restriction on the overall correlation struncture. Even when the overall correlation is low, the incoherence level can be not low enough.*

Lemma 5.3 *Let* $\lambda = O(\sqrt{\log p + m})$. *For a local solution* $\hat{A}$ *to* (5.2) *with the penalty* (5.4), *if the following conditions satisfied,*

$$\|(\Sigma_{\mathcal{M}^*} + W_{\mathcal{M}^*})^{-1}X^T_{\mathcal{M}^*}E\|_{max} < \min_{i\in\mathcal{M}^*,j\in[m]} |A^*_{ij}| - \|(\Sigma_{\mathcal{M}^*} + W_{\mathcal{M}^*})^{-1}W_{\mathcal{M}^*}A^*_{\mathcal{M}^*}\|_{max}$$
$$\|X^T_{\mathcal{M}^{*c}}(I - X_{\mathcal{M}^*}\Sigma^{-1}_{\mathcal{M}^*}X_{\mathcal{M}^*})E\|_{2,\infty} \leq \lambda k_0 - \|\Sigma_{\mathcal{M}^{*c},\mathcal{M}^*}(\Sigma_{\mathcal{M}^*} + W_{\mathcal{M}^*})^{-1}W_{\mathcal{M}^*}A^*_{\mathcal{M}^*}\|_{2,\infty}$$

then $sign(\hat{A}) = sign(A^*)$).

Remark 5.6

(1) W, *the weight of* l_2 *penalty takes the important role in selection either. Roughly speaking, if the weight on the true support set,* $W_{\mathcal{M}^*}$ *is assigned as very small compared to* $\Sigma_{\mathcal{M}^*}$, *then the right hand side of the first condition would be close to the minimum strength. Also, for the second condition, if* $W_{\mathcal{M}^*}$ *is small then the right side is close to* λk_0. *It also weaken the effect of* $\Sigma_{\mathcal{M}^{*c},\mathcal{M}^*}$ *therefore the incoherence restriction would be relaxed in this case.*
(2) If $W_{\mathcal{M}^*}$ *is properly assigned (or* $\|\Sigma_{\mathcal{M}^*}\|_2$ *is comparably large as pointed out and therefore if the right hand sides of two inequalities are close to* $\min|A^*_{\mathcal{M}^*}|$ *and* λk_0, *then* $p(sign(\hat{A}) = sign(A^*))$ *would be greater than that of* L_0 *regularization. Note that the the diagonal entries of* $(\Sigma_{\mathcal{M}^*} + W_{\mathcal{M}^*})^{-1}\Sigma_{\mathcal{M}^*}(\Sigma_{\mathcal{M}^*} + W_{\mathcal{M}^*})^{-1}$ *are smaller than* $\Sigma^{-1}_{\mathcal{M}^*}$. *Thus, the probability of the first condition is smaller. We can say, it requires relaxed signal strength restriction.*
(3) Note that these results can be obtained if the correlation between the relevant predictors is high (that is, $\|\Sigma_{\mathcal{M}^*}\|_2$ *is large). Generally, high correlation is not desirable condition*

because solutions are unstable and accurate selection is hard on this condition. However, $L_0 + L_2$ gives better result if the correlation between relevant subset is high and even it weaken the restriction on the signal strength. We expect that this method can work better than L_0 regularization when correlation is high and signal strength is low.

Proof. *Theorem 5.2* From the hard-TISP, we get the multivariate version of extended KKT condition as given by

$$\hat{A}^j + \frac{\lambda}{k_0}\hat{S}^j = \hat{A}^j - \frac{1}{k_0^2}\Sigma_{j,*}\hat{A} + \frac{1}{k_0^2}X_j^T Y,$$

where $\hat{S} \in \mathbb{R}^{p\times m}$ is multivariate version of generalized sign. Note that $\|\hat{S}^j\|_2 \leq 1$ if $j \notin \hat{\mathcal{M}}$, or zero otherwise. Therefore if $\mathcal{M}^* = \hat{\mathcal{M}}$ then

$$\begin{aligned}(\hat{A} - A^*)_{\mathcal{M}^*} &= \Sigma_{\mathcal{M}^*}^{-1} X_{\mathcal{M}^*} E \\ \|X_{\mathcal{M}^{*c}}^T (I - X_{\mathcal{M}^*}^T \Sigma_{\mathcal{M}^*}^{-1} X_{\mathcal{M}^*}) E\|_2 &\leq \lambda k_0\end{aligned}$$

where the second inequality holds for row-wise. Since $|(\hat{A} - A^*)_{i,j}| < |A^*_{i,j}|$, $\forall i \in [p], j \in [m]$ is required to achieve $sign(\hat{A}) = sign(A^*)$,

$$\|\Sigma_{\mathcal{M}^*}^{-1} X_{\mathcal{M}^*} E\|_{max} \leq \min_{i\in[p], j\in[m]} |A^*_{i,j}|$$

■

Proof. *Theorem 5.3* Similarly to the previous proof, a local solution $\hat{A}$ satisfies

$$\lambda k_0 \hat{S} = X^T Y - (\Sigma + W)\hat{A}$$

where $\|\hat{S}\|_2 = 0$ for $j \in \mathcal{M}^*$ or $\|\hat{S}^j\|_2 \leq 1$ otherwise. Therefore if $\mathcal{M}^* = \hat{\mathcal{M}}$,

$$\begin{aligned}(\Sigma_{\mathcal{M}^*} + W_{\mathcal{M}^*})(\hat{A} - A^*)_{\mathcal{M}^*} &= X_{\mathcal{M}^*}^T E - W A^* \\ \Sigma_{\mathcal{M}^{*c},\mathcal{M}^*}(\hat{A} - A^*)_{\mathcal{M}^*} &= X_{\mathcal{M}^{*c}}^T - \lambda k_0 \hat{S}_{\mathcal{M}^{*c}}.\end{aligned}$$

Since if $\min_{i\in\mathcal{M}^*, j\in[m]} |A^*_{ij}| > \|(\hat{A} - A^*)\|_{\hat{\mathcal{M}}}$ is satisfied in addition to the previous conditions, then $sign(A^*) = sign(\hat{A})$. Thus, if

$$\|(\Sigma_{\mathcal{M}^*} + W_{\mathcal{M}^*})^{-1} X_{\mathcal{M}^*}^T E\|_{max} < \min_{i\in\mathcal{M}^*, j\in[m]} |A^*_{ij}| - \|(\Sigma_{\mathcal{M}^*} + W)^{-1} W_{\mathcal{M}^*} A^*_{\mathcal{M}^*}\|_{max}$$

$$\|X_{\mathcal{M}^{*c}}^T (I - X_{\mathcal{M}^*}\Sigma_{\mathcal{M}^*}^{-1} X_{\mathcal{M}^*}) E\|_{2,\infty} \leq \lambda k_0 - \|\Sigma_{\mathcal{M}^{*c},\mathcal{M}^*}(\Sigma_{\mathcal{M}^*} + W_{\mathcal{M}^*})^{-1} W_{\mathcal{M}^*} A^*_{\mathcal{M}^*}\|_{2,\infty}$$

then $sign(A^*) = sign(\hat{A})$.

■

5.4 Minimax Rates

Assumption 5.1 *There exist positive constants k_1 and k_2 such that for any vector $\Delta \in \mathcal{C}_\delta$ where $\mathcal{C}_\delta := \{\Delta \in \mathbb{R}^{p\times m} : |\Delta| \leq \delta\}$ for any $\delta > 0$, we have*

$$(a) \quad \frac{\|X\Delta\|^2}{n\|\Delta\|^2} \geq k_{1,\delta}^2 \qquad (b) \quad \frac{\|X\Delta\|^2}{n\|\Delta\|^2} \leq k_{2,\delta}^2$$

Proposition 5.1 *[24] For fixed $q \geq 1$, if Assumption 2.1 (a) is satisfied, then*

$$\inf_{\hat{A}} \sup_{\|A^*\|_{2,0}\leq s} E(\|\hat{A} - A^*\|_{2,q}) \geq \frac{\sigma}{k_2}\frac{s^{1/q}}{\sqrt{n}}(m + \log(ep/s))^{1/2}$$

In addition, if Assumption 2.1 (b) is satisfied, then

$$\inf_{\hat{A}} \sup_{\|A^*\|_{2,0}\leq s} E(\|X\hat{A} - XA^*\|_{2,2}) \geq \frac{\sigma}{k_2}s^{1/2}(m + \log(ep/s))^{1/2}$$

Note that the minimax rate in Proposition 5.1 is given under the restricted vector space, $GS_s = \{A^* \in \mathbb{R}^{p\times m} : \|A^*\|_{2,0} \leq s\}$ for given $s \leq p/2$. We now consider minimax rate on the more restricted vector space, $GSR_{s,r} = \{A^* \in \mathbb{R}^{p\times m} : \|A^*\|_{2,0} \leq s \text{ and } \text{rank}(A^*) \leq r\}$. On the former restricted space, the degree of freedom is less than or equal to $s \cdot m$. On the latter space, it would be less than $r(s + m - r)$. We show the improved rate defined on the rank restricted vector space. Note that the minimax rate can be well-defined if the two conditions of Lemma A.4 are satisfied with appropriately defined packing set Θ of matrix space $GSR_{r,s}$. The Lemma A.5 and Lemma A.6 are used to define the dimensionality of packing sets. Under restricted condition on both of rank and sparsity, the number of unknown parameter would be $r(r_0 \wedge s + m - r)$ for $r_0 = \text{rank}(A^*) \leq r$. By using that fact, we will show the improved minimax rate compared to the Lounici et al's.

Theorem 5.4 *Consider the model (5.1) for $p \geq 2$ and $m, n \geq 1$. Suppose that $s \leq p/2$ and let part (b) of Assumption 5.1 be satisfied. Define*

$$\psi_n = \frac{\sigma}{k_2}\frac{1}{\sqrt{n}}\left(r\frac{r_0 \wedge s + m - r}{m} + s\frac{\log(ep/s)}{m}\right)^{1/2},$$

where and $s \geq |\mathcal{M}^*|$ and $r_0 = rank(A^*)$. Then with high probability greater than $1 - \frac{e^{-\alpha}}{1-e^{-\alpha}}$, there exist positive constants b, c which depend only on $l(\cdot)$ such that

$$\inf_{\hat{A}} \sup_{A^*\in GSR_{s,r}} \mathbb{E}l(b\psi_n^{-1}\frac{1}{\sqrt{m}}\|\hat{A} - A^*\|_F) \geq c,$$

where $GSR_{s,r} = \{A^* \in \mathbb{R}^{p\times m} : \|A^*\|_{2,0} \leq s$ *and* $rank(A^*) \leq r\}$ *and in addition, part (a) of Assumption 5.1 is satisfied, then there exist positive constants* b, c *which depend only on* $l(\cdot)$ *such that*

$$\inf_{\hat{A}} \sup_{A^* \in GSR_{s,r}} \mathbb{E} l(b\psi_n^{-1} \frac{1}{k_1\sqrt{nm}} \|X(\hat{A} - A^*)\|_F) \geq c.$$

Remark 5.7 *Without rank restriction, that is* $r = s \wedge m$*, the minimax rate of prediction error is given by*

$$\begin{aligned}\inf_{\hat{A}} \sup_{A^* \in GSR_{s,s\wedge m}} E(\|X(\hat{A} - A^*)\|_F^2 &\geq \sigma^2 \frac{k_1^2}{k_2^2}((s \wedge m)(s \vee m) + s\log(\frac{ep}{s})) \\ &= \sigma^2 \frac{k_1^2}{k_2^2}(sm + s\log(\frac{ep}{s}))\end{aligned}$$

which is the same rate of Lounici's.

Proof. *Theorem 5.4* From (5.9), with possible solutions $\hat{A}_1$ and $\hat{A}_2$, the difference $\hat{A}_1 - \hat{A}_2$ is included in the vector space defined in Assumption 5.1 for $\delta = 4$. We consider first the case $r(r_0 \wedge s + m - r) \leq s\log(ep/s)$. Set $\mathbf{0} = (0, ..., 0) \in \mathbb{R}^m$, $\mathbf{1} = (1, ..., 1) \in \mathbb{R}^m$. Define the set of vectors

$$\Omega = \{\omega \in \mathbb{R}^{p\times m} : \omega^j \in \{\mathbf{0}, \mathbf{1}\}, j = 1, ..., p, |\omega| \leq 2s\},$$

and

$$\mathcal{C}(\Omega) = \{\upsilon\psi_n/s^{1/2}\omega : \omega \in \Omega\},$$

where υ is an absolute constant to be chosen later. Note that $\mathcal{C}(\Omega) \subset GSR_{s,1} \subset GSR_{s,r}$, for any $r \geq 1$.

We have $|\omega_1 - \omega_2| \leq 4s$, $\forall \omega_1, \omega_2 \in \Omega$. Let $A_1 = \upsilon\psi_n\omega_1/s^{1/2}$, $A_2 = \upsilon\psi_n\omega_2/s^{1/2}$. Then $A_1, A_2 \in \mathcal{C}(\Omega)$ and

$$\begin{aligned}\frac{1}{\sqrt{m}}\|A_1 - A_2\|_F &= \upsilon\psi_n/s^{1/2} \cdot \rho(\omega_1, \omega_2)^{1/2} \\ &= \frac{\upsilon\sigma}{k_2\sqrt{ns}} \left(r\frac{s \wedge r_0 + m - r}{m} + s\frac{\log(ep/s)}{m}\right)^{1/2} \cdot \rho(\omega_1, \omega_2)^{1/2}\end{aligned}$$

where $\rho(\omega_1, \omega_2) = \sum_{j=1}^p I\{\omega_1^j \neq \omega_2^j\}$.

For $\theta \in \mathbb{R}^{nm}$, we denote by P_θ the probability distribution of $N(\theta, \sigma^2 I_{nm\times nm})$ Gaussian random vector. We denote by $\mathcal{K}(P, Q)$ the Kullback-Leibler divergence between the

probability measures P and Q. Then, under part(b) of Assumption 5.1,

$$\begin{aligned}\mathcal{K}(P_{XA_1}, P_{XA_2}) &= \frac{1}{2\sigma^2}\|XA_1 - XA_2\|_F^2 \\ &\leq \frac{k_2^2 n}{2\sigma^2}\frac{\upsilon^2\psi_n^2}{s}\rho(\omega_1,\omega_2)m \\ &\leq \upsilon^2[r(s\wedge r_0 + m - r) + s\log(ep/s)]/2 \\ &\leq \upsilon^2 s\log(ep/s) \end{aligned} \tag{5.24}$$

where we used that $\rho(\omega_1,\omega_2) \leq 4s$ for all $\omega_1,\omega_2 \in \Omega$ and $s\log(ep/s) \geq r(s\wedge r_0 + m - r)$. Then Lemma A.5 guarantees the existence of a subset $\mathcal{N}$ of Ω such that

$$\begin{aligned}&\log(|\mathcal{N}|) \geq \bar{c}2s\log(1 + ep/(2s)) \\ &\rho(\omega_1,\omega_2) \geq (1+2s)/4,\ \forall\omega_1,\omega_2 \in \mathcal{N},\ \omega_1 \neq \omega_2,\end{aligned} \tag{5.25}$$

for some absolute constant $\bar{c} > 0$. Then, for all $A_1, A_2 \in \mathcal{C}(\mathcal{N}), A_1 \neq A_2$,

$$\begin{aligned}\frac{1}{\sqrt{m}}\|A_1 - A_2\|_F &\geq \upsilon\psi_n/s^{1/2}\cdot s^{1/2}/\sqrt{2} \\ &= \frac{\upsilon\sigma}{\sqrt{2}k_2\sqrt{n}}\left(r\frac{s\wedge r_0 + m - r}{m} + s\frac{\log(ep/s)}{m}\right)^{1/2} \\ &\geq \upsilon\psi_n/\sqrt{2}\end{aligned}$$

and under part (a) of Assumption 5.1,

$$\begin{aligned}\frac{1}{mn}\|X(A_1 - A_2)\|_F^2 &\geq \frac{k_1^2}{m}\|A_1 - A_2\|^2 \\ &\geq k_1^2\upsilon^2\psi_n^2/2\end{aligned}$$

Furthermore, by (5.24) and (5.25), we have

$$\mathcal{K}(P_{XA_1}, P_{XA_2}) \leq 1/16\log(|\mathcal{N}|) = 1/16\log(|\mathcal{C}(\mathcal{N})|)$$

for all $A_1, A_2 \in \mathcal{C}(\mathcal{N})$ and for an absolute constant υ chosen small enough such that $\upsilon^2/4\bar{c} \leq 1/16$. Thus, two conditions of Lemma A.4 are satisfied, and minimax rate can be defined on the subset $\mathcal{N}$ which is subset of $GSR_{s,r}$.

Consider now the other case $r(s\wedge r_0 + m - r) > s\log(ep/s)$. For rank restricted matrix such that $\text{rank}(A^*) = r_0 \leq r$, the maximal number of unknown parameter would be $r(s\wedge r_0 + m - r)$. Introduce the set of vectors

$$\Omega_1 = \{\omega \in \mathbb{R}^{r\times(s\wedge r_0+m-r)} : \omega = (\omega^1,\cdots,\omega^r), \omega^j \in \{0,1\}^{(s\wedge r_0+m-r)}\}$$

and the associated set $\mathcal{C}(\Omega_1)$ defined as

$$\mathcal{C}(\Omega_1) = \{\upsilon\psi_n/r^{1/2}\omega : \omega \in \Omega\}.$$

We define $\rho_1(\omega_1, \omega_2) := \sum_{i\in[r]} \sum_{j\in[s\wedge r_0+m-r]} I_{\{\omega_{1,ij}\neq\omega_{2,ij}\}}$. Then $\rho_1(\omega_1, \omega_2) \leq 2r(s \wedge r_0 + m - r) \leq 2rm$. Note that $\mathcal{C}\Omega_1 \subset GRS_{s,r}$. We assume first that $r(s \wedge r_0 + m - r) \geq 8$. Then Lemma A.6 guarantees that there exists a subset $\mathcal{N}_1$ of Ω_1 such that

$$\begin{aligned} |\mathcal{N}_1| &\geq 2^{r(s\wedge r_0+m-r)/8}, \\ \rho(\omega_1, \omega_2) &\geq r(s \wedge r_0 + m - r)/8, \ \forall\omega_1, \omega_2 \in \mathcal{N}_1, \ \omega_1 \neq \omega_2. \end{aligned}$$

Define $A_1, A_2 \in \mathcal{C}(\mathcal{N}_1)$ as $A_1 = \upsilon\psi_n\omega_1/r^{1/2}$ and $A_2 = \upsilon\psi_n\omega_2/r^{1/2}$, for $\omega_1, \omega_2 \in \mathcal{N}_1$. Then, for all $\omega_1, \omega_2 \in \mathcal{N}_1$ such that $\omega_1 \neq \omega_2$, then for the A_1 and A_2, Then under part (a) and (b) of Assumption 5.1,

$$\begin{aligned} \frac{1}{n}\|X(A_1 - A_2)\|_F^2 &\geq \frac{k_1^2\upsilon^2\phi_n^2\rho_1(\omega_1 - \omega_2)}{r} \\ \frac{1}{n}\|X(A_1 - A_2)\|_F^2 &\leq \frac{k_2^2\upsilon^2\phi_n^2\rho_1(\omega_1 - \omega_2)}{r} \end{aligned}$$

and

$$\begin{aligned} \frac{1}{\sqrt{m}}\|A_1 - A_2\|_F &\geq \frac{\upsilon\psi_n}{\sqrt{rm}}\|\omega_1 - \omega_2\|_F \\ &\geq \upsilon\psi_n\frac{(r(s \wedge r_0 + m - r)/8)^{1/2}}{\sqrt{rm}} \\ &\geq \upsilon\psi_n/\sqrt{8} \end{aligned}$$

since $s \wedge r_0 + m - r \leq m$, and therefore under part (a) of Assumption 5.1,

$$\frac{1}{mn}\|X(A_1 - A_2)\|_F^2 \geq \frac{k_1^2}{m}\|A_1 - A_2\|_F^2 \geq \upsilon^2 k_1^2\psi_n^2/8$$

Furthermore, for all $A_1, A_2 \in \mathcal{C}(\mathcal{N}_1)$ under part (b) of Assumption

$$\begin{aligned} \mathcal{K}(P_{XA_1}, P_{XA_2}) &= \frac{1}{2\sigma^2}\|XA_1 - XA_2\|_F^2 \\ &\leq \frac{nk_2^2\upsilon^2\phi_n^2}{2\sigma^2 r}\rho(\omega_1, \omega_2) \\ &\leq \upsilon^2\frac{r(s \wedge r_0 + m - r) + s\log(2p/s)}{m}\frac{2rm}{2r} \\ &\leq 2\upsilon^2 r(s \wedge r_0 + m - r) \end{aligned} \tag{5.26}$$

Therefore

$$\mathcal{K}(P_{XA_1}, P_{XA_2}) \leq 2\upsilon^2 r(s + m - r) \leq 1/16\log(|\mathcal{C}(\mathcal{N}_1)|),$$

where the last inequality holds for an small absolute constant $\upsilon > 0$ chosen small enough. Then again by Lemma A.4 the constant minimax rate is defined on the subset $\mathcal{C}(\mathcal{N}_1)$.

Finally, we consider the other case where $r(s+m-r) > s\log(ep/s)$. Still (5.26) is satisfied and $r(s \wedge r_0 + m - r) < 8$. Also $\rho_1(\omega_1, \omega_2) \leq 2r(s \wedge r_0 + m - r) \leq 16$. Consider a subset $\mathcal{N}_2$ such that

$$\rho(\omega_1, \omega_2) \geq r(s \wedge r_0 + m - r), \forall \omega_1 \neq \omega_2 \in \mathcal{N}_2$$

then $|\mathcal{C}(\mathcal{N}_2)| \geq 2^{r(s \wedge r_0 + m - r)}$. Therefore

$$\begin{aligned} \mathcal{K}(P_{XA_1}, P_{XA_2}) &\leq 2\upsilon^2 r(s \wedge r_0 + m - r) \\ &\leq 1/16 \log|\mathcal{C}(\mathcal{N}_2)|) \end{aligned}$$

for some small constant υ. Then again from Lemma A.4 we can calculate the constant minimax rate. ■

CHAPTER 6

SIMULATION AND APPLICATION TO THE LEUKAEMIA DATA

6.1 Simulation

We have studied l_0 and $l_0 + l_2$ regularization compared to l_1 regularization. The main improvement via l_0 or l_0+l_2 compared to Lasso is that the required conditions on the design matrix or the true target model are relaxed, both for prediction and feature selection. The conditions for prediction accuracy and selection accuracy are related at some point as we showed in the comparison of the conditions, but they are intended for the different aims.

For Lasso, to control the prediction performance, the restricted sub matrices should be positive definite. Therefore, the sparsity constraint on the target vector and the corresponding restricted eigen value condition are required for acquiring prediction accuracy. For l_0 or l_0+l_2, the required conditions are relaxed than those for Lasso (For prediction performance, we do not need any restrictions). In order to achieve selection accuracy, a large signal-to-noise strength and small incoherence condition should be satisfied. The small incoherence does not mean exactly the small correlation among covariates. For the selection accuracy, the relevant subset $\mathcal{J}^*$ and irrelevant subset $\mathcal{J}^{*c}$ should have low correlation in order to be separated easily. For Lasso, the strict incoherence condition is required, but for L_0 or $L_0 + L_2$ the restriction is relaxed.

We follow the similar simulation setup, method and the evaluation criterions as [30]. We will compare Lasso, l_0 and $l_0 + l_2$ on various cases. The sample size and dimensionality are fixed as $n = 100, p = 500$, to see the performance on high dimensional setup such that $p > n$. The sparsity of true model is fixed as 5 and the i th column of A^*, $A_i^* \in \mathbb{R}^p$ is $[d, d, 0, d, d, d, 0, ...]$ for any $i \in [m]$. On this sparsity setup (5 out of 500), the restriction on sparsity for Lasso would not be violated. The strength of signal is set as $d = 0.5, 2.5$.

Also, the correlation of ith and jth predictors is defined as $\rho^{|}i-j|$, for any $i,j \in [p]$. To examine the regularization methods on the situation where the condtion is violated (RE condition or incoherence restriction for Lasso), we set the correlation level ρ is 0.1, 0.5, or 0.9. Therefore, in our setup, we can compare the performance of the three methods on different correlation and incoherence levels, and with different signal strengths. Here is the summary of experiments.

ex1: $[corr(X_i, X_j) = 0.1^{|i-j|}, \min|\hat{A}_{\mathcal{M}^*}| = 2.5 \ / \ p = 200, n = 100, m = 100/ \ |\mathcal{M}^*| = 5]$
ex2: $[corr(X_i, X_j) = 0.5^{|i-j|}, \min|\hat{A}_{\mathcal{M}^*}| = 2.5 \ / \ p = 200, n = 100, m = 100/ \ |\mathcal{M}^*| = 5]$
ex3: $[corr(X_i, X_j) = 0.9^{|i-j|}, \min|\hat{A}_{\mathcal{M}^*}| = 2.5 \ / \ p = 200, n = 100, m = 100/ \ |\mathcal{M}^*| = 5]$
ex4: $[corr(X_i, X_j) = 0.1^{|i-j|}, \min|\hat{A}_{\mathcal{M}^*}| = 0.5 \ / \ p = 200, n = 100, m = 100/ \ |\mathcal{M}^*| = 5]$
ex5: $[corr(X_i, X_j) = 0.5^{|i-j|}, \min|\hat{A}_{\mathcal{M}^*}| = 0.5 \ / \ p = 200, n = 100, m = 100/ \ |\mathcal{M}^*| = 5]$
ex6: $[corr(X_i, X_j) = 0.9^{|i-j|}, \min|\hat{A}_{\mathcal{M}^*}| = 0.5 \ / \ p = 200, n = 100, m = 100/ \ |\mathcal{M}^*| = 5]$

For unskewed comparison between the methods, we let tune the parameters λ, w over wide grid lines so to give the optimal result for each regularization method. To select w optimally, we first find the optimal w for ridge regression, then with the reference w_R, we tune λ for fixed $w = [0.5w_R, 0.05w_R, 0.005w_R]$. After tuning λ we tune w again with the fixed λ. Also, for validation, we generate another 10^4 datasets and also test the solution $\hat{A}$ on an another 10^4 datasets to give a precise result. We evaluate the results on the similar criterion as in [30]. We run 50 times of each combination of regularization method and the setup, and report only the middle 10 results out of 50 (the rest information are also used for ordering). The prediction performance can be evaluated by the prediction error $\|Y - X\hat{A}\|_F^2$, sde $(= 100\{\sum_{i\in[n]} \log f(Y_i; \hat{A})/\sum_{i\in[n]} \log f(Y_i; A^*) - 1\})$ and the estimation error $\|\hat{A} - A^*\|_F^2$. For selection performance, we show masking error $(= |\mathcal{M}^* \backslash \hat{\mathcal{M}}|)$ denoted by M, and swamping error$(= |\hat{\mathcal{M}} \backslash \mathcal{M}^*|)$ denoted by S, and sign accuracy $(= P(sign(\hat{A}) = sign(A^*)))$. Note that smaller M and S, and greater sign accuracy are better. We first show the result when the minimum signal strength is not low ($d = 2.5$). In Table 6.1, both L_0 and $L_0 + L_2$ show better performance than Lasso on both the prediction and selection (even though when correlation is high, Lasso has smaller estimation error than the others). Especially all methods do not have masking error which is more serious error than swamping error, on this relatively high signal strength. However, Lasso has greater swamping error than the others for every cases and less sign accuracy. Note that in this high signal strength setup, L_0 regularization slightly works better than $L_0 + L_2$ regularization.

Table 6.2 shows the results on the relatively low signal strength setup. In comparison to the previous case, $L_0 + L_2$ works slightly better than L_0 both on prediction and selection

Table 6.1: Minimum signal strength = 2.5

		Lasso		l_0		$L_0 + L_2$	
	$\|\hat{A} - A^*\|_F^2$	0.00038	(0.00003)	0.00011	(0)	0.0001	(0)
ex1	$\|Y - X\hat{A}\|_F^2$	0.06816	(0.0045)	0.05221	(0.00073)	0.05193	(0.00076)
	sde	6.769	(0.46324)	5.15782	(0.0542)	5.12613	(0.05971)
$\rho = 0.1$	M	0.00%	(0.00%)	0.00%	(0.00%)	0.00%	(0.00%)
init strength = 2.5	S	3.32%	(3.47%)	0.00%	(0.00%)	0.00%	(0.00%)
	$P(sign(\hat{A}) = sign(A^*))$	96.71%	(3.43%)	100.00%	(0.00%)	100.00%	(0.00%)
	$\|\hat{A} - A^*\|_F^2$	0.00018	(0.00001)	0.00015	(0)	0.00014	(0)
ex2	$\|Y - X\hat{A}\|_F^2$	0.08766	(0.00239)	0.05197	(0.00079)	0.0525	(0.00082)
	sde	8.66565	(0.26311)	5.14119	(0.04729)	5.21185	(0.07337)
$\rho = 0.5$	M	0.00%	(0.00%)	0.00%	(0.00%)	0.00%	(0.00%)
init strength = 2.5	S	11.32%	(2.64%)	0.00%	(0.00%)	0.00%	(0.00%)
	$P(sign(\hat{A}) = sign(A^*))$	88.79%	(2.62%)	100.00%	(0.00%)	100.00%	(0.00%)
	$\|\hat{A} - A^*\|_F^2$	0.00023	(0.00001)	0.00093	(0.00003)	0.00143	(0.00013)
ex3	$\|Y - X\hat{A}\|_F^2$	0.12199	(0.00443)	0.06074	(0.00062)	0.08999	(0.01186)
	sde	12.05459	(0.43856)	5.97938	(0.05192)	8.92127	(1.14825)
$\rho = 0.9$	M	0.00%	(0.00%)	0.00%	(0.00%)	0.00%	(0.00%)
init strength = 2.5	S	13.22%	(2.76%)	0.14%	(0.13%)	0.43%	(0.07%)
	$P(sign(\hat{A}) = sign(A^*))$	86.90%	(2.73%)	99.86%	(0.13%)	99.57%	(0.07%)

(it shows the smallest prediction, estimation errors, and sde both for moderately or highly correlated predictors. See the second and the third rows of the table). Also, overall performance of Lasso both for prediction and selection is worse than the others. Of couse, Lasso shows better performance when the correlation is low (but still not as good as the others).

To see why $L_0 + L_2$ works better than L_0, we examine two regularization methods on extreme cases. Note that Lasso shrinks large values so it can achieves smaller prediction or estimation errors. First is $\min|A^*_{\mathcal{M}^*}| = 20$ and $\rho = 0.1$ (ex7), and the other is $\min|A^*_{\mathcal{M}^*}| = 0.1$ and $\rho = 0.9$ (ex8). Table 6.3 shows when the signal is strong (20) then L_0 works much better than Lasso on selection, even though on prediction the results are comparable. However, when signal is low (=0.1) (and correlation is high), L_0 underselect ($|\mathcal{M}^*| = 2$) and masking error is too high. Since masking error should be considered more seriously, Lasso would be better method in this case since the masking error is zero even though it still overselects. In this setup, L_0+L_2 shows much better performance than the others on selection, the masking error is zero, and $|\hat{\mathcal{M}}| = 6$ which is much smaller than the cardinality of selected subset via Lasso. As we discussed before, both L_0 and Lasso have one regularization parameter used both for selection and prediction, and it is usually tuned in terms of prediction accuracy.

Table 6.2: Minimum signal strength = 0.5

		Lasso		l_0		$L_0 + L_2$	
	$\|\hat{A} - A^*\|_F^2$	0.00012	(0.00001)	0.00011	(0)	0.0001	(0)
ex4	$\|Y - X\hat{A}\|_F^2$	0.03826	(0.00088)	0.05409	(0.00126)	0.05176	(0.00161)
	sde	3.81928	(0.07768)	5.33908	(0.12144)	5.11467	(0.13354)
$\rho = 0.1$	M	0.00%	(0.00%)	0.00%	(0.00%)	0.00%	(0.00%)
init strength = 0.5	S	6.38%	(1.68%)	0.00%	(0.00%)	0.00%	(0.00%)
	$P(sign(\hat{A}) = sign(A^*))$	93.69%	(1.67%)	99.95%	(0.10%)	99.94%	(0.12%)
	$\|\hat{A} - A^*\|_F^2$	0.00013	(0)	0.00018	(0.00001)	0.00015	(0)
ex5	$\|Y - X\hat{A}\|_F^2$	0.06956	(0.00214)	0.06081	(0.00085)	0.05309	(0.0007)
	sde	6.87595	(0.24599)	6.00214	(0.09806)	5.22406	(0.08439)
$\rho = 0.5$	M	0.00%	(0.00%)	0.00%	(0.00%)	0.00%	(0.00%)
init strength = 0.5	S	10.05%	(2.46%)	0.14%	(0.13%)	0.14%	(0.13%)
	$P(sign(\hat{A}) = sign(A^*))$	89.79%	(2.55%)	99.85%	(0.14%)	99.85%	(0.14%)
	$\|\hat{A} - A^*\|_F^2$	0.00019	(0.00001)	0.00107	(0.00002)	0.00034	(0.00001)
ex6	$\|Y - X\hat{A}\|_F^2$	0.10344	(0.0034)	0.06613	(0.00109)	0.0375	(0.00085)
	sde	10.26445	(0.33531)	6.55319	(0.12796)	3.71069	(0.08725)
$\rho = 0.9$	M	0.00%	(0.00%)	0.00%	(0.00%)	0.00%	(0.00%)
init strength = 0.5	S	12.54%	(2.64%)	0.20%	(0.00%)	0.20%	(0.00%)
	$P(sign(\hat{A}) = sign(A^*))$	87.58%	(2.61%)	99.76%	(0.00%)	99.80%	(0.00%)

Since Lasso shrinks large values, therefore for strong signal, λ can be selected too small for Lasso and it results in a over-selection. On the other hand, when signal is too low, then λ for L_0 can be selected too large (since it does not shrink large values), and it may result in the under-selection. $L_0 + L_2$ can give better result on selection, because it has two regularization parameters of which one works for selection (λ) and the other works for shrinkage to enhance the prediction (w).

Figure 6.1 shows the solution paths via $l_0 + l_2$, l_0 and l_1 regularization when the sparsity is 5 and SNR is 4. See that the convergence speed is slower for Lasso than for the others. Also, It is apprantely shown that the solution of l_0 regularization is the most sparse. Before the convergence, $l_0 + l_2$ fluctuates more than l_0 since the weight for l_2 is updated so the selected subset is keep changed.

6.2 Leukaemia Data Analysis

We apply the $l_0 + l_2$ regularization to the Leukemia Data [48] to see the performance of the regularization on the highly correlated design matrix. The Leukemia data consist of

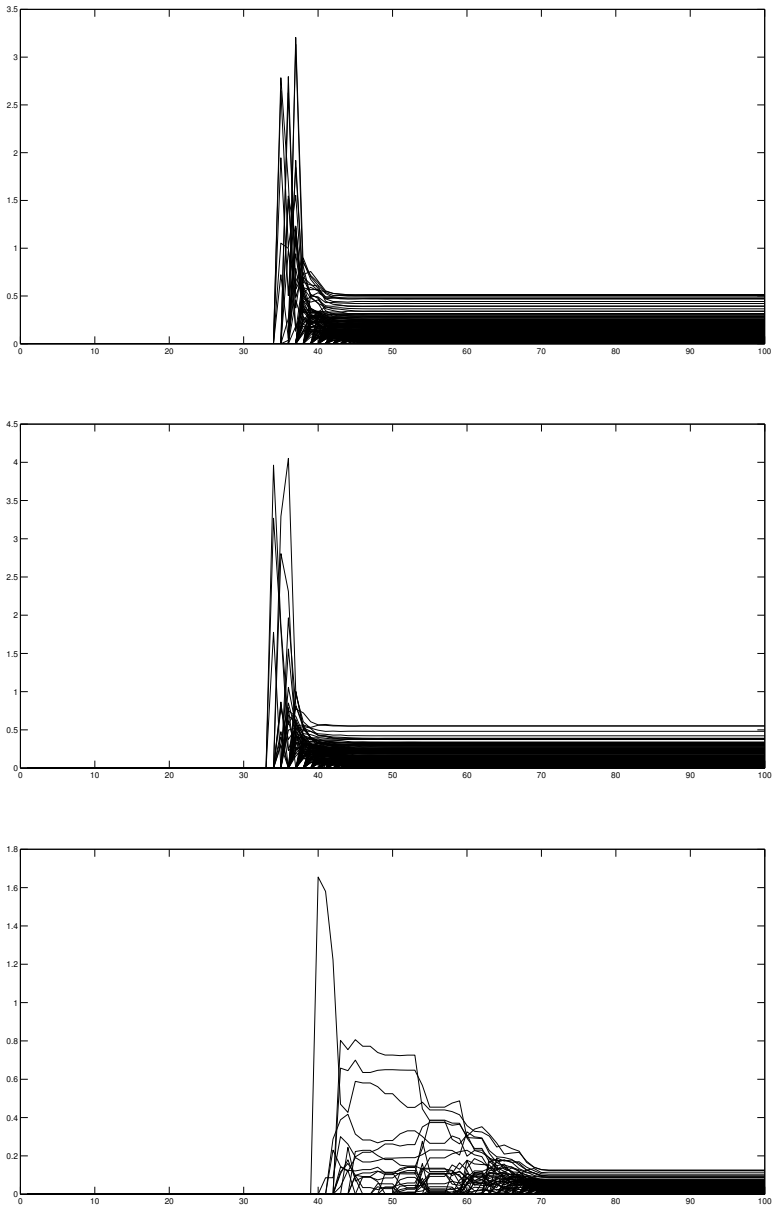

Figure 6.1: Solution path via $l_0 + l_2$, l_0 and l_1 regularization, when $J^* = 5$ and SNR $= 4$

Table 6.3: Special cases

		Lasso	l_0	$L_0 + L_2$
ex7	$\|\hat{A} - A^*\|_F^2$	$2.14E-04$	$1.04E-04$	
$\min\|A^*_{\mathcal{M}^*}\| = 20$	$\|Y - X\hat{A}\|_F^2$	0.11163	0.05151	
	M	0	0	
$\rho = 0.1$	S	0.105	0	
	$P(sign(\hat{A}) = sign(A^*))$	0.896	1	
	$\|\hat{\mathcal{M}}\|$	57	5	
ex8	$\|\hat{A} - A^*\|_F^2$	$2.92E-05$	$3.77E-04$	$3.47E-05$
$\min\|A^*_{\mathcal{M}^*}\| = 0.1$	$\|Y - X\hat{A}\|_F^2$	$1.86E-02$	$3.25E-02$	$1.12E-02$
	M	0	0.8	0
$\rho = 0.9$	S	0.0404	0.0020	0.0020
	$P(sign(\hat{A}) = sign(A^*))$	0.9599	0.9898	0.9976
	$\|\hat{\mathcal{M}}\|$	25	2	6

Table 6.4: Comparison with other methods

Method	10-fold CV error	Test error	Number of genes
Golub	3/38	4/34	50
Support vector maching-recursive feature elimination	2/38	1/34	31
Penalized Logistic regression-recursive feasure elimination	2/38	1/34	26
Nearest shrunken centroids	2/38	2/34	21
Elastic net	3/38	0/34	45
$L_0 + L_2$	2/38	3/34	3

7129 genes and 72 samples of which 38 samples are training set, and the rest 34 samples are test set [17] The response variable is a type of leukaemia (type 1 leukaemia is Acute lymphoblastic leukaemia and type 2 leukaemia is acute mycloid leukaemia). The goal is to correctly classify the type of leukaemia by construncting the classification rule with genes. There are 27 of type 1 leukaemia and 11 of type 2 leukaemia in the training set. We apply $l_0 + l_2$ regularization in the similar way as done in [48]. First we coded type 1 leukaemia as 0 and the other as 1. We prescreened the data based on the t-value of each variable and select the 1000 most significant variables, then applied the $l_0 + l_2$ regularization. Table 6.4 shows the results in comparison with other methods done in the past. We got the 2/38 of ten fold cross validation error as shown in the table, which is a comparable results to the other methods listed on the table. However the test error is very low and selected set is only 3 which is very sparse solution compared to the others. This result shows that the $l_0 + l_2$ regularization performs poorly for the classification when the data is highly

correlated. The purpose of our method is to select a sparse subset and achieve prediction accuracy, while for this data analysis the aim is to construct classification rule. For instance, if the covariates are highly correlated and the underlying model is not linear then classifying correctly by $l_0 + l_2$ regularization is difficult since the estimator via the method is acquired by optimizing penalized least square based on a linear model. Also, note that the elastic net only deducts the same amount from every selected values at each iteration, therefore the pairwise correlation is not changed, while $l_0 + l_2$ shrinks the selected values proportionally therefore the pairwise correlation is getting smaller. Therefore, it is hard to select the whole group within which covariates are pairwisely correlated. Indeed, $l_0 + l_2$ might be able to select sparse subset and converges faster than the others by de-correlating the pairwise correlation at each step. Figure 6.2 shows the solution path via $l_0 + l_2$ regularization. See that the solution path of some covariates are not monotonely decreasing or increasing. The solution lines fluctuate much and as the iteration increases they diverge. From this graph we can see the possibility of nonlinear underlying model(indeed the response is binary variable therefore this analysis may be inappropriate). Since elastic net can preserve the pairwise correlation among covariates therefore the classification would be much easier than $l_0 + l_2$, but $l_0 + l_2$ cannot.

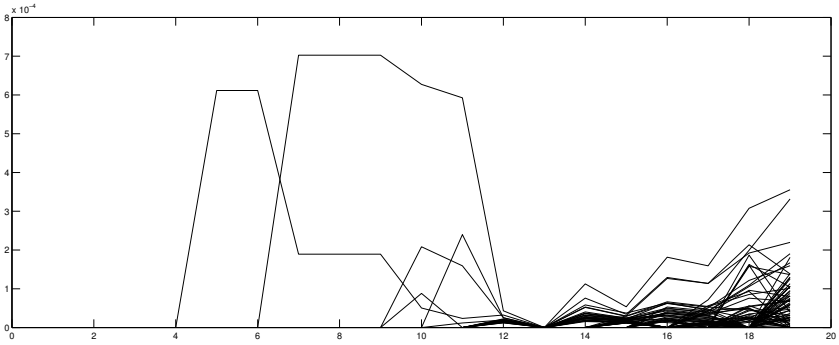

Figure 6.2: Solution path via $l_0 + l_2$ regularization on data analysis

CHAPTER 7

DISCUSSION

Our goal is to achieve accurate prediction and selection consistency as well as to reduce dimensionality. We focused on penalized least square estimation methods as one of shrinkage methods, of which many are based on l_1 penalty functions. Group L_1 regularization which uses group L_1 penalty, gives sparse solution therefore we can reduce dimensionality but usually prediction is not accurate and performance on selection is poor as well. Since for prediction accuracy, the L_1 regularization requires very stringent conditions on design matrix and restrictions on sparsity (such that correlation between predictors should be low enough and the sparsity of true model should be small), in most cases it does not predict accurately. Also, strict incoherence condition (correlation between irrelevant predictors and relevant predictors should be low) and high signal strength are required for L_1 to achieve the sign agreement (sign of solution is equal to sign of true model). To relax the required conditions and obtain better method which works in general cases (even when the correlation is high or the true model is not sparse), we came to nonconvex penalty. One of nonconvex penalty we came to L_0 regularization which gives more parsimonious model and involves no bias on the selected subset.

Our non-asymptotic studies showed that Group L_0 regularization obtained prediction accuracy and selection consistency in more relaxed condition such as highly correlated design matrix with large sparsity. The oracle inequalities are obtained without any assumption for the prediction error and selection error and with empirically much relaxed condition than that of group L_1 on the design matrix (and without sparsity restriction).

$$\frac{\|X(\hat{A} - A^*)\|_F^2}{\log p + m} + \frac{\|\hat{A} - A^*\|_{2,q}^2}{((\log p + m)/n)^{q/2}} + |\hat{\mathcal{M}}| = O_p(|\mathcal{M}^*|).$$

The error rates are optimal in the minimax sense, and identical to those via group L_1. The difference is that for L_1 these rates are able to be obtained only under restricted conditions.

There are inflation factor $\log p + m$, the number of predictors is in logarithm form and the number of response variables in in plain form. Asymptotically speaking, to achieve estimation consistency $\log p + m \lesssim o(n)$ and m should have smaller order than $\log p$ to remain the meaningful quantative realtionship between p and n such that $p \lesssim o(exp(n))$. Also, note that these optimal rates hold for any approximately sparse model (that is there are small number of strong signals and large number of weak signals close to zero) as well since we do not make any assumption on the true sparsity and the oracle inequalities (what we have shown earlier) hold for any model matrix.

Furthermore, for selection consistency, Group L_0 can work even when the incoherence (that is correlation between the relevant predictors' group and the irrelevant predictors' group) is high. For Group L_1, stringent incoherence condition is required (e.g. mutual incoherence condition, irrepresentable condition) and the required minimum signal strength depends on the incoherence condition so if the incoherence condition is weaker then the signal-to-noise ratio (SNR) is required at larger rate which includes $|\mathcal{M}^*|$ term so to gives high SNR when the sparsity is not low. For Group L_0, the incoherence condition is relaxed and it only requires lower overall correlation. Also, SNR is not strictly required preliminarily but affects only on the probability of sign agreement.

Therefore, L_0 can work better than L_1 regularization both on prediction and selection in most cases. However, empirically L_0 showed poor performance on selection in some extreme cases where signal is too low and correlation is high. Since there is only one regularization parameter which controlls both selection and prediction and it is tuned in terms of prediction accuracy, the parameter is not optimally tuned for selection. Therefore, in the extreme cases, λ is chosen to be too large so it may under-select and causes masking error (error of missed selection). Usually L_0 does better job but in this cases L_1 can be prefered, so we want to solve this issue (which is not only for L_0 but also for L_1).

Group $L_0 + L_2$ uses the combined penalty function of l_0 and l_2, and there are two regularization parameters which controlls selection and shrinkage to enhance prediction accuracy independently. Therefore, the parameters can be tuned optimally for its uses respectively, and it shrinks the selected subset via l_0 so that it could gives smaller MSE as refered to James-Stein phenomenon (especially when the selected subset is not sparse or the correlation between selected subset is high). We can check these fact from our

non-asymptotic studies as well. The rates of errors are obtained as follows

$$\|X(A^* - \hat{A})\|_F^2 \lesssim O_p(\|W^{1/2}A^*\|_F^2 + (\log p + m)|\mathcal{M}^*|)$$
$$\|A^* - \hat{A}\|_F^2 \lesssim O_p(\|WA^*\|_F^2 w_{min}^{-2} + (\log p + m)|\mathcal{M}^*|w_{min}^{-1})$$
$$|\hat{\mathcal{M}} \backslash \mathcal{M}^*| \lesssim O_p(\|WA^*\|_F^2 (w_{min}(\log p + m))^{-1} + |\mathcal{M}^*|).$$

There is additional term involved in the upper bound. To obtain the optimal error rate, the weight of l_2 penalty W should be assigned properly. At least $\|WA^*\|_F^2 w_{min}^{-1}$ should be smaller rate than $(\log p + m)/n$ to obtain the same optimal rate as L_0 regularization. However, there is possibility to achieve lower rate of estimation error from the smart choice of W. If we can ideally choose W with the assumption that σ, A^* are known, then estimation error can be upper bounded by smaller rate or at least smaller rist can be achieved. We also tried to adopt the empirical Bayesian idea for choosing W (as an analogue of James-Stein estimator) and smaller rist can be obtained. All optimal choice of W can be interpreted as σ^2 to squred signal ratio.

For selection, $L_0 + L_2$ regularization shows its potential either. Our theories show that if $W_{\mathcal{M}^*}$ is comparably smaller than $\Sigma_{\mathcal{M}^*}$ (or $\Sigma_{\mathcal{M}^*}$ is large) then the effect of incoherence term is vanished and the required minimum signal strength is also able to be lower. That means, if correlation between the relevant predictors is high, then $L_0 + L_2$ can do better job than L_0 regularization even when signal strength is low. Therefore, $L_0 + L_2$ can be applied to the cases where L_0 cannot work well. Our simulation study showed this fact, but not so clearly because the tuning grid line of W was coarse so the optimal choice of W could not be achieved.

We also studied the minimax rate with the constraint of rank and sparsity. The prediction loss is $r(r_0 \wedge s + m - r) + s\log(p/s)$ where r_0 is true rank of A^*, r and s are the constraints of rank and sparsity. The rate is improved than the minimax rate with sparsity constraint [24]. If true model has low rank (lower than the sparsity), then it has smaller rate.

CHAPTER 8

FUTURE WORKS

For $L_0 + L_2$ method, the choice of W, the weight of l_2 penalty, takes the important roles, such that it can lower the rate of estimation loss or weaken the incoherence condition and lower the required minimum signal strength for achieving sign agreement. Therefore, we want to study further on the smart choice of W, since we only show that its potential with ideal choices of W (assume some of true parameters are known). An empirical choice of W (as an analogue of James-Stein estimator) achieves a less risk than L_0 under the assumption that σ is known. However, it also can be applicable with different loss function.

$$F(A) = \frac{1}{2}\rho + p_g(\|A^j\|_2; \lambda)$$

where ρ is a robust regression function. For instance, if $\rho = l_1$ norm [1], then λ may not include unknown parameter σ term and the oracle upper bound contains no σ term (however, the loss function is not strictly convex, so numerical difficulty may be caused). Therefore, we expect the empirical choice of W can be applicable in this way. The empirical choice of W may be able to be combinded to Hard-Ridge TISP, by updating the value at each iterate. Also, we expect better estimation accuracy by applying block thresholding to this model. As shown by the optimal choice of W, it is possible to reduce the rate of estimation error. Even though A^* or σ are assumed to be known, it may be able to improve the estimation accuracy (at least up to constant) when the range of signal strength is wide (for example, small number of strong signals and large number of weak signals). Especially when the correlation between predictors is high, this method may be able to work better.

In this dissertation, we assume a Gaussian noise only. We will extend our research to non-Gaussian noise such as multivariate GLM or multivariate logistic regression. On the analysis of Leukaemia data, $L_0 + L_2$ did not do a better job than elastic net, since the underlying model of the data may be not linear one and elastic net is better for classification.

However, with the assumption of non-Gaussian noise, we can extend the method to nonlinear model.

We have shown the Hellinger distance for studying the convergence rate, but the result is not optimal one. Since the bracket of the domain vector space was not tight enough, we got the lower convergence rate even than that of Kullback-Leiber distance which should be greater than Hellinger distance. We will revise the bracketing by adopting a soft sparsity (l_q-ball) defined in [27].

APPENDIX A

ANCILLARY RESULTS

Lemma A.1 *Let W_d be a χ^2_d random variable with d degrees of freedom. Then*

$$P[W_d > (1+\epsilon)\cdot d] \leq \exp\left[-\frac{d}{2}\left\{\epsilon - \log(1+\epsilon)\right\}\right] \leq \exp\left[-\frac{d\epsilon^2}{4(1+\epsilon)}\right]$$

for some positive ϵ

Proof. *Proof of Lemma A.1* (See [10], p856-857.)

$$P[\eta_\upsilon > x] \leq exp[\varphi_\upsilon(x)] \ , \quad \varphi_\upsilon(x) = \frac{1}{2m^2(\upsilon)}\log[1+\sqrt{2}xm(\upsilon)] - \frac{x}{\sqrt{2}m(\upsilon)}.$$

where $\eta_\upsilon = (\sqrt{2}\|\upsilon\|)^{-1}\sum_{i=1}^{\infty}\upsilon_i(\xi_i^2-1)$, $\xi \sim$ i.i.d $N(0,1)$ and $m(\upsilon) = \sup_i |\upsilon_i|/\|\upsilon\|$. Apply $\upsilon = \mathbf{1}_d$, $\eta_\upsilon = (\sqrt{2d})^{-1}(W_d - d)$, and $x = d(\sqrt{2d})^{-1}\epsilon$. Then

$$\varphi_\upsilon(x) = \frac{d}{2}\log\left[1+x\sqrt{\frac{2}{d}}\right] - x\sqrt{\frac{d}{2}} = \frac{d}{2}\log(1+\epsilon) - \frac{d}{2}\epsilon = -\frac{d}{2}\left[\epsilon - \log(1+\epsilon)\right]$$

Also, for $\epsilon > 0$, we have the second inequality as following.

$$\log(1+\epsilon) - \epsilon = \epsilon\int_0^1\left(-\frac{\tau\epsilon}{1+\tau\epsilon}\right)d\tau \leq -\int_0^1\frac{\tau\epsilon^2}{1+\epsilon}d\tau = -\frac{\epsilon^2}{2(1+\epsilon)}.$$

■

Lemma A.2 *Let $\varepsilon \sim N(0,\sigma^2 I_N)$ and $P_\mathcal{J} = Z_\mathcal{J}(Z_\mathcal{J}^T Z_\mathcal{J})^- Z_\mathcal{J}^T$ for any $\mathcal{J} \subset [K]$. Define events*

$$A_{\lambda_0,\mathcal{J}} = \left\{\|\varepsilon^T P_\mathcal{J}\|_2^2 \leq \lambda_0^2|\mathcal{J}|\right\},$$

and

$$\mathcal{A}_{\lambda_0} := \bigcap_{\{\mathcal{J}\subset[K]\}\cup\{\mathcal{J}=\emptyset\}} A_{\lambda_0,\mathcal{J}}$$

for given λ_0. For given $\alpha > 0$, if λ_0 is chosen as

$$\lambda_0 \geq \sigma\sqrt{\alpha + 2 + \log K},$$

then the probability of $\mathcal{A}_{\lambda_0}$ is greater than $1 - \frac{e^{-\alpha}}{1-e^{-\alpha}}$.

Proof.

Let $P = Z(Z^TZ)^-Z^T$, where $(Z^TZ)^-$ is its Moore-Penrose inverse. Similarly, let $P_{\mathcal{J}} = Z_{\mathcal{J}}(Z_{\mathcal{J}}^TZ_{\mathcal{J}})^-Z_{\mathcal{J}}^T$, for any index set $\mathcal{J}$. Let $U\Lambda U^T \equiv U_2U_2^T$ be the spectral decomposition of $P_{\mathcal{J}} \in \mathbb{R}^{N\times N}$, where $U_2 \in \mathbb{R}^{N\times q_{\mathcal{J}}}$, for $q_{\mathcal{J}} = \text{rank}(Z_{\mathcal{J}}) \leq J \wedge N \leq J$. Since $\varepsilon \sim N(0, \sigma^2 I_N)$, $U_2^T\varepsilon$ can be written as a $q_{\mathcal{J}} \times 1$ column vector with Gaussian entries on top of $N - q_{\mathcal{J}}$ vector of zeros. Therefore, $\|\varepsilon^T P_{\mathcal{J}}\|^2 = \varepsilon^T U_2U_2^T\varepsilon \sim \sigma^2 \cdot \chi^2_{q_{\mathcal{J}}}$ and $E(\|\varepsilon^T P_{\mathcal{J}}\|^2) = \sigma^2 q_{\mathcal{J}}$.

Now, consider the $A_{\lambda_0,\mathcal{J}}$. We apply Lemma A.1 to upper bound the probability of $\mathcal{A}^c_{\lambda_0}$. For fixed $\lambda_0^2 > 1$ and for $W_J \sim \chi^2_J$,

$$\begin{aligned}
P(\mathcal{A}_{\lambda_0}) &\leq \sum_{\mathcal{J}\subset[K]} P(A^c_{\lambda_0,\mathcal{J}}) + P(A^c_{\lambda_0,\emptyset}) \\
&\leq \sum_{J=1}^{K} \binom{K}{J} P\left[W_J > (1 + \lambda_0^2/\sigma^2 - 1)J\right] + 0 \\
&\leq \sum_{J=1}^{K} \binom{K}{J} \exp\left[-J\frac{(\lambda_0^2/\sigma^2 - 1)^2}{4\lambda_0^2/\sigma^2}\right] \\
&\leq \sum_{J=1}^{K} \exp\left[-J\left\{\rho^2 - \log\left(\frac{Ke}{J}\right)\right\}\right] \\
&\leq \frac{e^{-\alpha}}{1 - e^{-\alpha}} \quad \text{, for some } \alpha > 0
\end{aligned}$$

where $\rho = (\lambda_0/\sigma - \sigma/\lambda_0)/2$. Note that $P(A^c_{\lambda_0,\emptyset}) = 0$ since $\|\varepsilon P_{\mathcal{J}}\|_2 = 0$ and we applied the fact that $\binom{K}{J} \leq \left(\frac{Ke}{J}\right)^J$. From the last inequality, since $\min_{J\in[K]}[(\lambda_0/\sigma - \sigma/\lambda_0)^2/4 - \log(Ke/J)] \geq \alpha$,

$$(\lambda_0/\sigma - \sigma/\lambda_0) \geq 2(\alpha + \log K + 1)^{1/2},$$

and thereby

$$\lambda_0 \geq \frac{2\sigma\sqrt{\alpha + 1 + \log K} + 2\sigma\sqrt{\alpha + 2 + \log K}}{2}.$$

Therefore, for given $\alpha > 0$ if λ_0 is chosen as,

$$\lambda_0 \geq 2\sigma\sqrt{\alpha + 2 + \log K},$$

then the probability of $\mathcal{A}_{\lambda_0}$ is greater than $1 - \frac{e^{-\alpha}}{1-e^{-\alpha}}$.

■

Lemma A.3 *[32] Let $z \sim \mathcal{N}(0, D_{d\times d})$ with Δ as its diagonal matrix. If $x \sim \mathcal{N}(0, \Delta)$ then for any positive numbers $c_1, \cdots, c_d$,*

$$P\left[\ \cap_{i\in[d]}\{|z_i| \leq c_i\}\ \right] \geq \prod_{i\in[d]} P[|x_i| \leq c_i]$$

Lemma A.4 *(Theorem 2.7 in [34]) Let w ge a loss function and let $A > 0$ be such that $w(A) > 0$. Assume that Θ contains elements $\theta_0, ..., \theta_M, M \geq 1$, such that:*

$$(a) \quad d(\theta_j, \theta_k) \geq 2s > 0, \quad \forall 0 \leq j < k \leq M$$

$$(b) \quad P_j \gg P_0, \quad \forall j = 1, .., M \quad \textit{and}$$

$$\frac{1}{M}\sum_{j=1}^{M} K(P_j, P_0) \leq \alpha \log M$$

$$\textit{with} \quad 0 < alpha < 1/8 \quad \textit{and} \quad P_j = P_{\theta_j}, j = 0, ..., M$$

Then for $\psi = s/A$ we have

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} E_\theta[w(\psi^{-1} d(\hat{\theta}, \theta))] \geq c(\alpha) w(A).$$

Lemma A.5 *(Rigollet and Tsybakov [28]) For any $p \geq 1$, $k \in \{1, ..., p-1\}$, let ω_k^p be the subset of Ω defined by:*

$$\Omega := \left\{ w \in \{0,1\}^p : \sum_{j=1}^{p} w_j = k \right\}$$

Let $p \geq 2$ and $1 \leq k \leq p$ be two integers and define $\bar{k} = min(k, p-k)$. Then there exists a subset $\mathcal{N}$ of Ω such that the Hamming distance $\rho(w, w') = \sum_{j=1}^{p} I(w_j \neq w'_j)$ satisfies

$$\rho(w, w') \geq \frac{\bar{k}+1}{4}\ , \quad \forall w, w' \in \mathcal{N} : w \neq w'$$

and

$$\log(|\mathcal{N}|) \geq C_1 \bar{k} \log\left(1 + \frac{ep}{\bar{k}}\right)$$

for some numerical constant $C_1 \geq 9 \cdot 10^{-4}$

Lemma A.6 *(Varshamov-Gilbert bound). Let $p \geq 8$. The there exists a subset $\{\omega_0, ...w_M\}$ of Ω such that $\omega_0 = (0, ..., 0)$.*

$$\rho(w_j, w_i) \geq p/8, \forall 0 \leq i < j \leq M$$

and

$$M \geq 2^{p/8}$$

Lemma A.7 *(Lemma A.3 in [24]) Let $Ts \geq 8$. If $\omega_1, \omega_2 \in \mathcal{N}_1$ such that $\rho(\omega_1, \omega_2) \geq Ts/8$, then the cardinality of the set $J(\omega_1, \omega_2) = \{\sum_{j=1}^{Ts} > T/16\}$ is greater than or equal to $s/16$.*

REFERENCES

[1] Alexandre Belloni and Victor Chernozhukov. 1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, February 2011.

[2] Peter J. Bickel. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, August 2009.

[3] Peter J. Bickel and Yaácov Ritov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, August 2009.

[4] Florentina Bunea. Honest variable selection in linear and logistic regression models via l_1 and $l_1 + l_2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194, August 2008.

[5] Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[6] T. Tony Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *The Annals of Statistics*, 27(3):898–924, June 1999. ISSN 0090-5364.

[7] T. Tony Cai and Ming Yuan. Adaptive covariance matrix estimation through block thresholding. (1211.0459), November 2012. Annals of Statistics 2012, Vol. 40, No. 4, 2014-2042.

[8] Emmanuel Cands and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, December 2007.

[9] Emmanuel J. Cands. Modern statistical estimation via oracle inequalities, 2006.

[10] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *The Annals of Statistics*, 30(3):843–874, June 2002.

[11] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, August 1994.

[12] David L. Donoho, Michael Elad, and Vladimir N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE TRANS. INFORM. THEORY*, 52(1):618, 2006.

[13] Bradley Efron. Empirical bayes and the james-stein estimator. University Lecture, 2000.

[14] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 2001.

[15] Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2001.

[16] Ildiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, May 1993.

[17] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, October 1999. PMID: 10521349.

[18] Mohamed Hebiri and Sara van de Geer. The smooth-lasso and other 1+2-penalized methods. *Electronic Journal of Statistics*, 5:1184–1226, 2011. ISSN 1935-7524. Mathematical Reviews number (MathSciNet): MR2842904.

[19] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. ISSN 0040-1706.

[20] Jian Huang and Huiliang Xie. Asymptotic oracle properties of SCAD-penalized least squares estimators. *IMS Lecture Notes-Monograph Series*, 55:149–166, September 2007.

[21] Jinzhu Jia and Bin Yu. On model selection consistency of the elastic net when p>> n. *Statistica Sinica*, 20:595–611, 2010.

[22] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378, October 2000.

[23] Karim Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

[24] Karim Lounici, Massimiliano Pontil, Alexandre B. Tsybakov, and Sara van de Geer. Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, 39 (4):2164–2204, 2011.

[25] Nicolai Meinshausen. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, February 2009.

[26] Mina Ossiander. A central limit theorem under metric entropy with l2 bracketing. *The Annals of Probability*, 15(3):897–919, July 1987. ISSN 0091-1798.

[27] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression overl_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, October 2011.

[28] Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *Annals of Statistics*, 39(39):731–771, March 2011.

[29] Yiyuan She. Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of Statistics*, 3:384–415, 2009.

[30] Yiyuan She. An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Comput. Stat. Data Anal.*, 56(10):29762990, October 2012.

[31] Xiaotong Shen. On methods of sieves and penalization. *The Annals of Statistics*, 25 (6):2555–2591, December 1997. ISSN 0090-5364.

[32] Zbynek Sidăk. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, June 1967.

[33] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[34] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 1st edition, November 2008. ISBN 0387790519.

[35] Sara van de Geer and Peter Bhlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[36] M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183 –2202, May 2009.

[37] David Williams. *Probability with Martingales*. Cambridge University Press, February 1991. ISBN 0521406056.

[38] Fei Ye and Cun-Hui Zhang. Rate minimaxity of the lasso and dantzig selector for the l_q loss in l_r balls. *Journal of Machine Learning Research*, 11:3519–3540, December 2010.

[39] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):4967, 2006. ISSN 1467-9868.

[40] Cun-Hui Zhang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, August 2008.

[41] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2), February 2010.

[42] Cun-hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, August 2008.

[43] Cun-Hui Zhang and Tong Zhang. A general theory of concave regularization for high dimensional sparse estimation problems. *arXiv:1108.4988*, August 2012.

[44] Tong Zhang. Some sharp performance bounds for least squares regression with l1 regularization. *The Annals of Statistics*, 37(5):2109–2144, October 2009.

[45] Tong Zhang. Multi-stage convex relaxation for feature selection. *arXiv:1106.0565*, June 2011.

[46] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

[47] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[48] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

BIOGRAPHICAL SKETCH

The author was born in Korea and got a Bachelor and Master degree in Statistics. After that she came to Florida State University for pursuing Doctoral degree in Statistics.