HIGH CONFIDENCE SET REGULARIZATION IN SPARSE HIGH DIMENSIONAL LOGISTIC REGRESSION WITH MEASUREMENT ERROR

by

Maorong Rao

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2018

Approved by:

____

Dr. Jiancheng Jiang

____

Dr. Hui-Kuan Tseng

____

Dr. Weihua Zou

____

Dr. Zhiyi Zhang

ABSTRACT

MAORONG RAO. HIGH CONFIDENCE SET REGULARIZATION IN SPARSE HIGH DIMENSIONAL LOGISTIC REGRESSION WITH MEASUREMENT ERROR. (Under the direction of DR. JIANCHENG JIANG)

The nature of complexity of high dimensional data diminishes the efficacy of the classical statistics inference. Regularization technique has been actively developed in response to derive revolution inference.

l_1 based regularization such Lasso [13] and Dantizg Selector [5] succeed in two aspects. First, the inherent sparsity of l_1 accords with the underlying nature of high dimensional data; second, the convexity essence paving the way to computational feasibility in high dimension. Based on the idea provided by Dantzig Selector, James, G. M. and P. Radchenko extended an algorithm [33] to solve Dantzig Selector for generalized linear model. Fan [8] abstracted this framework to the set of convex loss function as High Confidence Set. To fill the gap of theoretical support within this framework, we derive the bound of prediction error and parameter error beyond the scope of logistic loss. We termed this classifier as High Confidence Set Selector (HCS). An implicit assumption of high confidence set selection is that the data is collected precisely. However, the data is inevitable to process with measurement error in reality. In response to this challenge, a new methodology (MHCS) accounts for measurement error was introduced. We further derive the theory and algorithm.

Our simulation study provides strong numerical support that compared with other popular regularization methods, e.g., LASSO, Ridge, and HCS, MHCS ad-

vances in restore information from measurement error. And due to embedded linearity instinct, HCS and MHCS is versatile to connect with state of art technique such as word vectors, deep network, transfer learning, etc. We demonstrate the cutting edge applications in two real examples.

ACKNOWLEDGMENTS

First and foremost, I would like to express my special appreciation and thanks to my advisor, Dr. Jiancheng Jiang, for his trust, support and patience in the past 6 years. It has been a rich but stressful journey. Studying in mathematics with a Medicine background isn't the easiest thing to do. It was his guidance and motivation kept me moving forward and achieved this point.

I would like to thank the rest of my dissertation committee, Dr. Zhiyi Zhang, Dr. Weihua Zou, and Dr. Hui-Kuan Tseng, for serving as my committee members, providing me necessary assistant during this process. Thanks for all your dedication and commitment to high level educations.

I want to express my especial thanks to Dr. Hongyi Li, Dr. Robert Tibshirani, Dr. Martin Wainwright and Dr. Larry Wasserman. I would not have been possible to get here without the online courses and knowledge shared by you.

My humble gratitude also goes to everyone who has been part of my life for the past 6 years. I wouldn't have been here without all your companion, support and criticism.

I am grateful to all my loving and supportive friends, Anqi, Bingqing, Chen, Jianing, Kaifan, Siwei, Yaqi, Yongming, I am so lucky to have you in my life.

Finally, to my parents, I want to say thank you, but it is far less than enough. This is the 7th year that I am away from home, they have been nothing but support for every step that I made. Thank you for standing behind me for everything, even though we are thousands of miles apart. I couldn't have been luckier to be your

daughter.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

"High dimensional data are nowadays rule rather than exception." [4]

In high dimensional setting, the dimensions d is larger than sample size n, sometimes even grows faster with the sample size increasing. For example, in many contemporary applications, microarray data is frequently in thousands or beyond, while the sample size n is typically in the order of tens. "The central conflict in high dimensional setup is that the model complexity is not supported by limited access to data." Fan points out the essential challenge in high dimensional statistics [8]. In other words, the "variance" of conventional models is high in such new settings, and even simple models such as LDA need to be regularized.

This limit inclines the chance of overfitting. Basically, if the number of parameters is larger than the sample size, with un-regularized empirical risk minimization approach, a model can be selected with perfect performance in training simply by memorizing the training sample other than generalize the trend of signal from population. In other words, it may fail severely to predict the unseen data.

Basically, if the number of parameters is larger than the sample size, with un-regularized empirical risk minimization approach, a model can be selected with perfect performance in training simply by memorizing the training sample other than generalize the trend of signal from population. In other words, it may fail severely to predict the unseen data.

In order to develop statistical inference in high dimensional setting, which lead to reasonable accuracy or asymptotic consistency. It is crucial to pare down the high degree of complexity to its bare essentials.

A natural underlying form of simplicity in high dimension is sparsity, we hope that the nature of the world is not so complex as it might be. Loosely speaking, a sparse statistical model is one in which only a relatively small number of parameters (or predictors) play an important role. "it's possible to develop high dimensional statistical inference, if $log(p) \times (sparsity(\beta)) << n$."[4]

We refer to Hastie et al. [13] and Buhlmann et al. [4] for overviews of statistical challenges associated with high dimensionality.

In addition to the embedded simplicity, the other principle in high dimension stat is efficiency in algorithm.

The convexity of l_1 norm bring success in the efficiency of optimization, accompany with embedded sparsity, l_1 regularization prevails decades in recovering the underlying signal in high dimension data.

l_1 constrain enjoys two important properties. First, it is naturally sparse, i.e., it has a large number of zero components. Second, it is computationally feasible even for high-dimensional data whereas classical procedures such as BIC are not feasible when the number of parameters becomes large.

Fan[8] introduces a closely related regularization methodology in high dimension stat, which the fundamental idea is to select the sparsest member measured by l_1 norm in a set which carries the information of data, termed as high confidence set. We elaborate the idea as follow:

Assume a random sample from the population (X, Y) are collected in the form $(X_1, Y_1), \ldots, (X_n, Y_n)$, the loss function $\rho_\beta(X, Y)$ has the form $\rho_\beta(X, Y) = \rho(X^T\beta, Y)$, which is assumed to be convex.

$\beta^* \in \mathbb{R}^d$ is the target parameter which minimizes the expected loss $E\rho(X^T\beta, Y)$, that is:

$$\beta^* = \arg\min_{\beta \in R^d} E\rho(X^T\beta, Y)$$

Our target is to find an estimate of β^* through empirical risk minimization. Denote the empirical loss as

$$L_n\rho(\beta) = \frac{1}{n}\sum_{i=1}^{n} \rho(X_i^T\beta, Y_i);$$

and the gradient with respect to β as $\nabla_\beta L_n\rho(\beta)$, the high confidence set is constructed as follow:

$$C_\lambda = \{\beta \in \mathbb{R}^d : \|\nabla L_n\rho(\beta)\|_\infty < \lambda\},$$

where the tuning parameter λ is chosen related to the confidence level viz

$$Pr(\beta^* \in C_\lambda) = Pr\{\|\nabla L_n\rho(\beta)\|_\infty < \lambda\} > 1 - \delta$$

The high confidence set C_λ inherits the information about β^* from sample data. In addition, as we discuss above, if we impose the sparsity on the underlying parameter β^*, with this assumption, a natural solution is selecting the sparsest solution in the high confidence set, viz.

$$\beta = \arg\min_{\beta \in C_\lambda} \|\beta\|_1$$

With this generalized framework, several works can be considered as examples of high confidence set selection with specific loss measure. For instances, Dantzig Selector [5] can be viewed as high confidence set estimation for linear regression with quadratic loss; Cai and Liu [6] propose Linear programming discriminant rule (LPD) for two Multi-Gaussian distributed data, which apply the high confidence set selection with measured of log likelihood ratios of Bayes rule. Barut [2] extends the above linear discriminant rule through high confidence set selection under measurement error scenario.

Inspired by this idea, we apply this method to regularize high dimensional logistic regression. We term this method as High Confidence Set Selector (HCS).

An implicit assumption of HCS is that the data is collected precisely, however, in reality, the measure is inevitable to process with noise and missing value. In many real application, such as image recovery and speech recognition, most problems are subject to measurement error.

There are various studies concern on correction of measurement error. Within the context of estimate distribution of measurement error, estimators proposed in studies [34], [35], [36], [41] yield sound asymptotic results by approach maximum likelihood.

However, under high dimensional setting, the distribution of measurement error is too complex to capture. Methods proposed in [40], [42], [32] which accounts for measurement error without requiring estimation of its distribution, stand out in practical application in high dimensional setting.

In order to account for measurement error, we develop the model with addititve measurement error proposed in [40] to generalized linear model with logistic loss.

We denote the modified classifier as MHCS.

Through out this paper, we will introduce High Confidence Set Selector and its theoretical properties in Chapter 2. The extended method accounts for measurement error (MHCS) are introduced in Chapter 3. Implementation algorithm and numerical simulation are elaborated in Chapter 4; Applications in real world data are illustrated in Chapter 5.

CHAPTER 2: HIGH CONFIDENCE SET ESTIMATION

2.1 Model Setup and Methodology

Consider a measurable space $\mathcal{M} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{0, 1\}$, and $(x_i, y_i)_{i=1}^n \in \mathcal{M}$ is a set of n i.i.d. random pairs of observations; $\phi(\cdot)$ is a set of bounded real value functions $\phi = (\phi_1, \ldots, \phi_d)$, $\|\phi(\cdot)\|_\infty < M_d$ [15], which maps original features from $\mathcal{X}$ to $\mathcal{Z} \in \mathbb{R}$, e.g., $\phi : \mathcal{X} \to \mathcal{Z} \in \mathbb{R}^d$.

Defined the parametric space $\Omega : (f, \phi)$, for a given ϕ, let $Z = \phi(X)$, then the generalized logistic regression model defined in parametric space $\Omega : (f, \phi)$ can be modeled as:

$$Pr(Y = 1|Z) = \frac{\exp f(Z)}{1 + \exp f(Z)} \, ,$$

where $f : \mathcal{Z} \to \mathbb{R}$, is the log odds ratio, i.e.,

$$f(Z) = \log \frac{Pr(Y = 1|Z)}{Pr(Y = 0|Z)},$$

Denote ρ_f as the loss function of generalized logistic regression given $Z = \phi(X)$, then,

$$\rho_f(Z, Y) = Yf\ (Z) - \log \Big\{ 1 + \exp\ \ f\ (Z) \Big\};$$

denote the corresponding empirical loss as L_n, then

$$L_n\, \rho_f = \frac{1}{n} \sum_{i=1}^{n} \ \Big\{ \ Y_i f(Z_i) - \log \big[\ 1 + \exp\ \ f\ \ (Z_i)\ \big] \ \Big\} \, .$$

The expected risk $L\,\rho_f$ is the expectation of loss given f, it holds

$$L\,\rho_f = E\,(\,\rho_f\,) = E\,(\,L_n\,\rho_f\,).$$

Given $Z = \phi(X)$, denote f_0 as the best parameter in Ω which minimizes $L_n\ \rho_f$, e.g.:

$$f_0 = \arg\min_{f \in \Omega} L\,\rho_f.$$

For a set of given ϕ, consider the linear subspace $\Omega_\beta(\phi, f_\beta) \subset \Omega(\phi, f)$, such that:

$$f_\beta(Z) = \beta^T Z.$$

Correspondingly, in this linear subspace, the loss function is

$$\rho_\beta(Z, Y) = Y\beta^T Z - \log\left[\,1 + \exp\left(\,\beta^T Z\,\right)\right];$$

and the empirical loss is

$$L_n\,\rho_\beta\,(Z, Y) = \frac{1}{n}\sum_{i=1}^{n}\ \rho_\beta(Z, Y) = \frac{1}{n}\sum_{i=1}^{n}\ \left\{\ Y_i\,\beta^T Z_i - \log\left[\,1 + \exp\left(\beta^T Z_i\right)\right]\ \right\}.$$

Denote the expected loss as $\ L\,\rho_\beta$,

$$L\,\rho_\beta\,(Z, Y) = E\left[\,\rho_\beta\,(Z, Y)\,\right] = E\ \left[\ L_n\,\rho_\beta\,(Z, Y)\ \right]$$

The optimal parameter β^* in linear subspace is defined as the one minimizes the expected loss, e.g.,

$$\beta^* = \arg\min_{\beta \in \Omega_\beta} L\,\rho_\beta\,(Z, Y); \tag{1}$$

It holds that:

$$\left.\frac{\partial\, L\, \rho_{\beta}\,(Z,Y)}{\partial \beta}\right|_{\beta^*} = 0 \tag{2}$$

In classical statistics setting, with fixed dimension of β, as $n \rightarrow \infty$, by asymptotic theory, we can achieve $\nabla_{\beta}\, L_n\, \rho_{\beta^*}(Z,Y) \rightarrow 0$ in probability. However, in high dimensional statistics, d is larger than n, sometimes even grows faster than n, we cannot expect $\nabla_{\beta}\, L_n\, \rho_{\beta^*}(Z,Y) = 0$ to hold exactly, however, we would expect $\|\, \nabla_{\beta}\, L_n\, \rho_{\beta^*}(Z,Y)\, \|_{\infty} \leq \lambda$ with large probability when appropriate λ is chosen. Therefore, it's straightfoward to define the high confidence set as follow:

$$\mathcal{C}(\lambda) = \{\beta \in \mathbb{R}^d : \|\, \nabla_{\beta}\, L_n\, \rho_{\beta}\,(Z,Y)\, \|_{\infty} \leq \lambda\}; \tag{3}$$

where λ is chosen such that

$$Pr\left\{\ \beta^* \in \mathcal{C}(\lambda)\ \right\} = Pr\left\{\ \|\, \nabla_{\beta}\, L_n\, \rho_{\beta^*}(Z,Y)\, \|_{\infty} \leq \lambda\ \right\} \geq 1-\delta \tag{4}$$

for a positive sequence $\delta \rightarrow 0$.

Then we select the solution with minimum l_1 norm in $\mathcal{C}(\lambda)$ as a proxy of β^*, we termed this estimatior as High Confidence Set Selector (HCS):

$$\hat{\beta}_{HCS} = \underset{\beta \in \mathcal{C}(\lambda)}{\arg\min}\ \|\ \beta\ \|_1 \tag{5}$$

2.2 Theoretic Property of High Confidence Set Estimation

In this section we investigate the theoretical properties of High Confidence Set Selector in three aspects. First, we show that, with appropriate choice of λ, β^* falls in $C(\lambda)$ with high probability. Second, we derive the generalized prediction error bound of High Confidence Set Selector in terms of excess risk. With the assumptions of sparsity and restricted strong convexity [29], we derive the parameter error bound in third result.

The following assumptions are used in theoretical study:

Assumption. *A_1: $(Z_i, Y_i)_{i=1}^n$ $are \;\; i.i.d.$*

Assumption. *A_2: $\|\phi(\cdot)\|_\infty < M_d$;*

Remark. Assumption A_1 and Assumption A_2 are general assumptions in the literature regards generalized error bound in l_1 regularization and learning theory ([15], [16], [17], [18], [19]). In pratical, various data collected bounded, such as the image data which ranges from 0 to 255 in RGB; Sets of base function $\{\phi\}$ can outputs in nature, such as sigmoid function, softmax function ranges from 0 to 1; The output of feature transformed based on the similarity such as wordvector, neural networks with certain activate function, ranges from (-1,1). Addition advantage of this setting is that, X is distribution free, which avoids the complexity of density estimation in high dimension statistics.

Assumption. *A_3: $M_d\sqrt{log2d} \sim \mathcal{O}(\sqrt{n})$*

Assumption. *A_4: Construct a sequence $\{a_j\}_{j=0}^{J-1}$, $\;a_j = 2a_{j-1}$, for $\;\forall a_0 > 0$, there exists* *a positive integer $J < \infty$, such that,*

$$a_{J-1} = a_0\, 2^J \geq 2\|\beta^*\|_1$$

Remark. Assumption A_3 and Assumption A_4 are technique assumption.

Assumption. A_5*:* $\| \ \beta^* \ \|_0 \ \leq \ s$

Remark. Assumption A_5 assume the target parameter β^* is s-sparse, which means the maximum number of nonzero components of β^* is s, i.e., $\|\beta^*\|_0 = s$.

This assumption is widely used in high dimensional setting, we refer [43] for general reviews.

Assumption (A_6 Restricted Strong Convexity)**.**

$$\delta\, L_n\, \rho_{(\,\Delta,\, \beta^*\,)}\, (\,Z,\, Y\,) \geq\ \kappa \,\|\, \Delta\, \|_2 \|\beta^*\|_1 < \infty.$$

Remark. The restricted strong convexity assumption is the key assumption in derivation of parameter error bound. Define the support set S by a mapping nonzero components of β^* to the index set as follow: S :=$\{j : \beta^*_j \neq 0\}$, $|S|$ =s.

Denote Δ as deviation in the neighbor of β^*, $\Delta = \beta - \beta^*$;

The Restricted Strong Convexity Assumption is defined as [29]:

$$\begin{aligned} &\delta\, L_n\, \rho_{(\,\Delta,\, \beta^*\,)}\, (\,Z,\, Y\,) \\ &= L_n\, \rho_{(\,\beta^*+\Delta\,)}(\,Z,\, Y\,) - L_n\, \rho_{\beta^*}(\,Z,\, Y\,) - \left\langle\, \nabla_\beta L_n\, \rho_{\beta^*}(\,Z,\, Y\,),\, \Delta\,\right\rangle \\ &\geq \kappa\, \|\, \Delta\, \|_2 \qquad (6) \\ &for \quad \|\Delta_{S_c}\|_1 \leq \|\Delta_S\|_1. \end{aligned}$$

where S is the index set we defined before, S^c is complementary set of S.

The strong convexity in geometry is the curvature of loss function, we use the empirical loss to track the population performance, once we have the estimator $\hat{\beta}$, we prefer it is robust against the perturbation in empirical loss. If strong convexity exists, the solution to the parameter estimation will not change much to a small perturbation in empirical loss, it's therefore a stable solution. While in weak curvature, opposite effect occurs, small perturbation in empirical loss would cause parameter shifts enormously in parameter space.

From Theorem 2.2, we can see the excess risk is tracked by the l_1 norm of $\|\hat{\beta}-\beta^*\|$, in high dimension scenario, where $n << p$, there exists space with low curvature such that β^* is far away from $\hat{\beta}$, but it will not arouse fluctuation in empirical loss function, the main idea is to restrict the target parameter lies in these directions. By l_1 regularization, $\|\hat{\beta}\|_1 \leq \|\beta^*\|_1$, apply the lemma from basis pursuit [45], we have following property for $\hat{\Delta}$: $\|\hat{\Delta}_{S_c}\|_1 \leq \|\hat{\Delta}_S\|_1$.

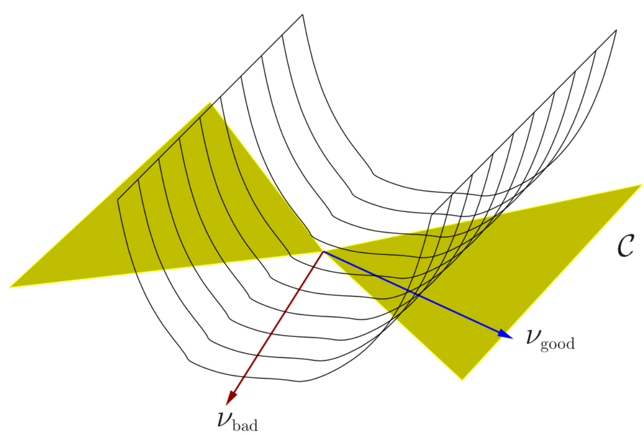


Figure 1: Illustration of Restricted Strong Convexity [43]

In high dimension setting, while we can't expect the strong convexity exists in every directions, we can expect it exists in the direction of $\hat{\Delta} : \|\hat{\Delta}_{S_c}\|_1 \leq \|\hat{\Delta}_S\|_1$. In Figure 1 we illustrate the 'restricted direction', where the shadow direction is the desiered.

To be simplify, the notation used in next section are listed below.

Notation:

$$\lambda^* \equiv \sqrt{2}\, M_d \sqrt{\frac{\log\,(\,2\,d\,) + \log n}{n}}\;;$$
$$\delta_1 \equiv \frac{2M_d}{n};$$
$$\delta_2 \equiv 2\,M_d \sqrt{\frac{2\,log\,2d}{n}}$$
$$\delta_0 = \delta_1 + \delta_2$$

2.2.1 High Confidence Set

Recall that the high confidence set defined in (3), as discuss in previous section, we expect the optimal linear solution β^* falls in the high confidence set with high probability when appropriate λ is chosen. Define

$$Event\ \ A := \{\beta^* \in C_\lambda\},$$

then we have following theorem

Theorem 2.1 (Event A)**.** *With* β^* *defined in* (1)*, and* C_λ *defined in* (3)*, under Assumption* $A_1 - A_3$*, if* $\lambda > \lambda^*$*, it holds that:*

$$P\;(\;\beta^* \in \mathcal{C}_\lambda\;) > 1 - \frac{1}{n}.$$

2.2.2 Prediction Error

Define our solution set:

$$\mathcal{B}_\lambda := \left\{ \beta \in R^d : \beta = \underset{\beta \in \mathcal{C}(\lambda)}{\arg\min} \; \| \beta \|_1 \right\} \tag{7}$$

The relationship between optimal linear solution β^*, solution to HCS ($\hat{\beta}_{HCS}$), linear parameter space (Ω_β), High Confidence Set (C_λ) and Solution Set of HCS ($\mathcal{B}_\lambda$) is illustrated in Figure 2.

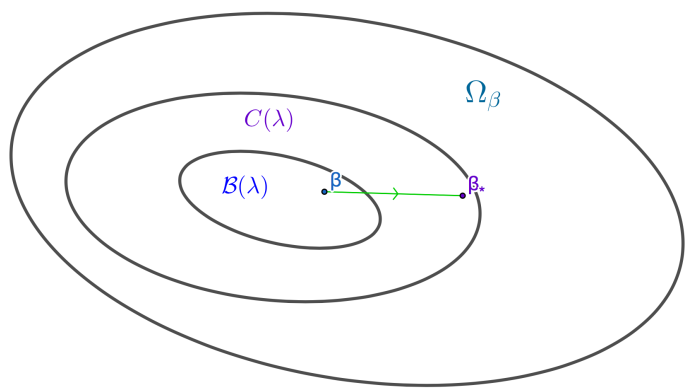


Figure 2: The relationship between β^*, $\hat{\beta}_{HCS}$, Ω_β, C_λ and $\mathcal{B}_\lambda$

We defined the excess risk of $\hat{\beta} \in \Omega_\beta$ as:

$$\mathcal{E}(\hat{\beta}) = L\rho_{\hat{\beta}} - L\rho_{\beta^*}.$$

The prediction error bound in terms of excess risk is derived in Theorem 2.2.

Theorem 2.2 (Prediction Error Bound)**.** *Denote the solution to HCS as β_{HCS}, under Assumption $A_1 - A_4$, when $\lambda > \lambda^*$, with probability at least $1 - 2J\, e^{-2n} - \frac{1}{n}$, where J is* *a positive integer satisfies Assumption A_4, it holds that:*

$$\mathcal{E}(\hat{\beta}_{HCS}) \leq (\,\lambda + \delta_0\,)\; \|\, \beta^* - \hat{\beta}_{HCS}\, \|_1 \;+\; \delta_0\, a_0$$

2.2.3 Parameter Error

Theorem 2.3 (Parameter Error Bound)**.** *Under Assumption $A_1 - A_6$, when $\lambda \geq \lambda^*$, with probability at least $1 - \frac{1}{n}$, it holds:*

$$(i) \quad \|\, \hat{\beta}_{HCS} \;-\; \beta^* \,\|_2 \;\leq\; \frac{4\,\lambda\,\sqrt{s}}{\kappa};$$

$$(ii) \quad \|\, \hat{\beta}_{HCS} \;-\; \beta^* \,\|_1 \;\leq\; \frac{8\,\lambda\,s}{\kappa}$$

Corollary (Prediction Error Bound)**.** *Under Assumption $A_1 - A_6$, when $\lambda > \lambda^*$, with probability at least $1 - 2J\,e^{-2n} - \frac{1}{n}$, it holds that:*

$$\mathcal{E}(\hat{\beta}_{HCS}) \leq \; \frac{8\;\lambda\,s}{\kappa}\,(\,\lambda + \delta_0\,) + \; \delta_0\; a_0$$

CHAPTER 3: THEORETICAL STUDY OF HIGH CONFIDENCE SET ESTIMATION WITH MEASUREMENT ERROR

3.1 Background and Model setup

As discuss in Chapter 1, the measurement error is inevitable in reality.

Consider model with additive measurement error. Instead of (X, Y), we observe (U, Y), where $X \in \mathcal{X}, and \ \ U \in \mathcal{X}$.

Analogous to model setup in Chapter 2, $\phi(\cdot) : \mathcal{X} \to \mathcal{Z}$ is a set of base function with $\|\phi(\cdot)\|_\infty \leq M_d$. After features transformation by $\phi(\cdot)$, we have (W, Y), where

$$W = \phi(U);$$

And the additive measurement error Ξ is defined as:

$$\Xi = \phi(U) - \phi(X)$$

For simplicity, denote $Z = \phi(X)$, thus,

$$W = Z + \Xi.$$

According to Theorem 2.1, β^* is feasible in C_λ if λ is chosen appropriately. However, the presence of measurement error leads the high confidence set lost its efficacy.

To see this, if we roughly plug the achievable measure W into the high confidence set,

$$C_\lambda = \{\beta : \|\nabla_\beta L_n(W, Y, \beta)\|_\infty < \lambda\}$$

$E(\nabla_\beta L_n(X, Y, \beta^*)) = 0$ thus for $\forall \lambda > 0$, $\nabla_\beta L_n(X, Y, \beta^*) \to 0$ if $n \to \infty$ however, $E(\nabla_\beta L_n(W, Y, \beta^*))$ is not necessary to be 0, thus for given λ, β^* may not in C_λ even $n \to \infty$.

In the case of linear regression, Rosenbaum and Tsybakov [40] introduced an addition parameter γ to bound the magnitude of the measurement error in the matrix uncertainty selector (MUS), which yielding the following two bounds:

$$\|W\epsilon\|_\infty < \lambda$$

and

$$\|\Xi\|_\infty < \gamma$$

where W is obeservation, Ξ is the measurement error and ϵ is the residual. These bounds are sufficient condition for β^* is feasible with high probability in following set:

$$\{\beta : \|W(Y - W\beta)\|_\infty < \lambda + \gamma\|\beta\|_1\}.$$

Inspired by this idea, we develop a modified high confidence set $C(\lambda, \gamma)$ for logistic regression. Note that logistic loss can be expressed in the form of mean function $\mu(Z\beta)$, thus it can be expressed in the following form:

$$\|\nabla_\beta \, L_n \rho_\beta \, (W, Y)\|_\infty = \frac{1}{n} \, \| \, W^T \, [\, Y - \mu \, (W\beta) \,] \, \|_\infty;$$

where

$$\mu\,(\,W\beta\,) = \frac{\exp\,(\,W\beta\,)}{1+\exp\,(\,W\beta\,)} \;\in\; (\,0,1\,).$$

By model assumption,

$$W\beta = Z\beta + \Xi\beta;$$

Thus by Taylor expansion and Cauchy residual theorem,

$$\mu\,(\,W\beta\,) = \mu\,(\,Z\beta\,) + \mu'\,(\,\xi\,)\,(\,\Xi\beta\,)$$

where ξ lies in the segment between $W\beta$ and $Z\beta$.

Then by triangle inequality, $\frac{1}{n}\;\|\,W^T\;[\;Y-\mu\;(\,W\beta\,)\;]\;\|_\infty$ can break into two parts:

$$\begin{aligned}\frac{1}{n}\;\|\,W^T\;[\;Y-\mu\;(\,W\beta\,)\;]\;\|_\infty =& \frac{1}{n}\,\|\,W^T\,[\,Y-\mu\,(\,Z\beta\,)-\mu'\,(\,\xi\beta\,)\,(\,\Xi\beta\,)\,]\,\|_\infty\\ \leq& \frac{1}{n}\,\|\,W^T\,[\,Y-\mu\,(\,Z\beta\,)\,]\,\|_\infty + \frac{1}{n}\,\|\,W^T\,\mu'\,(\,\xi\,)\,(\,\Xi\beta\,)\,\|_\infty\\ \leq& \frac{1}{n}\,\|\,W^T\,[\,Y-\mu\,(\,Z\beta\,)\,]\,\|_\infty + \frac{1}{n}\,\|\,W^T\,\mu'\,(\,\xi\,)\,\Xi\,\|_\infty\|\beta\|_1\end{aligned}$$

Thus it's intuitive to construct the high confidence set which accounts for measurment error as follow:

$$C(\lambda,\gamma)\;=\{\;\frac{1}{n}\,\|\,W^T\,[\,Y-\mu\,(\,W\beta\,)\,]\,\|_\infty \leq \lambda\;+\;\gamma\,\|\,\beta\,\|_1\;\};$$

Where λ and γ are the high-confidence upper bound of $\frac{1}{n}\,\|\,W^T\,[\,Y-\mu\,(\,Z\beta\,)\,]\,\|_\infty$ and $\frac{1}{n}\,\|\,W^T\,\mu'\,(\,\xi\,)\,\Xi\,\|_\infty$ respectively.

Then alike HCS, we select the member in $C(\lambda,\gamma)$ with minimal l_1 norm:

$$\hat{\beta} = \underset{\beta\,\in\,\mathcal{C}\,(\,\lambda,\gamma\,)}{\arg\min}\;\|\,\beta\,\|_1.$$

We termed this estimator as High Confidence Set Selector with Measurment Error, abbreviated as MHCS.

3.2 Theoretical Properties of MHCS

Analogous to HCS, we extend the study of high confidence set property, prediction error bound and parameter error bound to MHCS. The modified assumptions and notations used in this chapter are listed below.

Assumption and Notation

Assumption (C_1)**.** $(Z_i, Y_i)_{i=1}^n\ \ are\ \ i.i.d.$*, and* $(W_i, Y_i)_{i=1}^n$ *are i.i.d.;*

Assumption (C_2)**.** $W = Z + \Xi$*, and* $E(W) = 0$*.*

Assumption $(\ C_3)$**.** $\|\phi(\cdot)\|_\infty < M_d$*; i.e.,* $\| \ Z\ \|_\infty \leq M_d$*; and* $\| \ W\ \|_\infty \leq M_d$*;*

Assumption $(\ C_4)$**.** $M_d\sqrt{log 2d^2} \sim \mathcal{O}(\sqrt{n})$*;*

Assumption (C_5)**.** *For* $\ \forall a_0 > 0, \exists J < \infty,\ such\ \ that, a_{J-1} = a_0\, 2^J \geq 2\|\beta^*\|_1$*;*

Assumption (C_6)**.** $\| \ \beta^*\ \|_0\ \leq\ s$*;*

Assumption $(\ C_7)$**.** $\delta\ L_n\, \rho_{(\ \Delta,\ \beta^*\ )}\ (\, W,\ Y\ ) \geq\ \kappa \ \|\ \Delta\ \|_2$

Remark. The model assumption of additive measurement error C_2 has been illustrated in Section 3.1, other assumptions are analogous to in Section 2.2.

Notation:

$$\lambda^* \equiv \sqrt{2}\ M_d \sqrt{\frac{\log(2d) + \log n}{n}};$$

$$\gamma^* \equiv M_d^2 \sqrt{\frac{\log(2d^2) + \log n}{2n}},$$

$$\delta_1 \equiv \frac{2M_d}{n};$$

$$\delta_2 \equiv 2\, M_d \sqrt{\frac{2\, log\, 2d}{n}}$$

$$\delta_0 = \delta_1 + \delta_2$$

We have following properties for MHCS:

Theorem 3.1 (Event B)**.**

Under Assumption $C_1 - C_4$, when $\lambda > \lambda^*$, $\gamma > \gamma^*$,

$$P\,[\,\beta^* \in \mathcal{C}_{(\lambda,\gamma)}\,] > 1 - \frac{2}{n}.$$

Theorem 3.2 (Excess Risk)**.**

Under Assumption $C_1 - C_5$, when $\lambda > \lambda^*$, $\gamma > \gamma^*$, with probability at least $1 - \frac{2}{n}$, it holds:

$$\mathcal{E}(\hat{\beta}_{MHCS}) \leq \left(3\,\lambda + 2\,\gamma \,\|\, \beta^* \,\|_1 + \delta_0\right) \|\, \beta^* - \hat{\beta}_{MHCS} \,\|_1 + \delta_0\, a_0\,.$$

Theorem 3.3 (Parameter Error Bound)**.**

Under Assumption $C_1 - C_7$, when $\lambda \geq \lambda^*$ and $\gamma \geq \gamma^*$, with probability at least $1 - \dfrac{2}{n}$,

it holds:

$$(i) \quad \|\, \hat{\beta}_{MHCS} - \beta^* \,\|_2 \leq \frac{4\,(\,\lambda + \gamma \,\|\, \beta^* \,\|_1\,)\sqrt{s}}{\kappa};$$

$$(ii) \quad \|\, \hat{\beta}_{MHCS} - \beta^* \,\|_1 \leq \frac{8\,(\,\lambda + \gamma \,\|\, \beta^* \,\|_1\,)\, s}{\kappa}.$$

Corollary. *Under Assumption $C_1 - C_7$, when $\lambda > \lambda^*$, $\gamma > \gamma^*$, with probability at least $1 - 2J\, e^{-2n} - \frac{2}{n}$, it holds that:*

$$\mathcal{E}(\hat{\beta}_{MHCS}) \leq \frac{8s(\,\lambda + \gamma\|\beta^*\|_1\,)\,(3\,\lambda\, + 2\gamma\|\beta^*\|_1 + \delta_0)}{\kappa} + \,\delta_0\, a_0$$

CHAPTER 4: NUMERICAL STUDY OF HIGH CONFIDENCE SET ESTIMATION

4.1 Implementation

We propose an algorithm utilize Newton-Raphson method to solve this optimization problem, which involves in a sequence of non-convexity approximations to the high confidence set. In the following we introduce the main idea.

Notice that simple algebra leads to:

$$L_n'(Z,Y,\beta) \equiv \nabla_\beta \, L_n \, \rho_\beta(Z,Y) \; = n^{-1} \sum_{i=1}^{n} \Big\{ -Y_i Z_i + \frac{\hat{Z}_i \exp(\beta^T Z_i)}{1+\exp(\beta^T Z_i)} \Big\}$$

and

$$L_n''(Z,Y,\beta) \equiv \frac{\partial^2 L_n \rho_\beta(Z,Y)}{\partial \beta^2} = n^{-1} \sum_{i=1}^{n} \frac{\hat{Z}_i \hat{Z}_i^T \exp(\beta^T Z_i)}{\{1+\exp(\beta^T Z_i)\}^2}.$$

Given an initial value $\hat{\beta}^{(0)}$, by Taylor's expansion, we have

$$L_n'(Z,Y,\beta) \approx L_n'(Z,Y,\hat{\beta}^{(0)}) + L_n''(Z,Y,\hat{\beta}^{(0)})(\beta - \hat{\beta}^{(0)}) \equiv \delta_0 + \Sigma_0 \beta,$$

where $\delta_0 = L_n'(Z,Y,\hat{\beta}^{(0)}) - L_n''(Z,Y,\hat{\beta}^{(0)})\hat{\beta}^{(0)}$ and $\Sigma_0 = L_n''(Z,Y,\hat{\beta}^{(0)})$. Then $\mathcal{C}(\lambda)$ can be approximated by

$$\mathcal{C}(\lambda;\hat{\beta}^{(0)}) = \{\beta \in \mathbb{R}^d : \|\delta_0 + \Sigma_0 \beta\|_\infty \leq \lambda\}.$$

Then we obtain the one-step approximation to $\hat{\beta}(\lambda)$:

$$\hat{\beta}^{(1)} = \arg\min_{\beta} \Big\{ \|\beta\|_1 : \beta \in \mathcal{C}(\lambda;\hat{\beta}^{(0)}) \Big\}. \tag{8}$$

Using the above estimator as an updated initial value, we obtain a two-step approximation. Repeat this procedure up to convergence.

The remaining problem is to solve optimization problem (8). This requires solving a non-convex program which can be written as the following form:

$$\begin{aligned}
&\min \mathbf{1}_p^T(\beta^+ + \beta^-)\\
&s.t.\ \Sigma_0\beta^+ - \Sigma_0\beta^- + \delta \le \lambda\\
&\Sigma_0\beta^+ - \Sigma_0\beta^- + \delta \ge -\lambda\\
&\beta^+, \beta^- \ge 0\\
&\beta_j^+\beta_j^- = 0,\ \text{for } j = 1, \ldots, p.
\end{aligned}$$

The convex relaxation of this problem can be obtained by dropping the final constraint. Furthermore, the relaxed problem is a linear program with $2d$ variables and $4d$ constraints. This linear program can be solved very efficiently using a large set of methods such as interior point method or the dual simplex method.

It's straightforward to extend this algorithm to the case of MHCS as follow:

$$L_n'(W, Y, \beta) \equiv n^{-1}\sum_{i=1}^{n}\Big\{-Y_iW_i + \frac{W_i\exp(\beta^TW_i)}{1+\exp(\beta^TW_i)}\Big\}$$

and

$$L_n''(W, Y, \beta) = n^{-1}\sum_{i=1}^{n}\frac{W_i\hat{Z}_i^T\exp(\beta^TW_i)}{\{1+\exp(\beta^TW_i)\}^2}.$$

Given an initial value $\hat{\beta}^{(0)}$, by Taylor's expansion, we have

$$L_n'(W, Y, \beta) \approx L_n'(W, Y, \hat{\beta}^{(0)}) + L_n''(W, Y, \hat{\beta}^{(0)})(\beta - \hat{\beta}^{(0)}) \equiv \delta_0 + \Sigma_0\beta,$$

where $\delta_0 = L'_n(W, Y, \hat{\beta}^{(0)}) - L''_n(Z, Y, \hat{\beta}^{(0)})\hat{\beta}^{(0)}$ and $\Sigma_0 = L''_n(Z, Y, \hat{\beta}^{(0)})$. Then $\mathcal{C}(\lambda, \gamma)$ can be approximated by

$$\mathcal{C}(\lambda, \gamma; \hat{\beta}^{(0)}) = \{\beta \in \mathbb{R}^d : \|\delta_0 + \Sigma_0\beta\|_\infty \leq \lambda + \gamma\|\beta\|_1\}.$$

Then we obtain the one-step approximation to $\hat{\beta}^{(1)}$ for next implementation:

$$\hat{\beta}^{(1)} = \arg\min_{\beta}\Big\{\|\beta\|_1 : \beta \in \mathcal{C}(\lambda, \gamma; \hat{\beta}^{(0)})\Big\}. \tag{9}$$

Using the above estimator as an updated initial value, we obtain a two-step approximation. Repeat this procedure up to convergence.

The remaining problem is to solve optimization problem (9). This requires solving a non-convex program which can be written as the following form:

$$\begin{aligned}
&\min \mathbf{1}_p^T(\beta^+ + \beta^-)\\
&s.t.\ (\Sigma_0 - \gamma)\beta^+ - (\Sigma_0 + \gamma)\beta^- + \delta \leq \lambda\\
&(\Sigma_0 + \gamma)\beta^+ - (\Sigma_0 - \gamma)\beta^- + \delta \geq -\lambda\\
&\beta^+, \beta^- \geq 0\\
&\beta_j^+\beta_j^- = 0,\ \text{for } j = 1, \ldots, d.
\end{aligned}$$

4.2 Simulation Experiment

In this section, we conduct simulation experiments to investigate prediction error and parameter error of proposed methods (HCS, MHCS). Specifically, follow [2], we evaluate the performance of our classifier in scopes of following measures, and compared the results with competitive l_1, l_2 regularization approach, i.e., LASSO and Ridge. The performance measures are:

1. CE: Classification Error;

2. $Deviance$: Cross Entropy:

$$Deviance = -\frac{1}{n}\sum_{i}^{n}[y_i log(\hat{y_i}) + (1-y_i)log(1-\hat{y_i})];$$

where $\hat{y} = 1/(1+\exp(-x\hat{\beta}))$;

3. L_1: l_1 norm of the difference between standardized $\hat{\beta}$ and β^*;

$$L_1 = \left\| \frac{\hat{\beta}}{\|\hat{\beta}\|_2} - \frac{\beta^*}{\|\beta^*\|_2} \right\|_1;$$

4. L_2: l_2 norm of the difference between standardized $\hat{\beta}$ and β^*;

$$L_2 = \left\| \frac{\hat{\beta}}{\|\hat{\beta}\|_2} - \frac{\beta^*}{\|\beta^*\|_2} \right\|_2;$$

5. FN: False Negative Ratio, i.e., the number of zero coefficient of $\hat{\beta}$ for which β^* is non-zero

$$FN = \frac{s_0 - \|\hat{\beta}_J\|_0}{s_0}.$$

6. FP: False Positive Ratio, i.e., the number of non-zero coefficient of $\hat{\beta}$ for which β^* is zero

$$FP = \frac{\|\hat{\beta}_{J^c} - \beta^*_{J^c}\|_0}{p - s_0}$$

Binomial distributed sample data set are generated as follow:

$$
\begin{aligned}
&GenerateX \sim MultiGaussian(\mathbf{0}^d, \mathbf{\Sigma}) \\
&\beta^* = [1^{s_0}, 0^{d-s_0}]^T, \ ; \\
&Pr = \frac{1}{1+e^{-\mathbf{X}\beta^*}} \\
&Y = Binomial(n, 1, Pr)
\end{aligned} \tag{10}
$$

In our experiment setting, dimension $d = 200$; sparsity parameter $s_0 = 10$; training sample size $n_{training} = 100$; and testing sample size $n_{testing} = 100$;

Three types of correlation matrix are taken into account:

Type 1: Identity Matrix: $\Sigma_{d\times d} = diag(d)$;

Type 2: Equal Correlation Matrix: $\Sigma : \Sigma_{i,j} = \rho^{1\{i\neq j\}}$;

Type 3: Toeplitz Matrix: $\Sigma : \Sigma_{i,j} = \rho^{|i-j|}$;

For each type of correlation matrix, we consider following measurement error scenarios:

Scenario 1. Missing Value: which randomly replace a certain proportion (10%, 30%, 50%) of data entries with 0;

Scenario 2. Perturbation: standard Gaussian noise are randomly added to a certain proportion(10%, 30%, 50%) of original data.

We denote the modified training dataset as W_{train}, testing dataset as W_{test}, and the original training dataset as Z_{train}, testing dataset as Z_{test}. In measurement error experiments, classifiers are trained on (W_{train}, Y_{train}) and performance measure (1-6) are tested on (W_{test}, Y_{test}) and (Z_{test}, Y_{test}) respectively. The corresponding

Classification Error and Deviance measures are denoted as $CE(Z_{test})$, $CE(W_{test})$, $Deviance(Z_{test})$, $Deviance(W_{test})$ in result table.

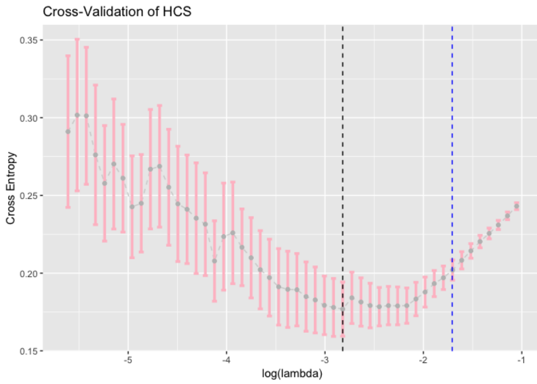
For regularization parameter selection, we sample tuning parameter from grid, and conduct 5-fold cross validation on training set to select tuning parameter. The effect of regularization parameters on β and cross validation will be illustrated in following Experiment.

Experiment 1: Regularization Approach on different level of Perturbation

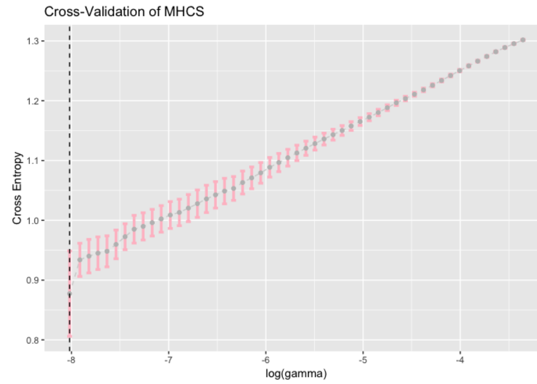
Follow the process of general simulation setup, we generate Type 1 data with different level of perturbation, (10%, 30%, 50%). Figure 3 summarized the 5-fold cross validation error varying with tuning parameter from 0% to 30% of perturbation error. Graphs in left column illustrate cross validation error of HCS varying along with λ. The black dash reference line on left column indicates the optimal λ, denoted as λ^*, which minimize the cross-validation error. The blue reference line denotes λ^* plus standard error. For MHCS tuning, we fixed the $\lambda = \lambda^*$, where λ^* is attained from HCS cross-validation, then conduct 5-fold cross validation on γ grid. The black dash reference line on right column indicates the optimal $\gamma = \gamma^*$ which minimize the cross-validation error. Figure 3 presents that, as perturbation level increases, the reference line of λ^* and γ^* slide to right. In right column, in order to illustrate how cross validation error and tuning parameter γ differs as measurement error increases, graphs (b), (d), (f), (h) starts with ($\lambda = \lambda^*$, $\gamma = 0$), which is the solution to HCS, the corresponding cross validation error is plotted at the most beginning of x-axis (e^{-7}) instead of $\gamma = 0$, since the x-axis is log scale. From (b), (d) in Figure 3 it's seen that, for data without measurement error or with low perturbation level (10%), $\gamma^* = 0$, which implies tuning parameter λ is capable to capture the residual error to some extent. However, as the measurement error aggravates in (f) and (g), γ^* increases in response. This result strongly supports our theory in chapter 3.

In Figure 4, we trace β route varying with regularization parameters, where the colored lines indicate β_j for $j \in S_0$ ($S_0 = \{j : \beta_j^* \neq 0\}$), while for $j \in S_0^c$, β_j lines in light grey. In our experiment, only first ten elements are colored. The figures on left column trace β route move along with λ. Black dash line denotes the position where λ^* is. The figures on right column trace β route regard to γ tuning process with fixed λ^*. It's seen that as the regularization parameter (λ, γ) increase, the magnitude of all β_j decay. As the perturbation level increases, the route of β trumbles further along λ.

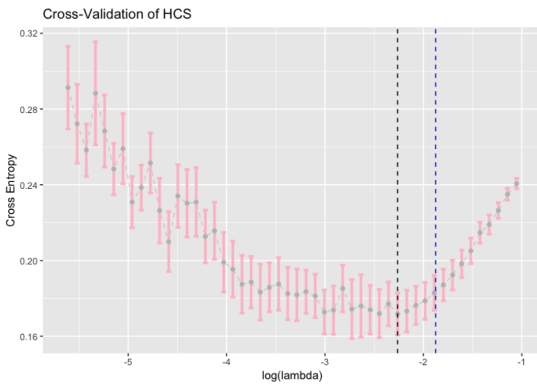
For right column graphs, black dash reference line denotes the $(\lambda = \lambda^*, \gamma = 0)$, while blue dash line denotes the position of γ^* at each perturbation level. The figures show both HCS and MHCS demonstrate the capability in feature selection. However, MHCS selects less features in a more critical way.



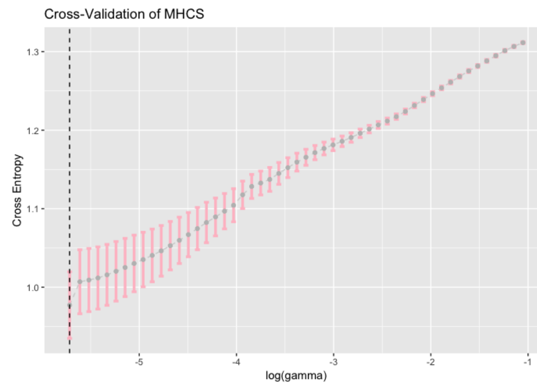
(a) Without measurement error, Type 1



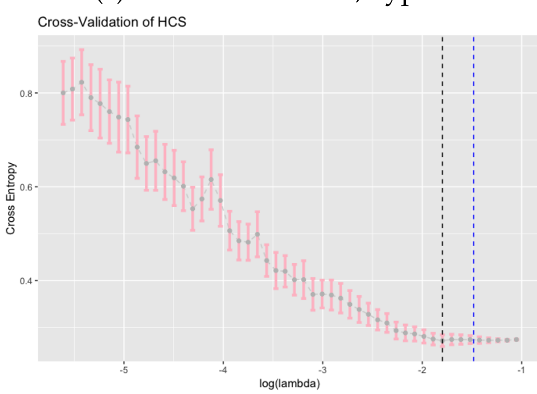
(b) Without measurement error, Type 1



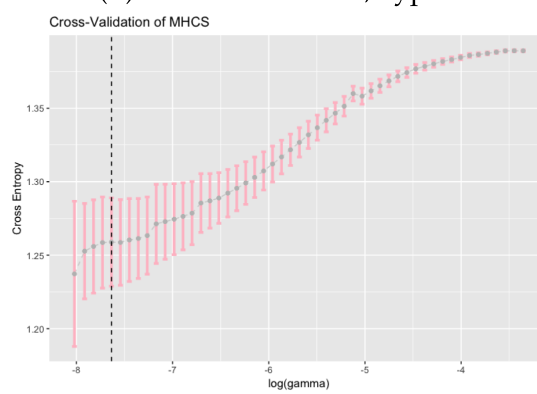
(c) 10% Perturbation, Type 1



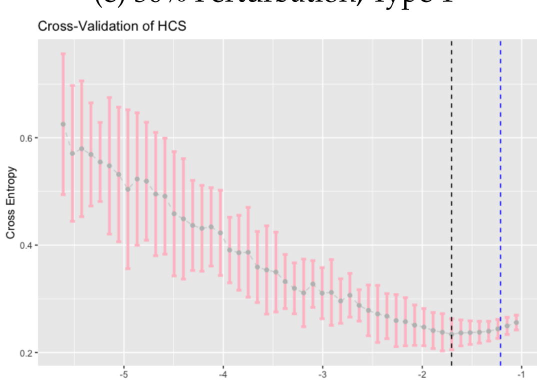
(d) 10% Perturbation, Type 1



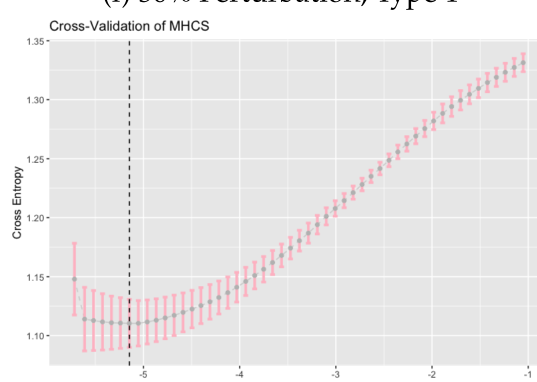
(e) 30% Perturbation, Type 1



(f) 30% Perturbation, Type 1



(g) 50% Perturbation, Type 1



(h) 50% Perturbation, Type 1

Figure 3: Tuning Parameter Selection Illustration: Cross Validation Error with different level of Perturbation, Type 1

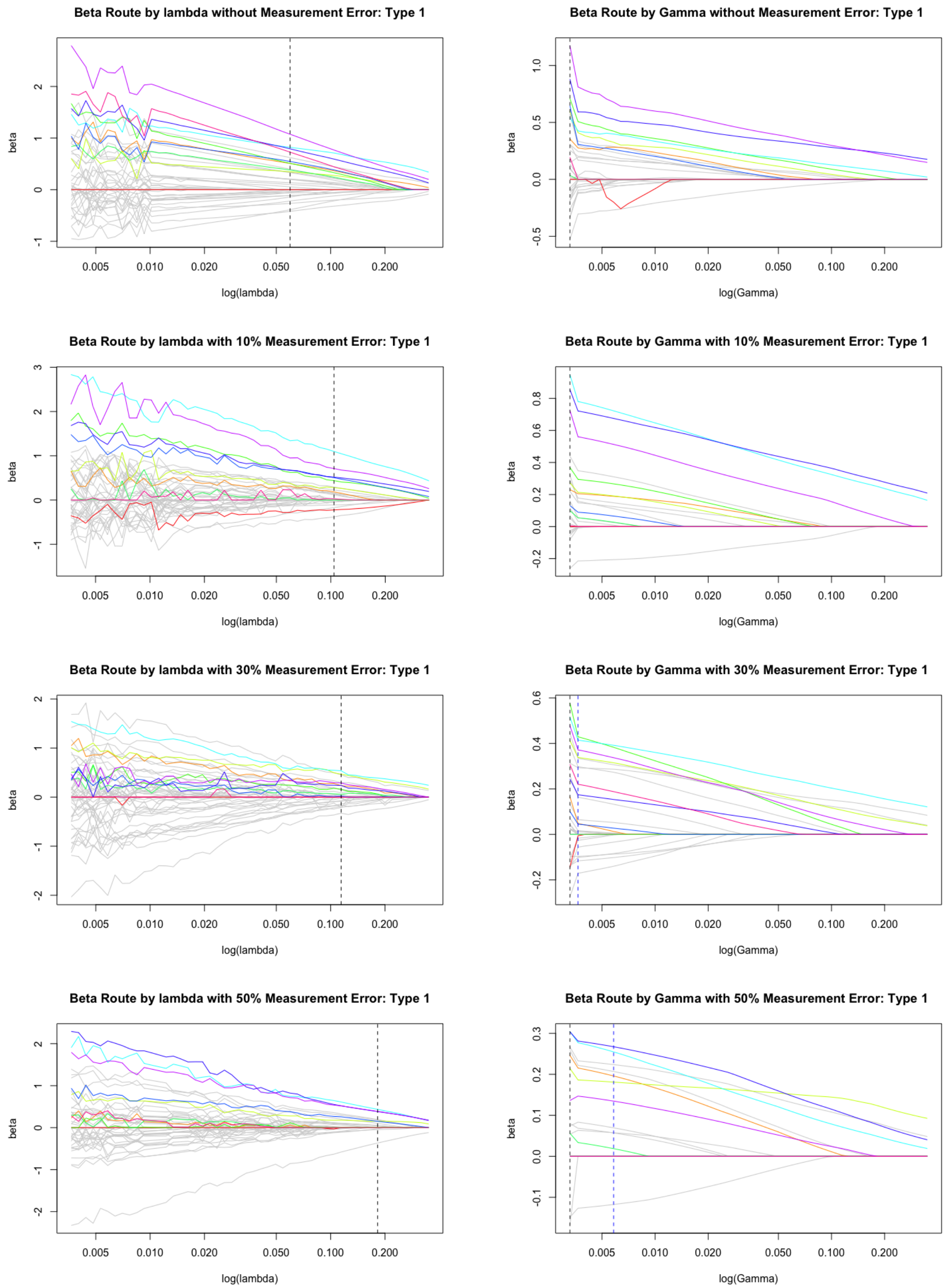


Figure 4: β route with different level of Perturbation

Experiment 2: Type 1 data in different scenarios:

In this experiment, we compare performance (1-6) of four regularized classifiers on the dataset generated from Type 1 correlation matrix, i.e., Identity correlation matrix. Table 1- Table 7 summarized the result of 7 different scenarios respectively. From Table 1 , 2 and 5, where no measurement error or mild measurement error(10%) exists, LASSO, HCS, MHCS perform comparably in terms of prediction error (CE, Deviance) and parameter error (L1 and L2).

Ridge regression has fair capacity in capture the classification error, however, it is not designed for sparse setting, which lead to large l1 norm and failed to conduct feature selection. With respect to features selection, LASSO tends to reduce the False Positive number at the cost of bringing up False Negative number; while HCS acts on opposite, MHCS play a moderate role in between. As the measurement error aggravates, which shown in Table 3, Table 4 Table 6, Table 7, all the performance measures worsen to some degree. However, it's seen that MHCS performs relatively more robust against measurement error than other classifiers. To see this, the margins of performance measure (CE, Deviance) between MHCS and LASSO, MHCS and HCS increase while measurement error rises.

Table 1: Result without Measurement Error, Type 1

	LASSO		RIDGE		HCS		MHCS	
CE	0.34	(0.05)	0.38	(0.03)	0.32	(0.04)	0.31	(0.05)
$Deviance$	1.24	(0.13)	1.33	(0.02)	1.34	(0.24)	1.24	(0.2)
L_1	3.45	(0.55)	11.12	(0.36)	4.52	(0.63)	4.14	(0.58)
$L2$	0.83	(0.2)	1.09	(0.08)	0.75	(0.17)	0.75	(0.17)
FN	0.3	(0.19)	0	(0)	0.13	(0.13)	0.14	(0.13)
FP	0.08	(0.04)	1	(0)	0.2	(0.03)	0.17	(0.03)

Table 2: Result of 10% Missing Value, Type 1

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.35	(0.06)	0.4	(0.03)	0.34	(0.06)	0.34	(0.04)
$Deviance(Z_{test})$	1.25	(0.1)	1.34	(0.02)	1.44	(0.22)	1.32	(0.18)
$CE(W_{test})$	0.37	(0.05)	0.41	(0.05)	0.34	(0.06)	0.34	(0.05)
$Deviance(W_{test})$	1.29	(0.11)	1.34	(0.02)	1.44	(0.23)	1.33	(0.19)
L_1	3.68	(0.53)	11.33	(0.33)	4.87	(0.75)	4.57	(0.65)
$L2$	0.88	(0.14)	1.13	(0.08)	0.84	(0.2)	0.84	(0.2)
FP	0.34	(0.11)	0	(0)	0.18	(0.15)	0.19	(0.14)
FN	0.08	(0.04)	1	(0)	0.22	(0.03)	0.18	(0.02)

Table 3: Result of 30% Missing Value, Type 1

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.37	(0.07)	0.4	(0.05)	0.37	(0.05)	0.35	(0.05)
$Deviance(Z_{test})$	1.38	(0.1)	1.33	(0.03)	1.85	(0.36)	1.42	(0.15)
$CE(W_{test})$	0.45	(0.05)	0.41	(0.04)	0.41	(0.04)	0.41	(0.04)
$Deviance(W_{test})$	1.43	(0.11)	1.35	(0.02)	1.81	(0.23)	1.43	(0.1)
L_1	4.25	(0.71)	11.72	(0.46)	4.37	(0.73)	4.25	(0.62)
$L2$	1.18	(0.19)	1.23	(0.09)	1.21	(0.2)	1.21	(0.2)
FP	0.59	(0.23)	0	(0)	0.29	(0.15)	0.35	(0.12)
FN	0.07	(0.05)	1	(0)	0.25	(0.02)	0.16	(0.02)

Table 4: Result of 50% Missing Value, Type 1

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.41	(0.05)	0.41	(0.04)	0.40	(0.06)	0.37	(0.05)
$Deviance(Z_{test})$	1.35	(0.08)	1.35	(0.03)	1.94	(0.36)	1.42	(0.18)
$CE(W_{test})$	0.42	(0.05)	0.44	(0.04)	0.42	(0.04)	0.41	(0.04)
$Deviance(W_{test})$	1.43	(0.16)	1.36	(0.02)	1.76	(0.23)	1.39	(0.08)
L_1	4.55	(0.62)	10.36	(3.81)	5.57	(0.43)	4.33	(0.52)
$L2$	1.22	(0.25)	1.25	(0.16)	1.21	(0.16)	1.21	(0.16)
FP	0.62	(0.2)	0	(0)	0.31	(0.1)	0.37	(0.07)
FN	0.05	(0.05)	1	(0)	0.27	(0.02)	0.16	(0.01)

Table 5: Result of 10% Measurement Error, Type 1

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.34	(0.04)	0.39	(0.03)	0.35	(0.05)	0.34	(0.04)
$Deviance(Z_{test})$	1.28	(0.11)	1.34	(0.02)	1.44	(0.2)	1.28	(0.14)
$CE(W_{test})$	0.36	(0.03)	0.38	(0.03)	0.36	(0.04)	0.36	(0.05)
$Deviance(W_{test})$	1.31	(0.12)	1.34	(0.02)	1.51	(0.19)	1.31	(0.14)
L_1	3.73	(0.74)	11.27	(0.46)	4.93	(0.89)	3.98	(0.73)
$L2$	0.88	(0.22)	1.12	(0.1)	0.87	(0.23)	0.87	(0.23)
FP	0.29	(0.17)	0	(0)	0.18	(0.15)	0.25	(0.12)
FN	0.09	(0.04)	1	(0)	0.21	(0.03)	0.12	(0.03)

Table 6: Result of 30% Measurement Error, Type 1

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.38	(0.06)	0.38	(0.04)	0.37	(0.06)	0.35	(0.06)
$Deviance(Z_{test})$	1.3	(0.08)	1.34	(0.03)	1.43	(0.18)	1.26	(0.12)
$CE(W_{test})$	0.38	(0.07)	0.41	(0.05)	0.38	(0.05)	0.36	(0.05)
$Deviance(W_{test})$	1.36	(0.11)	1.35	(0.02)	1.64	(0.28)	1.35	(0.17)
L_1	4.01	(0.71)	11.58	(0.34)	5.36	(0.53)	4.48	(0.52)
$L2$	1.01	(0.23)	1.2	(0.08)	0.95	(0.19)	0.95	(0.19)
FP	0.41	(0.25)	0	(0)	0.23	(0.11)	0.27	(0.08)
FN	0.09	(0.07)	1	(0)	0.23	(0.03)	0.14	(0.02)

Table 7: Result of 50% Measurement Error, Type 1

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.40	(0.05)	0.4	(0.05)	0.40	(0.05)	0.37	(0.06)
$Deviance(Z_{test})$	1.39	(0.18)	1.35	(0.03)	1.56	(0.2)	1.35	(0.11)
$CE(W_{test})$	0.42	(0.04)	0.43	(0.04)	0.41	(0.03)	0.42	(0.03)
$Deviance(W_{test})$	1.5	(0.21)	1.36	(0.03)	1.85	(0.2)	1.47	(0.12)
L_1	4.56	(0.75)	10.83	(2.73)	5.92	(0.67)	4.91	(0.62)
$L2$	1.08	(0.21)	1.19	(0.1)	1.14	(0.15)	1.14	(0.15)
FP	0.48	(0.25)	0	(0)	0.33	(0.13)	0.38	(0.17)
FN	0.12	(0.07)	1	(0)	0.23	(0.03)	0.14	(0.03)

Experiment 3: Type 2 data with different scenarios

In this experiment, we compare performance measures (1-6) of four regularized classifiers on the dataset generated from Type 2 correlation matrix, i.e., equal correlation matrix with $\rho = 0.5$. Table 8-Table 14 summarized results of 7 different scenarios respectively.

The result presents that classification error (i.e., $CE(Z_{test})$, Deviance) from all four classifiers are close to each other among all scenarios. In terms of classification error (CE), Ridge regression edges out other classifiers with a small lead, however, the performance of L1 and feature selection (FN, PN) in this dataset failed to exceed l_1 regularization. With respect to feature selection, the performance of LASSO, HCS, MHCS are consistently close to each other in every setting. The corresponding results exhibit that, False Negative ratio among all the l_1 regularized classifiers (LASSO, HCS, MHCS) exceeds 50%, and False Positive is relatively high compare to Type 1 and Type 3 dataset, each of which goes beyond 11%. As the measurement error levels up, parameter error (L1, L2) and feature selection measure (FP, FN) worsen to some extent, however, the classification risk appears consistently drift around 11% to 13%. The reason is Type 2 data is generated with equal correlation matrix, which all features are correlated to each other. With this inherent structure, the l_1 based classifier is more robust with respect to prediction error as measurement error increases, though pays the price of increment of parameter error.

Table 8: Result without Measurement Error, Type 2

	LASSO		RIDGE		HCS		MHCS	
CE	0.12	(0.04)	0.11	(0.03)	0.12	(0.04)	0.12	(0.03)
$Deviance$	0.56	(0.08)	0.64	(0.05)	0.56	(0.12)	0.56	(0.1)
L_1	5.08	(0.54)	14.23	(0.45)	5.5	(0.3)	5.43	(0.31)
$L2$	1.31	(0.17)	1.41	(0.05)	1.4	(0.15)	1.4	(0.15)
FP	0.56	(0.08)	0	(0)	0.53	(0.12)	0.55	(0.13)
FN	0.1	(0.02)	1	(0)	0.12	(0.02)	0.11	(0.02)

Table 9: Result of 10% Missing Value, Type 2

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.12	(0.04)	0.11	(0.03)	0.12	(0.04)	0.12	(0.03)
$Deviance(Z_{test})$	0.56	(0.12)	0.62	(0.05)	0.59	(0.14)	0.58	(0.13)
$CE(W_{test})$	0.13	(0.04)	0.11	(0.03)	0.13	(0.04)	0.12	(0.03)
$Deviance(W_{test})$	0.59	(0.15)	0.65	(0.05)	0.59	(0.14)	0.59	(0.13)
L_1	5.25	(0.42)	14.24	(0.44)	5.49	(0.67)	5.42	(0.58)
$L2$	1.38	(0.18)	1.42	(0.06)	1.41	(0.23)	1.41	(0.23)
FP	0.57	(0.11)	0	(0)	0.56	(0.12)	0.57	(0.08)
FN	0.11	(0.02)	1	(0)	0.11	(0.02)	0.11	(0.02)

Table 10: Result of 30% Missing Value, Type 2

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.12	(0.03)	0.11	(0.03)	0.11	(0.03)	0.11	(0.03)
$Deviance(Z_{test})$	0.54	(0.12)	0.57	(0.06)	0.57	(0.14)	0.57	(0.13)
$CE(W_{test})$	0.13	(0.04)	0.11	(0.03)	0.13	(0.04)	0.12	(0.03)
$Deviance(W_{test})$	0.59	(0.16)	0.67	(0.05)	0.59	(0.12)	0.58	(0.1)
L_1	5.66	(0.5)	14.3	(0.41)	5.82	(0.37)	5.88	(0.44)
$L2$	1.43	(0.16)	1.44	(0.05)	1.45	(0.13)	1.45	(0.13)
FP	0.53	(0.09)	0	(0)	0.52	(0.1)	0.6	(0.08)
FN	0.12	(0.01)	1	(0)	0.13	(0.02)	0.13	(0.02)

Table 11: Result of 50% Missing Value, Type 2

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.13	(0.04)	0.11	(0.03)	0.12	(0.03)	0.12	(0.04)
$Deviance(Z_{test})$	0.59	(0.18)	0.52	(0.07)	0.57	(0.24)	0.59	(0.23)
$CE(W_{test})$	0.15	(0.05)	0.12	(0.04)	0.15	(0.04)	0.16	(0.02)
$Deviance(W_{test})$	0.74	(0.23)	0.7	(0.04)	0.66	(0.15)	0.67	(0.12)
L_1	6.18	(0.81)	14.28	(0.39)	6.11	(0.58)	6.35	(0.66)
$L2$	1.57	(0.25)	1.48	(0.06)	1.56	(0.22)	1.56	(0.22)
FP	0.66	(0.13)	0	(0)	0.62	(0.14)	0.65	(0.16)
FN	0.15	(0.02)	1	(0)	0.14	(0.02)	0.14	(0.02)

Table 12: Result of 10% Measurement Error, Type 2

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.12	(0.03)	0.11	(0.03)	0.13	(0.03)	0.13	(0.03)
$Deviance(Z_{test})$	0.57	(0.1)	0.65	(0.05)	0.56	(0.12)	0.56	(0.12)
$CE(W_{test})$	0.12	(0.04)	0.11	(0.03)	0.12	(0.03)	0.12	(0.03)
$Deviance(W_{test})$	0.59	(0.11)	0.65	(0.05)	0.58	(0.13)	0.58	(0.13)
L_1	5.38	(0.5)	14.19	(0.41)	5.78	(0.54)	5.74	(0.57)
$L2$	1.43	(0.22)	1.42	(0.05)	1.46	(0.21)	1.46	(0.21)
FP	0.63	(0.17)	0	(0)	0.61	(0.15)	0.63	(0.13)
FN	0.11	(0.01)	1	(0)	0.13	(0.01)	0.13	(0.02)

Table 13: Result of 30% Measurement Error, Type 2

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.12	(0.02)	0.1	(0.03)	0.12	(0.02)	0.12	(0.03)
$Deviance(Z_{test})$	0.58	(0.1)	0.66	(0.05)	0.59	(0.11)	0.57	(0.12)
$CE(W_{test})$	0.14	(0.03)	0.11	(0.03)	0.15	(0.02)	0.14	(0.02)
$Deviance(W_{test})$	0.62	(0.1)	0.67	(0.05)	0.65	(0.13)	0.63	(0.13)
L_1	5.8	(0.94)	14.26	(0.39)	6.21	(0.76)	6.07	(0.76)
$L2$	1.54	(0.27)	1.45	(0.06)	1.6	(0.25)	1.6	(0.25)
FP	0.61	(0.14)	0	(0)	0.64	(0.2)	0.63	(0.19)
FN	0.11	(0.02)	1	(0)	0.13	(0.02)	0.13	(0.02)

Table 14: Result of 50% Measurement Error, Type 2

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.13	(0.03)	0.11	(0.03)	0.13	(0.04)	0.12	(0.02)
$Deviance(Z_{test})$	0.59	(0.07)	0.68	(0.04)	0.56	(0.1)	0.55	(0.09)
$CE(W_{test})$	0.15	(0.03)	0.11	(0.03)	0.15	(0.04)	0.13	(0.03)
$Deviance(W_{test})$	0.67	(0.09)	0.69	(0.04)	0.68	(0.12)	0.66	(0.11)
L_1	6.23	(0.64)	14.21	(0.29)	6.55	(0.76)	6.52	(0.8)
$L2$	1.69	(0.19)	1.46	(0.05)	1.67	(0.2)	1.67	(0.2)
FP	0.7	(0.17)	0	(0)	0.72	(0.13)	0.72	(0.16)
FN	0.12	(0.02)	1	(0)	0.14	(0.02)	0.14	(0.02)

Experiment 4: Type 3 data with different scenarios

In this experiment, we compare performance measures (1-6) of four regularized classifiers on the dataset generated from Type 3 correlation matrix, i.e., Toeplitz correlation matrix with $\rho = 0.5$. Table 15 - Table 21 summarized result of 7 different scenarios respectively.

The result presents that LASSO, HCS and MHCS perform comparably in terms of prediction error (CE, Deviance) in all settings, though MHCS appears to have a small lead when measurement error imposed.

With respect to parameter error, MHCS has lowest L1 error, while LASSO has lowest L2 error. the margin of L1 between MHCS and LASSO decreases while the margin of L2 between MHCS and LASSO increases as the measurement error levels up. As we discuss before, HCS is a specific solution to MHCS where $\gamma = 0$. With the appropriate γ, MHCS is apt to approach solution in the direction of L1 decaying, which dramatically improve the performance of FP with trade off in a small increment in FN, especially in the case of measurement error presents. To see this, compare FN and FP of MHCS with HCS in Table 17, Table 18, where missing value reach 30% and 50% respectively, MHCS amends to reduce the FN by 13% by only bringing up 1% increment to FP. As the measurement error leverages, all the performance margins between MHCS and HCS increase, which suggests that MHCS performs more robust against measurement error.

Table 15: Result without Measurement Error, Type 3

	LASSO		RIDGE		HCS		MHCS	
CE	0.17	(0.04)	0.24	(0.04)	0.18	(0.03)	0.17	(0.04)
$Deviance$	0.78	(0.11)	1.21	(0.03)	0.85	(0.19)	0.79	(0.08)
L_1	2.46	(0.29)	9.09	(0.34)	3.49	(0.27)	2.26	(0.32)
$L2$	0.48	(0.1)	0.67	(0.04)	0.55	(0.07)	0.55	(0.07)
FP	0.13	(0.08)	0	(0)	0.12	(0.1)	0.15	(0.08)
FN	0.06	(0.02)	1	(0)	0.16	(0.03)	0.04	(0.01)

Table 16: Result of 10% Missing Value, Type 3

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.19	(0.03)	0.24	(0.05)	0.2	(0.05)	0.19	(0.04)
$Deviance(Z_{test})$	0.81	(0.11)	1.21	(0.03)	0.93	(0.2)	0.8	(0.08)
$CE(W_{test})$	0.2	(0.03)	0.25	(0.05)	0.21	(0.05)	0.19	(0.04)
$Deviance(W_{test})$	0.84	(0.1)	1.22	(0.03)	0.95	(0.2)	0.84	(0.09)
L_1	2.87	(0.76)	9.38	(0.33)	3.84	(0.47)	2.42	(0.42)
$L2$	0.54	(0.14)	0.72	(0.05)	0.63	(0.12)	0.63	(0.12)
FP	0.16	(0.11)	0	(0)	0.14	(0.11)	0.16	(0.11)
FN	0.09	(0.05)	1	(0)	0.17	(0.02)	0.05	(0.02)

Table 17: Result of 30% Missing Value, Type 3

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.21	(0.03)	0.27	(0.05)	0.23	(0.03)	0.21	(0.04)
$Deviance(Z_{test})$	0.9	(0.1)	1.21	(0.03)	1.22	(0.27)	0.85	(0.1)
$CE(W_{test})$	0.25	(0.04)	0.29	(0.04)	0.29	(0.05)	0.22	(0.05)
$Deviance(W_{test})$	0.99	(0.12)	1.25	(0.03)	1.34	(0.23)	0.97	(0.09)
L_1	2.96	(0.68)	10.17	(0.35)	4.79	(0.58)	2.81	(0.51)
$L2$	0.58	(0.13)	0.84	(0.06)	0.78	(0.14)	0.78	(0.14)
FP	0.17	(0.13)	0	(0)	0.12	(0.06)	0.13	(0.09)
FN	0.08	(0.06)	1	(0)	0.21	(0.03)	0.08	(0.01)

Table 18: Result of 50% Missing Value, Type 3

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.19	(0.04)	0.26	(0.07)	0.23	(0.07)	0.2	(0.05)
$Deviance(Z_{test})$	0.85	(0.12)	1.19	(0.06)	1.35	(0.54)	0.83	(0.16)
$CE(W_{test})$	0.24	(0.05)	0.33	(0.05)	0.3	(0.04)	0.27	(0.05)
$Deviance(W_{test})$	1.02	(0.18)	1.29	(0.03)	1.43	(0.26)	1.07	(0.09)
L_1	3.23	(0.73)	10.61	(0.32)	5.26	(0.72)	3.23	(0.62)
$L2$	0.65	(0.13)	0.96	(0.09)	0.88	(0.17)	0.88	(0.17)
FP	0.2	(0.12)	0	(0)	0.14	(0.08)	0.15	(0.1)
FN	0.09	(0.05)	1	(0)	0.23	(0.02)	0.1	(0.02)

Table 19: Result of 10% Measurement Error, Type 3

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.17	(0.04)	0.23	(0.06)	0.18	(0.04)	0.17	(0.04)
$Deviance(Z_{test})$	0.82	(0.12)	1.22	(0.03)	0.86	(0.19)	0.82	(0.08)
$CE(W_{test})$	0.19	(0.04)	0.25	(0.06)	0.19	(0.04)	0.18	(0.04)
$Deviance(W_{test})$	0.85	(0.12)	1.23	(0.03)	0.95	(0.18)	0.85	(0.08)
L_1	2.67	(0.32)	9.27	(0.29)	3.9	(0.57)	2.47	(0.41)
$L2$	0.53	(0.12)	0.71	(0.04)	0.63	(0.13)	0.63	(0.13)
FP	0.17	(0.08)	0	(0)	0.16	(0.11)	0.18	(0.1)
FN	0.07	(0.02)	1	(0)	0.17	(0.03)	0.05	(0.02)

Table 20: Result of 30% Measurement Error, Type 3

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.2	(0.05)	0.24	(0.06)	0.22	(0.07)	0.19	(0.06)
$Deviance(Z_{test})$	0.89	(0.13)	1.24	(0.03)	1	(0.26)	0.87	(0.12)
$CE(W_{test})$	0.24	(0.04)	0.26	(0.04)	0.26	(0.06)	0.23	(0.04)
$Deviance(W_{test})$	0.94	(0.1)	1.24	(0.03)	1.2	(0.23)	0.93	(0.1)
L_1	2.75	(0.57)	9.78	(0.42)	4.57	(0.81)	2.82	(0.62)
$L2$	0.56	(0.17)	0.79	(0.07)	0.73	(0.18)	0.73	(0.18)
FP	0.19	(0.14)	0	(0)	0.13	(0.09)	0.14	(0.11)
FN	0.06	(0.03)	1	(0)	0.2	(0.03)	0.07	(0.02)

Table 21: Result of 50% Measurement Error, Type 3

	LASSO		RIDGE		HCS		MHCS	
$CE(Z_{test})$	0.19	(0.04)	0.27	(0.04)	0.21	(0.05)	0.2	(0.04)
$Deviance(Z_{test})$	0.91	(0.09)	1.26	(0.02)	0.96	(0.21)	0.91	(0.06)
$CE(W_{test})$	0.25	(0.04)	0.31	(0.06)	0.27	(0.04)	0.24	(0.05)
$Deviance(W_{test})$	1.03	(0.12)	1.27	(0.03)	1.29	(0.2)	1.01	(0.09)
L_1	2.85	(0.59)	10	(0.4)	4.36	(0.46)	2.75	(0.45)
$L2$	0.58	(0.13)	0.82	(0.07)	0.73	(0.14)	0.73	(0.14)
FP	0.17	(0.13)	0	(0)	0.11	(0.07)	0.17	(0.12)
FN	0.08	(0.04)	1	(0)	0.2	(0.03)	0.07	(0.02)

CHAPTER 5: REAL DATA ANALYSIS

Real Data Example 1: Sentiment Analysis of IMDb Movie Review

This example presents the proposed techniques (HCS, MHCS) to perform sentiment analysis in IMDB movie reviews, we compare the results with competing methods: penalized logistic regression(PLR), support vector machine (SVM).

Experiment setup

We download a sample data set developed by [27], the training data set contains 2000 movie reviews from IMDb, where 1000 reviews are labeled as positive (1), and 1000 reviews are labeled as negative (0); the testing data set contains 1000 reviews, with 500 reviews are labeled as negative, and 500 reviews are labeled as positive. The general techniques for text preprocessing include following steps: first remove all the links and punctuations in text, convert all words into lower case, and tokenize the text into a sequence of single word (unigrams).

Bag of Words representation is applied to this example, where each unique word serves as a feature. We also applied a rough dimension reduction technique by simply removing stopword according to NLTK stopword list and dropping features which occur less than 10 times over all samples, which result in 12,932 dimensions in total number of features.

Denote $n_w\{i,j\}$ as the number of occurrences of word j in review i and $n_d\{i\}$

as the total number of words in review i. then we denote the value of feature j in review i as:

$$z_{ij} = \frac{n_w\{i,j\}}{n_d\{i\}}$$

$y_i = 1$ if the review is positive, $y_i = 0$ if the review is negative.

The full data set is randomly separated to 70% for training, the remaining 30% for testing. Then we fit the different classifiers on training set, and record the numbers of non-zero coefficients ($\|\hat{\beta}\|_0$) and testing classification error (CE). The tuning parameter λ in HCS is selected by 5-fold cross validation on training set. For MHCS, λ and γ are sampled from a grid search, then similar to HCS, we select the one produces best result by 5-fold cross validation on training set. This process is repeated for 50 times, and the mean value of $\|\hat{\beta}\|_0$ and CE are summarized in Table 22:

The results shows that among all classifiers, SVM achieves lowest mean classification error, which is 0.1. HCS, MHCS perform comparatively with the mean classification error are 0.12 and 0.11 respectively. Although SVM performs slightly better in terms of mean classification error, from Table 22, it's seen MHCS performs more stable than SVM with standard error 0.01 while for SVM is 0.08.

In the aspect to the capability of feature selection, the proposed classifiers show prominent advantage of l_1 regularization in high dimensional setting. According to the number of non-zero, HCS and MHCS select 82 and 47 features among 12,932 features. while other classifiers do not present the power of feature selection.

Table 22: Performance measures of IMDB movie review

	HCS		MHCS		SVM		PLR	
CE	0.12	(0.02)	0.11	(0.01)	0.1	(0.08)	0.13	(0.03)
$\|\|\hat{\beta}\|\|_0$	82	(1.5)	47	(1.1)	12,932	(0)	12,932	(0)

We exhibit the features with top 10 positive and negative coefficient selected by MHCS in a test trail, according to the result shows in Table 23 , the positive and negative terms demonstrate a close match to human's emotional sentiment.

Table 23: Demo: Top 10 Positive and Negative Features Selected by MHCS

coef	word	coef	word
5.253178e-03	wonderful	4.628492e-03	poor
4.500304e-03	favorite	-2.106760e-03	worst
4.462270e-03	loved	-1.062546e-03	disappointing
4.112702e-03	excellent	-5.603305e-04	terrible
2.744693e-03	amazing	-5.047209e-04	waste
8.060263e-04	worth	-2.728851e-04	awful
7.088666e-04	enjoy	-2.314834e-04	boring
6.042885e-04	perfect	-9.118960e-05	save
4.579394e-04	best	-2.501499e-05	horrible
4.315558e-04	holiday	-5.956561e-06	disappointment

Missing Value Scenarios

In order to investigate the proposed classifiers' capability of dealing with measurement error, we randomly delete a certain proportion of word sequence which generates the original data set Z, denote this new data set as W, then we sample 70% data from W as training set, denote as W_{train}. The training process is the same as previous example. Then we apply the fitted model and tuning parameter selected by 5-fold cross validation to conduct prediction test on remaining testing data W_{test}, we test on Z which has the same index as W_{test} as well, denote as Z_{test}. Repeat this process 50 times, the mean and standard error are recorded in Table 24: the result presents that HCS, MHCS, SVM perform better than PLR over all settings, although the standard error of SVM is higher than PLR. As the missing proportion increases, the performance of all the classifiers worsen to some degree. Nevertheless, the result shows that MHCS performs better than other classifiers, demonstrates its robustness against missing value.

Table 24: Performance measures of IMDb movie review Missing Value Scenario

		HCS		MHCS		SVM		PLR	
		mean	sd	mean	sd	mean	sd	mean	sd
10%	W_{test}	0.12	(0.02)	0.11	(0.02)	0.11	(0.07)	0.15	(0.02)
	Z_{test}	0.13	(0.02)	0.12	(0.02)	0.10	(0.11)	0.14	(0.02)
30%	W_{test}	0.15	(0.03)	0.12	(0.01)	0.14	(0.1)	0.16	(0.02)
	Z_{test}	0.15	(0.02)	0.13	(0.02)	0.13	(0.09)	0.16	(0.01)
50%	W_{test}	0.16	(0.02)	0.15	(0.02)	0.17	(0.1)	0.18	(0.01)
	Z_{test}	0.15	(0.02)	0.14	(0.02)	0.17	(0.7)	0.2	(0.01)

Real Data Example 2: Cat vs. Dog Image Recognition

Image data is remarkably high dimensional and frequently process with noise. In this example, we use a small sample of labeled images of dog and cat from kaggle [26], our aim is to build a classifier automatically distinguish whether images contain either a dog or a cat. The original data set contains 25,000 images of dogs and cats in training fold, and 12,500 images in test folds. In order to demonstrate proposed classifier in $d > n$ setting, we only use a small sample in this example. We use 1,200 images from training fold, 600 are labeled as cat and 600 are labeled as dog, then randomly split the data to 1000 images for training, 200 images for testing. The main idea is input the image data ($3 \times 224 \times 224$) to a pre-trained network (VGG-16 [28]) for feature extraction, then foward the extracted features to the linear classifiers concerned. VGG-16[28] is a CNN network pre-trained on ImageNet data set, its architecture is illustrated in Figure 5. ImageNet[44] is large data set contains

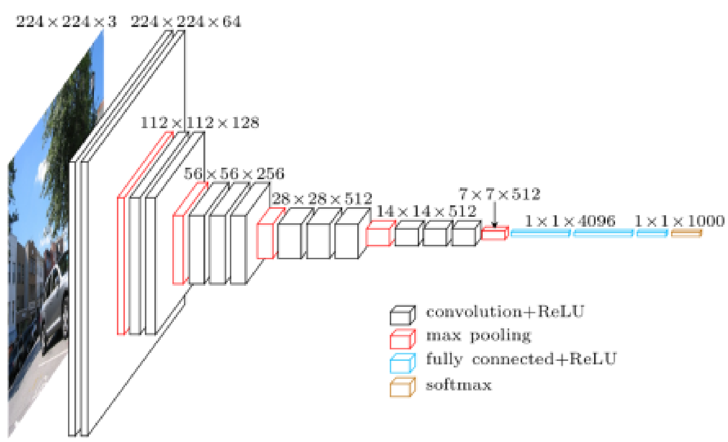


Figure 5: VGG-16 Architature [28]

1.2 M labeled images from 1000 categories.

In image recognition, pre-trained networks demonstrates strong capability to conduct new deep learning task via transfer learning. Besides computationally efficiency due to pre-trained weights, the first few layers of CNN in image recognition training usually capture universal features such as lines, edges, curves that related to other task.

To conduct the feature extraction, we freeze all weights, utilize the entire network as feature extractor, then forward the extracted features to the HCS, MHCS, SVM and PLR.

First block in Table 25 illustrates top 5 feature extracted by VGG-16 with the a sample cat image 'original Murphy' (Figure 6).

Perturbation Scenarios:

In order to investigate proposed classifiers' capability to cope with noise contaminated data, we add Gaussian noise to image data set on purpose. Figure 6 elaborates the effect with different proportion of noise, the corresponding top 5 feature extracted by VGG-16 listed in Table 25.

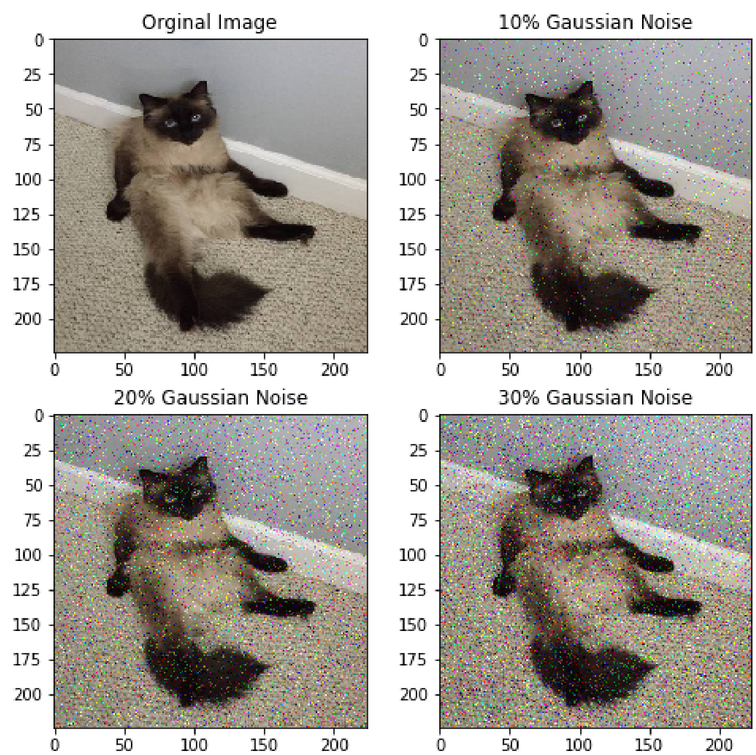


Figure 6: Murphy with different proportion Gaussian Noise

Table 25: Top 5 Features Extracted by VGG-16

Orignal Murphy		10% Noise	
Feature	Value	Feature	Value
Siamese cat	0.998657	Siamese cat	0.972666
paper towel	0.000367	cairn	0.014863
tub	0.000216	West Highland white terrier	0.001727
toilet tissue	0.000196	Scotch terrier	0.001670
lynx	0.000178	paper towel	0.001231
20% Noise		30% Noise	
Feature	Value	Feature	Value
Siamese cat	0.982971	West Highland white terrier	0.505062
cairn	0.004131	Scotch terrier	0.141582
West Highland white terrier	0.002864	cairn	0.134488
Scotch terrier	0.000798	Siamese cat	0.073840
giant panda	0.000571	Norwich terrier	0.037540

In second step, we train the classifiers on extracted features, this process is same as previous examples. Table 26 and Table 27 summarize the result of performance. It's seen that MHCS performs best among four classifiers, with lowest classification error (19%) and smallest number of selected features (17 out of 1,000), HCS also demonstrates the capability of feature selections which the number is 32 out of 1,000, while SVM and PLR selects all the features.

Table 26: Performance measures of Cat vs Dog Image Recognition

	HCS	MHCS	SVM	PLR
CE	0.21	0.19	0.2	0.23
$\|\|\hat{\beta}\|\|_0$	32	17	1,000	1,000

From Table 27, it's shown that HCS and MHCS perform more stable than SVM and PLR as the proportion of perturbation increases. In aspects of prediction error, and robustness against noise, MHCS surpass other classifiers.

Table 27: Performance Measures of Cat vs Dog with Noise

		HCS		MHCS		SVM		PLR	
		Mean	SE	Mean	SE	Mean	SE	Mean	SE
10%	Z_{test}	22.8	0.2	20.5	0.1	22.2	0.3	25.6	0.1
	W_{test}	22.3	0.2	21.4	0.1	23.1	0.3	25.1	0.1
20%	Z_{test}	23.5	0.1	21.2	0.2	23.2	0.3	25.8	0.2
	W_{test}	22.8	0.2	21.9	0.2	23.5	0.3	25.3	0.1
30%	Z_{test}	24.1	0.1	22.4	0.1	25.3	0.3	26.7	0.1
	W_{test}	24.5	0.1	22.7	0.1	25.4	0.2	26.4	0.1

REFERENCES

[1] Ackermann, M. and Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* **10**, 1471–2105.

[2] Barut, E., Bradic, J., Fan, J. and Jiang, J. (2016). High dimensional classification with errors-in-variables using high confidence sets. Manuscript.

[3] Bickel, P. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, "naive Bayes," and some alternatives when there are many more variables than observations. *Bernoulli* **10**, 989–1010.

[4] Bühlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data. Springer.

[5] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics, Vol. 35, 6, 2313–2351.

[6] Cai, T. and Liu, W. (2011). A Direct Estimation Approach to Sparse Linear Discriminant Analysis. *Journal of the American Statistical Association* **106**, 1566–1577.

[7] Larry Wasserman.(2015) Concentration of Measure *Lecture Notes* http://www.stat.cmu.edu/ larry/=sml/Concentration-of-Measure.pdf

[8] Fan, J. (2014). Features of Big Data and sparsest solution in high confidence set. In *Past, Present and Future of Statistical Science* (X, Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, J.-L. Wang, Eds.), 507–523.

[9] Fan, J., and Fan, Y. (2008). High-Dimensional Classification Using Features Annealed Independence Rules *Annals of Statistics* **36**, 2605–2637.

[10] Fan, J., Feng, Y., Jiang, J. and Tong, X. (2016). A Classification Rule of Feature Augmentation via Nonparametrics and Selection (FANS) in High Dimensional Space. *Jour. Amer. Statist. Assoc.*, to appear.

[11] Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Amer. Statist. Assoc.* **106**, 544–557.

[12] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33** 1–22.

[13] Hastie, T., Tibshirani, R. and Friedman, J. H. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition). Springer-Verlag Inc.

[14] Ledoux, M. and Talagrand, M.(1991). Probability in Banach Spaces: Isoperimetry and Processes. Springer, New York.

[15] Sara A. van de Geer (2008) High-dimensional generalized linear models and the lasso Annals of Statistics 2008, Vol. 36, No. 2, 614–645

[16] BÃijhlmann P., van de Geer S. (2011) Generalized linear models and the Lasso. In: Statistics for High-Dimensional Data. Springer Series in Statistics. Springer, Berlin, Heidelberg

[17] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007b). Sparsity oracle inequalities for the Lasso. Electron. J. Statist. 1 169–194.

[18] Y. Zhang, M. Wainwright, and M. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In COLT, pages 921–948, 2014.

[19] S. van de Geer. Estimation and Testing Under Sparsity: Ecole dEt ÌĄe de Probabilities de Saint-Flour XLV-2016. Springer Science Business Media, 2016.

[20] R. Tibshirani. Regression analysis and selection via the Lasso. Journal of the Royal Statistical Society Series B, 58:267–288, 1996.

[21] V. Koltchinskii, K. Lounici, and A.B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Annals of Statistics, 39(5):2302–2329, 2011.

[22] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007a). Sparse density estimation with l1 penalties. COLT 2007 4539 530–543.

[23] C. McDiarmid. On the method of bounded differences. In Surveys in Combinatorics, pages 148âĂŞ188. Cambridge University Press, 1989.

[24] Massart, P. (2000). About the constants in TalagrandâĂŹs concentration inequalities for empirical processes. Ann. Probab. 28 863âĂŞ884.

[25] Meier, L., Geer, V., and Bühlmann, P. (2009). High-Dimensional Additive Modeling. *Annals of Statistics* **37**, 3779–3821.

[26] https://www.kaggle.com/c/dogs-vs-cats

[27] Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng, Andrew Y. and Potts, Christopher(2011) Learning Word Vectors for Sentiment Analysis **1**42–150

[28] Karen Simonyan and Andrew Zisserman(2014) Very Deep Convolutional Networks for Large-Scale Image Recognition

1409–1556,

[29] Sahand Negahban and Martin J. Wainwright(2010) Restricted strong convexity and weighted matrix completion: Optimal bounds with noise

1009–2118

[30] Vanderbei, R. J. (2013). Linear Programming: Foundations and Extensions. Springer:New York.

[31] Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). Measurement Error in Nonlinear Models. Boca Raton, Florida, USA: Chapman & Hall/CRC.

[32] Chen, Y. and C. Caramanis (2013). Noisy and missing data regression: Distribution-oblivious support recovery. In JMLR Workshop and Conference Proceedings, Volume 28, pp. 381–391.

[33] James, G. M. and P. Radchenko (2009). A generalized Dantzig selector with shrinkage tuning. Biometrika 96(2), 323–337.

[34] Liang, H. and R. Li (2009). Variable selection for partially linear models with measurement errors. Journal of the American Statistical Associa- tion 104(485), 234–248.

[35] Loh, P.-L. and M. J. Wainwright (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. Annals of Statistics 40(3), 1637–1664.

[36] Ma, Y. and R. Li (2010). Variable selection in measurement error models. Bernoulli 16(1), 274–300.

[37] McCullagh, P. and J. Nelder (1989). Generalized Linear Models. Boca Raton, Florida, USA: Chapman & Hall / CRC.

[38] Meinshausen, N. and P. BuÌĹhlmann (2006). High-dimensional graphs and vari- able selection with the lasso. Annals of Statistics 34(3), pp. 1436–1462.

[39] Nguyen, N. and T. Tran (2013, April). Robust lasso with missing and grossly corrupted observations. Information Theory, IEEE Transactions on 59(4), 2036–2058.

[40] Rosenbaum, M. and A. B. Tsybakov (2010). Sparse recovery under matrix uncertainty. Annals of Statistics 38(5), 2620–2651.

[41] Rosenbaum, M. and A. B. Tsybakov (2013). Improved matrix uncertainty selector. In From Probability to Statistics and Back: High-Dimensional Models and Processes âĂŞ A Festschrift in Honor of Jon A. Wellner, pp. 276 –290. Beachwood, Ohio, USA: Institute of Mathematical Statistics.

[42] Zhu, H., G. Leus, and G. B. Giannakis (2011). Sparsity-cognizant total least squares for perturbed compressive sampling. Signal Processing, IEEE Transaction on 59(5).

[43] Hastie, Trevor and Tibshirani, Robert and Wainwright, Martin (2015). Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman & Hall/CRC.

[44] Deng, J. and Dong, W. and Socher, R. and Li, L.-J. and Li, K. and Fei-Fei, L.(2009) ImageNet: A Large-Scale Hierarchical Image Database.

[45] Shaobing Chen and David Donoho.(1994). Basis Pursuit.

A Proof of Chapter 2

Assumption and Notation used in proofs of Chapter 2

Assumption.

$$A_1 : (Z_i, Y_i)_{i=1}^{n} \;\; are \;\; i.i.d.$$

$$A_2 : \|\phi(\cdot)\|_\infty < M_d$$

$$A_3 : M_d \sqrt{log 2d} \sim \mathcal{O}(\sqrt{n})$$

$$A_4 : For \;\; \forall a_0 > 0, \exists J < \infty, \;\; such \;\; that, a_{J-1} = a_0\, 2^J \geq 2\|\beta^*\|_1$$

$$A_5 : \|\, \beta^* \,\|_0 \; \leq \; s$$

$$A_6 : \delta\, L_n\, \rho_{(\,\Delta,\, \beta^*\,)}\, (\, Z,\, Y\,) \geq \; \kappa \,\|\, \Delta \,\|_2$$

Notation:

$$\lambda^* \;\equiv\; \sqrt{2}\; M_d \sqrt{\frac{\log\,(\,2\,d\,) + \log n}{n}}\;;$$

$$\delta_1 \;\equiv\; \frac{2M_d}{n};$$

$$\delta_2 \;\equiv\; 2\, M_d \sqrt{\frac{2\, log\, 2d}{n}}$$

$$\delta_0 = \delta_1 + \delta_2$$

A1. Proof of Theorem 2.1

Theorem 2.1 [Event A] Under Assumption $A_1 - A_3$, if $\lambda > \lambda^*$, it holds that:

$$P\ (\ \beta^* \in \mathcal{C}_\lambda\ ) > 1 - \frac{1}{n}.$$

Proof.

$$L_n\, \rho_{\beta^*}\ (Z, Y) = \frac{1}{n} \sum_{i=1}^{n} \Big\{ Y_i\, Z_i \beta^* - \log\left[1 + \exp\left(Z_i \beta^*\right)\right] \Big\}.$$

Let $\nabla_{\beta_j}\ L_n\, \rho_{\beta^*}\ (Z, Y)$ denote the gradient of $L_n\, \rho_{\beta^*}\ (Z, Y)$ with respect to β_j, then for each j,

$$\nabla_{\beta_j}\ L_n\, \rho_{\beta^*}\ (Z, Y) = \frac{1}{n} \sum_{i=1}^{n} \Big\{ Z_{ij}\ [\ Y_i - \mu\ (\ Z_i \beta^*\ )\ ] \Big\}$$

where

$$\mu\,(\ Z_i \beta^*\ ) = \frac{\exp\left(Z_i \beta^*\right)}{1 + \exp\left(Z_i \beta^*\right)}\ \in\ (\,0, 1\,)$$

let $\epsilon_i = Y_i - \mu\ (\ Z_i \beta\ )$, then $\epsilon_i\ \in\ (\ -1,\ 1\ )$

$$\nabla_{\beta_j}\ L_n\, \rho_{\beta^*}\ (Z, Y) = \frac{1}{n} \sum_{i=1}^{n}\ Z_{ij}\, \epsilon_i$$

Since $\{\ Z_i,\ Y_i\ \}_{i=1}^{n}$ are $i.i.d$, then for each j, $\{\ Z_{ij} \epsilon_i\ \}_{i=1}^{n}$ is a set of n independent random variables.

We have following properties for $Z_{ij}\epsilon_i$:

According to Lemma 5.1, we have

$$E\ (\ Z_{ij} \epsilon_i\ ) = 0; \tag{11}$$

According to Assumption A_2, $\|Z\|_\infty \leq M_d$ and $\epsilon_i \in (-1,\ 1)$, we have

$$Z_{ij}\,\epsilon_i \;\in\; (-M_d,\ M_d) \tag{12}$$

Then, combine (11) and (12), we can apply Hoeffding's Inequality to $\{\, Z_{ij}\epsilon_i \,\}_{i=1}^n$,

$$\begin{aligned} & P\left\{ \left| \nabla_{\beta_j} L_n\, \rho_{\beta^*}\ (Z, Y) \right| > \lambda \right\} \\ &= P\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} Z_{ij}\epsilon_i \right| > \lambda \right\} \leq\ 2\ \exp\ \left[-\frac{2\, n^2\, \lambda^2}{\sum_{i=1}^n (\, 2\, M_d\,)^2} \right] \\ &= 2\ \exp\ \left(-\frac{n\, \lambda^2}{2\, M_d^2} \right) \end{aligned} \tag{13}$$

Then by De Morgan's Law, we abtain the union bound:

$$\begin{aligned} P\left\{ \beta^* \in \mathcal{C}_\lambda \right\} &= P\left\{ \left\| \nabla_\beta\, L_n\, \rho_{\beta^*}\ (Z, Y) \right\|_\infty \leq \lambda \right\} \\ &= P\left(\cap\ _{j=1}^{d} \left\{ \left| \nabla_{\beta_j}\ L_n\, \rho_{\beta^*}\ (Z, Y) \right| \leq \lambda \right\} \right) \\ &= 1 - P\Big(\cup\ _{j=1}^{d} \left\{ \left| \nabla_{\beta_j}\ L_n\, \rho_{\beta^*}\ (Z, Y) \right|\ >\ \lambda \right\} \Big) \\ &\geq 1 - \sum_{j=1}^{d}\ P\left\{ \left| \nabla_{\beta_j}\ L_n\, \rho_{\beta^*}\ (Z, Y) \right|\ >\ \lambda \right\} \end{aligned} \tag{14}$$

Plug the result of (13) into (14), it holds,

$$(14) \geq 1 - 2d\ \exp\ \left(-\frac{n\, \lambda^2}{2\, {M_d}^2} \right) = 1 - \exp\left[\left(-\frac{n\, \lambda^2}{2\, {M_d}^2} \right) + \log\left(\, 2\, d\,\right) \right] \tag{15}$$

therefore,

$$P\left\{ \left\| \nabla_\beta\, L_n\, \rho_{\beta^*}\ (Z, Y) \right\|_\infty \leq \lambda \right\} \geq 1 - \exp\ \left[\left(-\frac{n\, \lambda^2}{2\, {M_d}^2} \right) + \log\left(\, 2\, d\,\right) \right]$$

let

$$\lambda \;\geq\; \sqrt{2}\; M_d \sqrt{\frac{\log\,(\,2\,d\,)+\tau}{n}}\;;$$

then

$$P\;(\,\beta^*\;\in\,\mathcal{C}_\lambda\,) = P\,\left\{\;\Big\|\;\nabla_\beta\,L_n\,\rho_{\beta^*}\;(Z,Y)\;\Big\|_\infty \leq \sqrt{2}\;M_d\sqrt{\frac{\log\,(\,2\,d\,)+\tau}{n}}\;\right\} > 1-e^{-\tau}.$$

To be specific, set $\tau = \log 2d$, then with

$$\lambda \;\geq\; \sqrt{2}\; M_d \sqrt{\frac{\log\,(\,2\,d\,)+\log n}{n}}\;\equiv \lambda^*,$$

it holds that:

$$P\;(\,\beta^*\;\in\,\mathcal{C}_\lambda\,) \geq 1-\frac{1}{n}.$$

□

Lemma 5.1. *With same notations in Theorem 2.1,*

$$E\,(\;Z_{ij}\epsilon_i\;) = 0$$

Proof.

By definition of $L_n\,\rho_{\beta^*}\;(Z,Y)$,

$$L_n\,\rho_{\beta^*}\;(Z,Y)\;=\frac{1}{n}\sum_{i=1}^{n}\;\rho_{\beta^*}\;(\,Z_i,Y_i\,)$$

Thus for $1 \leq j \leq d$, the gradient w.r.t. β^* is:

$$\nabla_{\beta_j}\,L_n\,\rho_{\beta^*}\;(Z,Y)\;=\frac{\partial\,\frac{1}{n}\;\sum_{i=1}^{n}\;\rho_\beta(\,Z_i,Y_i\,)}{\partial\,\beta_j}\;\Big|_{\beta^*} = \frac{1}{n}\;\sum_{i=1}^{n}\;\frac{\partial\,\rho_\beta(\,Z_i,Y_i\,)}{\partial\,\beta_j}\;\Big|_{\beta^*}$$

Take the expectation of the gradient, combine with the definition of β^* and (2), we have:

$$
\begin{aligned}
&E\,[\,\nabla_{\beta_j}\, L_n\, \rho_{\beta^*}\,(Z, Y)\,] = E\left\{\,\frac{1}{n}\,\sum_{i=1}^{n}\,\frac{\partial\,\rho_{\beta}(\,Z_i, Y_i\,)}{\partial\,\beta_j}\,\Big|_{\beta^*}\,\right\} \\
&= E\left\{\,\frac{\partial\,\rho_{\beta}(\,Z, Y\,)}{\partial\,\beta_j}\,\Big|_{\beta^*}\,\right\} = \frac{\partial\, E\,[\,\rho_{\beta}\,(Z, Y)\,]}{\partial\,\beta_j}\,\Big|_{\beta^*} \\
&= \frac{\partial\; L\,\rho_{\beta^*}\,(Z, Y)}{\partial \beta_j}\bigg|_{\beta^*} = 0.
\end{aligned} \tag{16}
$$

Therefore,

$$
E\,(\,Z_{ij}\epsilon_i\,) = E\left[\,\frac{1}{n}\,\sum_{i=1}^{n}\,Z_{ij}\epsilon_i\,\right] = E\left[\,\nabla_{\beta_j}\;L_n\,\rho_{\beta^*}\,(Z, Y)\,\right] = 0;
$$

□

A2. Parameter Error Bound

2.1 Preliminary

2.1.1 Concentration Inequality

Theorem 5.1 (Hoeffding's Inequality)**.**

If $Z_1, Z_2, \ldots, Z_n$ are independent with $P\ (\ a_i \leq Z_i \leq b_i\ ) = 1$, then for any $t > 0$,

$$P\ (\ |\ \frac{1}{n}\ \sum_{i=1}^{n}\ Z_i - E(Z)\ |\ > \lambda\ ) \leq 2\, e^{-2n\lambda^2/c};$$

where $c = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(b_i - a_i)^2.$

Theorem 5.2 (McDiarmid Inequality[23])**.** *Let $Z_1, \ldots, Z_n \in \mathcal{Z}$ be independent random variables, a mapping $G : \mathcal{Z} \rightarrow R$, and there exist nonnegative numbers $c_1, \ldots, c_n$ such that* $\forall i \in \{1, , n\}$*, and* $\forall Z_1, \ldots, Z_n, Z'_k \in \mathcal{Z}$*, the function G satisfies*

$$\sup_{Z_1,\ldots,Z_n,Z'_i} \left| G(Z_1, \ldots, Z_i, \ldots, Z_n) - G(Z_1, \ldots, Z'_i, \ldots, Z_n) \right| \leq c_i \tag{17}$$

then,

$$P\left(\left|\ G(Z_1, \ldots, Z_n) - E\big[G(Z_1, \ldots, Z_n)\big] \right| \geq \delta \right) \leq 2 \exp\left(-\frac{2\delta^2}{\sum_{i=1}^{n}\ c_i^2} \right) \tag{18}$$

Lemma 5.2. *[7]*

Let Z be a random variable with mean 0, and $a \leq\ Z \leq b$. Then, for any t,

$$E\ (\ e^{tZ}\ ) \leq e^{t^2(b-a)^2/8}.$$

2.1.2 Measure of Complexity

To develop uniform bound, it's necessary to introduce a way to measure how complex the hypothesis class is. There are several approaches to measure the complexity such as VC Dimension, Covering, Rademacher Complexity, etc. In this theoretical study, we utilize Rademacher Complexity to measure the complexity of function class for high confidence set selection.

Definition 5.1 (Rademacher Random Variable)**.**

Rademacher Random Variable $\{r_1, r_2, \ldots, r_n\}$is a set of independent and identical random variables, with $P(r_i = 1) = P(r_i = -1) = 0.5$.

Definition 5.2 (Rademacher Complexity)**.** [7]

Rademacher Complexity of $\mathcal{F}$ is

$$Rad\ (\ \mathcal{F}\ ) = E\left[\ \sup_{f \in \mathcal{F}}\ \frac{1}{n} \sum_{i=1}^{n}\ r_i\ f(\ Z_i\ )\ \right]$$

The more complex the function class is , the larger the $Rad\ (\ \mathcal{F}\ )$ would be. Intuitively, if the function class is complex enough, it's possible to pick some $f \in \mathcal{F}$, which match the sign of Rademacher Random Variable, to make the $Rad(\mathcal{F})$ large. There are a lot of important properties of Rademacher Complexity. we introduce one useful Lemma below, which apply symmetrization technique.

Lemma 5.3 (Symmetrization Theorem [24])**.** *Let $Z_1, ..., Z_n$ be indepedent random variables with values in $\mathcal{Z}$, $r_1, ..., r_n$ be a Rademacher sequence independent of $Z_1, ..., Z_n$; f is* *a real valued functions on $\mathcal{Z}$, Then*

$$E\left(\sup_{f\in\mathcal{F}} \left| \, (\, L_n - L \,) \, f(Z_i) \, \right| \right) \; \leq \; 2\, E \left(\, \sup_{f\,\in\,\mathcal{F}} \, \left| \, \frac{1}{n} \sum_{i=1}^{n} r_i \, f \, (\, Z_i \,) \, \right| \right).$$

Lemma 5.4 (Contraction Theorem [14]). *Let $Z_1, ..., Z_n$ be non-random elements of some space $\mathcal{Z}$ and let $\mathcal{F}$ be a class of real valued functions on $\mathcal{Z}$, Consider Lipschitz functions $\rho_i : R \rightarrow R$, i.e.*

$$\left| \, \rho_i(x) - \rho_i(x') \, \right| \leq | \, x - x' \, |, \forall x, x' \in R,$$

Let $r_1, ..., r_n$ be a Rademacher sequence. Then for any function $\phi : \mathcal{Z} \rightarrow R$, we have:

$$E\left(\sup_{f\in\mathcal{F}} \left| \, \sum_{i=1}^{n} r_i \left[\, \rho_i(\, \phi(x_i) \,) - \rho_i(\, \phi'(x_i) \,) \, \right] \, \right| \right) \leq 2\, E\left(\sup_{f\in\mathcal{F}} \left| \, \sum_{i=1}^{n} r_i \left[\, \phi(x_i) - \phi'(x_i) \, \right] \, \right| \right).$$

2.2 Proof of Theorem 2.2

Theorem 2.2 Under Assumption $A1 - A4$, when $\lambda > \lambda^*$, with probability at least $1 - 2J\,e^{-2n} - \frac{1}{n}$, it holds that:

$$\mathcal{E}(\hat{\beta}_{HCS}) \leq (\,\lambda + \delta\,)\ \|\ \beta^* - \hat{\beta}_{HCS}\ \|_1\ +\ \delta\ a_0$$

proof of Theorem 2.2. Define the solution set of HCS:

$$\mathcal{B}_\lambda := \left\{\ \hat{\beta}\ \in\ R^d:\ \hat{\beta} = \underset{\beta\,\in\,\mathcal{C}_\lambda\,)}{\arg\min}\ \|\ \beta\ \|_1\ \right\} \tag{19}$$

Define a quantity V_λ:

$$V_\lambda = \sup_{\hat{\beta}\,\in\,\mathcal{B}_\lambda} \frac{(\,L_n - L\,)\,(\,\rho_{\beta^*} - \rho_{\hat{\beta}}\,)}{a_0 + \|\hat{\beta} - \beta^*\|_1} \tag{20}$$

Where L_n is the emperical loss operator, L is the expected loss operator;

$\rho_{\hat{\beta}}$, ρ_{β^*} are logistic loss with respect to $\hat{\beta}$ and β^* respectively;

a_0 is a small quantity which by assumption A_4 satisfies: $a_0 > \frac{\|\beta^*\|_1}{2^J}$.

Construct a partition set of $\mathcal{B}_\lambda$ according to the distance between β^* and $\hat{\beta}$:

$$\begin{aligned}
&\mathcal{B}_0 = \{\hat{\beta} : \hat{\beta} \in \mathcal{B}_\lambda, \|\hat{\beta} - \beta^*\|_1 \leq a_0\} \\
&\mathcal{B}_j = \{\hat{\beta} : \hat{\beta} \in \mathcal{B}_\lambda, a_{j-1} < \|\hat{\beta} - \beta^*\|_1 \leq a_j\}; (1 \leq j \leq J-1) \\
&\mathcal{B}_J = \{\hat{\beta} : \hat{\beta} \in \mathcal{B}_\lambda, \|\hat{\beta} - \beta^*\|_1 > a_{J-1}\}
\end{aligned} \tag{21}$$

For $1 \leq j \leq J-1$: $a_j = 2a_{j-1}$, by Assumption A_4, it holds $a_{J-1} \geq 2\|\beta^*\|_1$.

Then, we can derive the bound according to this partition $\mathcal{B}_\lambda$ as follow:

$$P\,(\,V_\lambda > \delta_0\,) = P\,\Big(\sup_{\hat\beta\,\in\,\mathcal{B}_\lambda}\;\frac{(\,L_n - L\,)\,(\,\rho_{\beta^*} - \rho_{\hat\beta}\,)}{a_0 + \|\hat\beta - \beta^*\|_1} > \delta_0\,\Big)$$

$$\le \sum_{j=0}^{J} P\,\Big(\sup_{\hat\beta\,\in\,\mathcal{B}_j}\;\frac{(\,L_n - L\,)\,(\,\rho_{\beta^*} - \rho_{\hat\beta}\,)}{a_0 + \|\hat\beta - \beta^*\|_1} > \delta_0\,\Big). \tag{22}$$

to be simplified, let

$$V_j = \sup_{\hat\beta\,\in\,\mathcal{B}_j}\;\frac{(\,L_n - L\,)\,(\,\rho_{\beta^*} - \rho_{\hat\beta}\,)}{a_0 + \|\hat\beta - \beta^*\|_1} \tag{23}$$

then (22) is equivalent to

$$P\,(\,V_\lambda > \delta_0\,) \le \sum_{j=0}^{J} P\,\Big(V_j > \delta_0\,\Big).$$

According to Lemma 5.5: For $0 \le j \le J-1$

$$P\,\Big(V_j > \delta_0\,\Big) < 2\,e^{-2n}$$

By Theorem 2.1, when $\lambda > \lambda^*$, it holds

$$P\,\Big(V_J > \delta_0\,\Big) \le P\,\Big(\,\|\hat\beta - \beta^*\|_1 > a_{J-1}\Big) \le P\,\Big(\,\|\hat\beta - \beta^*\|_1 > 2\,\|\beta^*\|_1\Big) \le P\,(\,\beta^* \notin \mathcal{C}_\lambda\,)\; \le\; \frac{1}{n}$$

Thus,

$$P\,\Big(V_\lambda > \delta_0\,\Big) < 2J\,e^{-2n} + \frac{1}{n}.$$

It comes to conclude that, with probability at least $1 - 2J\,e^{-2n} - \frac{1}{n}$, it holds:

$$V_\lambda := \sup_{\hat\beta\in\mathcal{B}_\lambda}\;\frac{|(\,L_n - L\,)\,(\,\rho_{\beta^*} - \rho_{\hat\beta}\,)|}{a_0 + \|\beta^* - \hat\beta\|_1}\; < \delta_0.$$

Since our estimator $\hat\beta_{HCS} \in \mathcal{B}_\lambda$, it holds

$$\frac{|\,(\,L_n - L\,)\,(\,\rho_{\beta^*} - \rho_{\hat\beta_{HCS}}\,)\,|}{a_0 + \|\beta^* - \hat\beta_{HCS}\|_1} \le \sup_{\hat\beta\in\mathcal{B}_\lambda}\;\frac{|(\,L_n - L\,)\,(\,\rho_{\beta^*} - \rho_{\hat\beta}\,)|}{a_0 + \|\beta^* - \hat\beta_{HCS}\|_1}\; \le \delta_0;$$

thus,

$$L_n\ (\ \rho_{\beta^*} - \rho_{\hat{\beta}_{HCS}}\ ) - L\ (\ \rho_{\beta^*} - \rho_{\hat{\beta}_{HCS}}\ ) \leq\ \ \delta_0\ a_0\ \ + \delta\ \|\ \beta^* - \hat{\beta}_{HCS}\ \|_1\ .$$

rearrange the orders, then

$$\mathcal{E}(\hat{\beta}_{HCS}) = L\ (\ \rho_{\hat{\beta}_{HCS}} - \rho_{\beta^*}\ ) \leq L_n\ (\ \rho_{\hat{\beta}_{HCS}} - \rho_{\beta^*}\ ) +\ \delta_0\ \|\ \beta^* - \hat{\beta}_{HCS}\ \|_1\ +\ \delta\ a_0$$

by (3): $L_n(\ \rho_{\hat{\beta}_{HCS}} - \rho_{\beta^*}\ )\ \leq\ \lambda\ \|\ \hat{\beta}_{HCS} - \beta^*\ \|_1$, thus

$$\mathcal{E}(\hat{\beta}_{HCS}) \leq (\ \lambda + \delta\ )\ \|\ \beta^* - \hat{\beta}_{HCS}\ \|_1\ +\ \delta_0\ a_0$$

□

Lemma 5.5. *As $\mathcal{B}_j$, V_j, defined in (21), (23), for $0 \leq j \leq J-1$, it holds:*

$$P\ (\ V_j > \delta_0\ ) < 2\ e^{-2n}.$$

Proof. The process to prove V_j is bouned contains following two steps: Step 1, prove V_j is concerntrated around its mean $E(\ V_j\ )$; Step 2, prove the mean $E(V_j)$ is bounded above.

Step 1: Concerntraction around mean:

$$P\ \Big(\ |\ V_j - E\ (\ V_j\ )\ | > \delta_1\ \Big)\ < 2\ e^{-2n}$$

Denote $\{D_i\}_{i=1}^n = (Z_i, Y_i)_{i=1}^n$. It suffices to apply McDiarmid Inequality (Theorem 5.2) to derive the concentration bound if it satisfies:

$$\sup_{D_i} \Big|\ V_j\ (\ D_1, \ldots, D_k, \ldots, D_n\ ) - V_j\ (\ D_1, \ldots, D'_k, \ldots, D_n\ )\ \Big| \leq c_i.$$

$$\text{Let}\ \ h_\beta = \frac{\rho^*_\beta - \rho_\beta}{a_0 + \|\hat{\beta} - \beta^*\|_1}\ \ \text{ and }\ \ \bar{h}_\beta = h_\beta - E(h_\beta) \tag{24}$$

then,

$$V_j(\, D_1, \ldots, D_n \,) = \sup_{\beta \in \mathcal{B}_j} \frac{(L_n - L)(\rho^*_\beta - \rho_\beta)\left\{\, D_1, \ldots, D_n \,\right\}}{a_0 + \|\, \hat{\beta} - \beta^* \|_1}$$
$$= \sup_{\beta \,\in\, \mathcal{B}_j} \frac{1}{n} \sum_{i=1}^{n} \bar{h}_\beta\,(\, D_i\,) \tag{25}$$

Construct a set $\left\{\, D'_i \,\right\}_{i=1}^{n}$, such that:

$$D'_i = \begin{cases} (\, Z'_i,\, Y'_i\,) & \text{when } i = k \\ \\ (\, Z_i,\, Y_i\,) & \text{when } i \neq k \end{cases}$$

then,

$$V_j\,(\, D'_1, \ldots, D'_k, \ldots, D'_n\,) = \sup_{\beta \,\in\, \mathcal{B}_j} \frac{1}{n} \sum_{i=1}^{n} \bar{h}_\beta\,(\, D'_i\,)$$

By definition of V_j in (25), for $\forall \beta_1 \in \mathcal{B}_j$ we have:

$$\frac{1}{n} \sum_{i=1}^{n} \bar{h}_{\beta_1}\,(\, D_i\,) \; - V_j\,(\, D_1, \ldots, D'_k, \ldots, D_n\,) = \; \frac{1}{n} \sum_{i=1}^{n} \bar{h}_{\beta_1}\,(\, D_i\,) \; - \sup_{\beta \,\in\, \mathcal{B}_j} \frac{1}{n} \sum_{i=1}^{n} \bar{h}_\beta\,(\, D'_i\,)\;; \tag{26}$$

For $\forall \beta_1 \in \mathcal{B}_j$, it holds $\sup_{\beta \,\in\, \mathcal{B}_j} \frac{1}{n} \sum_{i=1}^{n} \bar{h}_\beta\,(\, D'_i\,) > \frac{1}{n} \sum_{i=1}^{n} \bar{h}_{\beta_1}\,(\, D'_i\,)$, thus,

$$(26) \leq \frac{1}{n} \sum_{i=1}^{n} \bar{h}_{\beta_1}\,(\, D_i\,) \; - \; \frac{1}{n} \sum_{i=1}^{n} \bar{h}_{\beta_1}\,(\, D'_i\,) \; \leq \; \frac{1}{n}\left[\, \bar{h}_{\beta_1}\,(\, D_k\,) \; - \bar{h}_{\beta_1}\,(\, D'_k\,)\,\right] \tag{27}$$

Since D_k and D'_k are from same distribution, we have

$$E\left[\, h_{\beta_1}\,(\, D_k\,)\,\right] = E\left[\, h_{\beta_1}\,(\, D'_k\,)\,\right] \tag{28}$$

Therefore, by definiton of $\bar{h}_\beta$, (24):

$$
\begin{aligned}
(27) = \; & \frac{1}{n}\left\{\Big(h_{\beta_1}\,(\,D_k\,) - E\,[\,h_{\beta_1}\,(\,D_k\,)]\Big) - \Big(h_{\beta_1}\,(\,D'_k\,) - E\,[\,h_{\beta_1}\,(\,D'_k\,)\,]\Big)\right\} \\
= \; & \frac{1}{n}\,[\,h_{\beta_1}\,(\,D_k\,) \; - h_{\beta_1}\,(\,D'_k\,)\,]
\end{aligned}
$$

by definition of h_β (24),

$$
= \; \frac{1}{n}\left\{\frac{(\,\rho^*_\beta - \rho_{\beta_1}\,)\,(\,D_k\,) - \,(\,\rho^*_\beta - \rho_{\beta_1}\,)\,(\,D'_k\,)}{a_0 + \|\beta^* - \beta_1\|_1}\right\}
$$

by the triangle inequality,

$$
\leq \; \frac{1}{n}\left\{\frac{\big|\,(\,\rho^*_\beta - \rho_{\beta_1}\,)\,(\,Z_k,\,Y_k\,)\,\big| + \big|\,(\,\rho^*_\beta - \rho_{\beta_1}\,)\,(\,Z'_k,\,Y'_k\,)\,\big|}{a_0 + \|\beta^* - \beta_1\|_1}\right\}
$$

by Lipschitz property of ρ_β,

$$
\leq \; \frac{1}{n}\left\{\frac{\big|\,Z_k^T\beta^* - Z_k^T\beta_1\big| \; + \big|\,Z_k'^T\beta^* - Z_k'^T\beta_1\big|}{a_0 + \|\beta^* - \beta_1\|_1}\right\}
$$

by Holder's inequality,

$$
\leq \; \frac{1}{n}\left\{\frac{\|\,Z_k^T\,\|_\infty\|\,\beta^* - \beta_1\,\|_1 \;\; + \;\|\,Z_k'^T\,\|_\infty\|\,\beta^* - \beta_1\,\|_1}{a_0 + \|\beta^* - \beta_1\|_1}\right\}
$$

by Assumption:

$$
\leq \; \frac{2M_d}{n}\;.
$$

Since for $\forall \beta_1 \in \mathcal{B}_j$, it holds

$$
\frac{1}{n}\sum_{i=1}^{n}\,\bar{h}_{\beta_1}\,(\,D_i\,) \; - V_j\,(\,D_1,\ldots,D'_k,\ldots,D_n\,) \leq \frac{2M_d}{n}
$$

thus,

$$
\sup_{\beta\in\mathcal{B}_j}\,\frac{1}{n}\sum_{i=1}^{n}\,\bar{h}_\beta\,(\,D_i\,) \; - V_j\,(\,D_1,\ldots,D'_k,\ldots,D_n\,) \leq \frac{2M_d}{n}\;,
$$

by (25),

$$
V_j\,(\,D_1,\ldots,D_k,\ldots,D_n\,) = \sup_{\beta\in\mathcal{B}_j}\,\frac{1}{n}\sum_{i=1}^{n}\,\bar{h}_\beta\,(\,D_i\,)\;,
$$

thus,

$$V_j\ (\,D_1, \ldots, D_k, \ldots, D_n\,) - V_j\ (\,D_1, \ldots, D'_k, \ldots, D_n\,) \leq \frac{2M_d}{n}\ .$$

Analogously, it can be proved

$$V_j\ (\,D_1, \ldots, D'_k, \ldots, D_n\,) - V_j\ (\,D_1, \ldots, D_k, \ldots, D_n\,) \leq \frac{2M_d}{n}\ ,$$

thus,

$$\left|\ V_j\ (\,D_1, \ldots, D_k, \ldots, D_n\,) - V_j\ (\,D_1, \ldots, D'_k, \ldots, D_n\,)\ \right| \leq \frac{2M_d}{n}\ .$$

Since $D_1, \ldots, D_n$ are i.i.d,

$$\sup_{D_1, \ldots, D_k, \ldots, D_n, D_{k'}} \left|\ V_j\ (\,D_1, \ldots, D_k, \ldots, D_n\,) - V_j\ (\,D_1, \ldots, D'_k, \ldots, D_n\,)\ \right| \leq \frac{2M_d}{n}\ .$$

Thus, the condition (17) in McDiarmid Inequality (Theorem 5.2) meets with

$$c_i = \frac{2M_d}{n}\ .$$

By setting $\delta = \dfrac{2M_d}{n}$, it conclude that:

$$P\left(\,|\ V_j\ -\ E\,(\,V_j\,)\,|\ >\ \frac{2M_d}{n}\ \right)\ <\ 2e^{-2n} \tag{29}$$

Step 2: Upper Bounded $E(V_j)$

$$E\ (\,V_j\,)\ \leq\ \delta_2$$

$$E\,(\,V_j\,) = E\left(\ \sup_{\hat{\beta} \in \mathcal{B}_j} \frac{\left|\ (\,L_n - L\,)\,(\rho_{\beta^*} - \rho_{\hat{\beta}})\ \right|}{a_0 + \|\ \beta^* - \hat{\beta}\ \|_1}\right)$$

recall the definition of $\mathcal{B}_j$:

$$\mathcal{B}_j = \{\hat{\beta} : \hat{\beta} \in \mathcal{B}_\lambda, a_{j-1} < \| \beta^* - \hat{\beta} \|_1 < a_j\};$$

thus,

for $\quad j = 0$:

$$a_0 + \| \beta^* - \hat{\beta} \|_1 \geq a_0,$$

for $\quad 1 \leq j \leq J-1$:

$$a_0 + \| \beta^* - \hat{\beta} \|_1 \geq a_{j-1}.$$

therefore,

for $\quad j = 0$:

$$E\,(\,V_j\,) \leq \frac{1}{a_0}\,E\left(\sup_{\hat{\beta}\in\mathcal{B}_j}\,|\,(\,L_n - L\,)\,(\rho_{\beta^*} - \rho_{\hat{\beta}})\,|\right)$$

for $\quad 1 \leq j \leq J-1$:

$$E\,(\,V_j\,) \leq \frac{1}{a_{j-1}}\,E\left(\sup_{\hat{\beta}\in\mathcal{B}_j}\,|\,(\,L_n - L\,)\,(\rho_{\beta^*} - \rho_{\hat{\beta}})\,|\right)$$

Denote $\tilde{h}_{\hat{\beta}} = \rho_{\beta^*} - \rho_{\hat{\beta}}$, by Symmetrization Lemma (Lemma 5.3), it holds:

$$E\left(\sup_{\hat{\beta}\in\mathcal{B}_j}\,|\,(\,L_n - L\,)\,\tilde{h}_{\hat{\beta}}\,|\right) \;\leq\; 2\,Rad\,\{\,\tilde{h}_{\hat{\beta}},\;\hat{\beta}\in\mathcal{B}_j\,\}; \tag{30}$$

Where $Rad\,\{\,\tilde{h}_{\hat{\beta}},\;\hat{\beta}\in\mathcal{B}_j\,\}$ is Rademacher Complexity of $\{\,\tilde{h}_{\hat{\beta}},\;\hat{\beta}\in\mathcal{B}_j\,\}$:

$$Rad\,\{\,\tilde{h}_{\hat{\beta}},\;\hat{\beta}\in\mathcal{B}_j\,\} = E\left(\sup_{\hat{\beta}\,\in\,\mathcal{B}_j}\left|\,\frac{1}{n}\sum_{i=1}^{n} r_i\,\tilde{h}_{\hat{\beta}}\,(\,Z_i,\,Y_i\,)\,\right|\right);$$

and $\{r_i\}_{i=1}^n$ is a set of i.i.d Rademacher random variable.

Thus,

$$E \left(\sup_{\hat\beta \in \mathcal{B}_j} \mid (L_n - L) \, \tilde h_{\hat\beta} \mid \right) \le 2\, E \left(\sup_{\hat\beta \in \mathcal{B}_j} \left| \frac{1}{n} \sum_{i=1}^{n} r_i \, \tilde h_\beta \, (Z_i, \, Y_i) \right| \right)$$

By Contraction Theorem (Lemma 5.4),

$$\le 2\, E \left(\sup_{\hat\beta \in \mathcal{B}_j} \left| \frac{1}{n} \sum_{i=1}^{n} r_i \, Z_i \, (\beta^* - \hat\beta) \right| \right)$$

By Holders Inequality,

$$\le 2\, E \left(\sup_{\hat\beta \in \mathcal{B}_j} \left[\frac{1}{n} \; \| Z^T \, r \|_\infty \, \| \, \beta^* - \hat\beta \, \|_1 \; \right] \right)$$

since $\|\beta^* - \hat\beta\|_1 < a_j$

$$\le 2\, a_j \, E \; \left[\frac{1}{n} \; \| Z^T \, r \|_\infty \; \right]$$

By Lemma 5.6:

$$\le 2 a_j \, M_d \sqrt{\frac{2 \, log \, 2d}{n}}$$

Thus, for $j = 0$:

$$E \, (\, V_0 \,) \le \frac{2a_0}{a_0} \; M_d \sqrt{\frac{2 \, log \, 2d}{n}} \le 2 \, M_d \sqrt{\frac{2 \, log \, 2d}{n}};$$

for $1 \le j \le J - 1$:

$$E \, (\, V_j \,) \le \frac{2a_j}{a_{j-1}} \; M_d \sqrt{\frac{2 \, log \, 2d}{n}} \le 4 \, M_d \sqrt{\frac{2 \, log \, 2d}{n}};$$

which concludes that, for $0 \le j \le J - 1$:

$$E \, (\, V_j \,) \le 4 \, M_d \sqrt{\frac{2 \, log \, 2d}{n}} \equiv \delta_2. \tag{31}$$

Set $\delta_0 = \delta_1 + \delta_2$, combine (29) and (31), it concludes:

$$P \, (\, V_j > \delta_0 \,) < 2Je^{-2n} + \frac{1}{n} \tag{32}$$

□

Lemma 5.6. *With same notations of Lemma 5.5, it holds:*

$$E\left(\max_{1 \le j \le d} \frac{1}{n} \mid Z_j^T\, r \mid \right) \le M_d \sqrt{\frac{2\, log\, 2d}{n}}.$$

Proof.

Let $T_{ij} = \dfrac{1}{n}\, Z_{ij}\, r_i,$ then $E\,(\,T_{ij}\,) = 0,$ and $|T_{ij}| \;\le\; \dfrac{M_d}{n}.$

By Lemma 5.2,

$$E\;[\;exp\;(\;t\;T_{ij}\;)\;]\;\le exp\;(\;\frac{t^2\;{M_d}^2}{2\;n^2}\;). \tag{33}$$

Let $T_j = \sum_{i=1}^n T_{ij}$, then,

$$E\;[\;exp\;(\;t\;T_j\;)\;] = E\;[\;exp\;(\;t\;\sum_{i=1}^n\;T_{ij}\;)\;]$$

by independency of $\{T_{ij}\}_{i=1}^n$,

$$= \prod_{i=1}^n E\;[\;exp\;(\;t\;T_{ij}\;)\;]$$

by the result of (33),

$$\le \prod_{i=1}^n exp\;(\;\frac{t^2 {M_d}^2}{2n^2}\;) = exp\;(\;\frac{t^2 {M_d}^2}{2n}\;) \tag{34}$$

Thus, T_j is a subgaussian random variable with $\sigma = \frac{M_d}{\sqrt{n}}$.

Create a set $\{\;T'_j\;\}_{\;j=1}^{\;2d}$, with $2d$ elements, where

$$\{\;T'_j\;\}_{\;j=1}^{\;d} = \{\;T_j\;\}_{\;j=1}^{\;d};$$

$$\{\;T'_j\;\}_{\;j=d+1}^{\;2d} = \{\;-T_j\;\}_{\;j=1}^{\;d}.$$

Notice that,

$$\max_{1 \le j \le d} |\, T_j \,| = \max_{1 \le j \le 2d} T'_j;$$

thus,

$$exp\,[\; t\; E\;(\max_{1 \le j \le d} |\, T_j \,|\;)\;] = exp\,[\; t\; E\;(\max_{1 \le j \le 2d} T'_j\;)\;].$$

Then, by Jensen's Inequality and convexity of e^x, it holds:

$$exp\,[\; t\; E\;(\max_{1 \le j \le 2d} T'_j\;)\;] \le E\,[\; exp\;(\; t\; \max_{1 \le j \le 2d} T'_j\;)\;]$$

$$= E\,[\; \max_{1 \le j \le 2d} exp\;(\; t\, T'_j\;)\;]$$

$$\le \sum_{j=1}^{2d} E\,[\; exp\;(\; t\, T'_j\;)\;]$$

by the result of subgaussian tails in (34):

$$\le 2d\; exp\;(\; \frac{t^2\, {M_d}^2}{2\, n}\;).$$

take log for both sides, we have

$$E\;(\max_{1 \le j \le d} |\, T_j \,|\;) \le \frac{\log\, 2d}{t} + \frac{t\, {M_d}^2}{2\, n}.$$

setting $t = \sqrt{2n\, log\, 2d}\,/M_d$,

$$E\;(\max_{1 \le j \le d} |\, T_j \,|\;) \le M_d\, \sqrt{\frac{2\, log\, 2d}{n}}.$$

$\square$

A3. Proof of Theorem 2.3

Theorem 2.3: Under Assumption $A1 - A6$, when $\lambda \geq \lambda^*$, with probability at least $1 - \frac{1}{n}$, it holds:

$$(i) \quad \| \hat{\beta}_{HCS} - \beta^* \|_2 \leq \frac{4\lambda\sqrt{s}}{\kappa};$$

$$(ii) \quad \| \hat{\beta}_{HCS} - \beta^* \|_1 \leq \frac{8\lambda s}{\kappa}$$

proof of Theorem 2.3. Since $\hat{\beta}_{HCS}$ is the solution from $\mathcal{C}_\lambda$, by definition of $\mathcal{C}_\lambda$,

$$\| \nabla_\beta L_n \rho_{\hat{\beta}_{HCS}} (Z, Y) \|_\infty \leq \lambda$$

Under Event A, we have $\beta^* \in \mathcal{C}_\lambda$, therefore

$$\| \nabla_\beta L_n \rho_{\beta^*} (Z, Y) \|_\infty \leq \lambda;$$

then by the triangle inequality,

$$\| \nabla_\beta L_n \rho_{\hat{\beta}_{HCS}} (Z, Y) - \nabla_\beta L_n \rho_{\beta^*} (Z, Y) \|_\infty$$

$$\leq \| \nabla_\beta L_n \rho_{\hat{\beta}_{HCS}} (Z, Y) \|_\infty + \| \nabla_\beta L_n \rho_{\beta^*} (Z, Y) \|_\infty \leq 2\lambda; \tag{35}$$

let $\hat{\Delta} = \hat{\beta} - \beta^*$, then the first order Taylor error is

$$\delta L_n \rho_{(\hat{\Delta}, \beta^*)}(Z, Y) := L_n \rho_{(\beta^* + \hat{\Delta})}(Z, Y) - L_n \rho_{\beta^*}(Z, Y) - \left\langle \nabla_\beta L_n \rho_{\beta^*}(Z, Y), \hat{\Delta} \right\rangle$$

by first order derivative property of convexity function,

$$\leq \left\langle \nabla_\beta L_n \rho_{\hat{\beta}_{HCS}} (Z, Y), \hat{\Delta} \right\rangle - \left\langle \nabla_\beta L_n \rho_{\beta^*} (Z, Y), \hat{\Delta} \right\rangle$$

rearrange inner product,

$$= \left\langle \left[\nabla_\beta L_n \rho_{\hat{\beta}_{HCS}} (Z, Y) - \nabla_\beta L_n \rho_{\beta^*} (Z, Y) \right], \hat{\Delta} \right\rangle$$

by Holder's inequality,

$$\leq \| \nabla_\beta L_n \rho_{\hat{\beta}_{HCS}} (Z, Y) - \nabla_\beta L_n \rho_{\beta^*} (Z, Y) \|_\infty \| \hat{\Delta} \|_1$$

by the result of (35), it holds:

$$\leq 2\lambda \| \hat{\Delta} \|_1 \tag{36}$$

Since $\hat{\beta}_{HCS} = \arg\min_{\beta \in \mathcal{C}_\lambda} \| \beta \|_1$, thus $\| \hat{\beta}_{HCS} \|_1 \leq \| \beta^* \|_1$, similar to basis pursuit [45], we have following two properties for $\hat{\Delta}$:

$$\| \hat{\Delta}_{J_c} \|_1 \leq \| \hat{\Delta}_J \|_1 \tag{37}$$

$$\| \hat{\Delta} \|_1 \leq 2 \sqrt{s} \| \hat{\Delta} \|_2 \tag{38}$$

By $Assumption\ A_6$, $\hat{\Delta}$ satisfies restricted strong convexity assumption, that is,

$$\delta L_n \rho_{(\hat{\Delta}, \beta^*)} (Z, Y) \geq \kappa \| \hat{\Delta} \|_2^2;$$

combine with (36) and (38) we have,

$$\kappa \| \hat{\Delta} \|_2^2 \leq 2 \lambda \| \hat{\Delta} \|_1 \leq 4 \lambda \sqrt{s} \| \hat{\Delta} \|_2;$$

therefore,

$$\| \hat{\Delta} \|_2 \leq \frac{4 \lambda \sqrt{s}}{\kappa}; \tag{39}$$

plug (39) into (38), we have

$$\| \hat{\Delta} \|_1 \leq \frac{8 \lambda s}{\kappa}. \tag{40}$$

□

[Corollary] Under Assumption $A_1 - A_6$, when $\lambda > \lambda^*$, with probability at least $1 - 2J\, e^{-2n} - \frac{1}{n}$, it holds that:

$$\mathcal{E}(\hat{\beta}_{HCS}) \leq \frac{8\, \lambda\, s}{\kappa}\, (\, \lambda + \delta_0\,) + \; \delta_0\; a_0$$

Proof. Take the result of (38) into Theorem 2.2, the result can be achieved. □

B Proof of Chapter 3

Assumptions and Notations in Chapter 3

Assumption (C_1)**.** *$(Z_i, Y_i)_{i=1}^n$ are i.i.d., and $(W_i, Y_i)_{i=1}^n$ are i.i.d.;*

Assumption (C_2)**.** *$W = Z + \Xi$, and $E(W) = 0$.*

Assumption (C_3)**.** *$\|\phi(\cdot)\|_\infty < M_d$; i.e., $\| Z \|_\infty \leq M_d$; and $\| W \|_\infty \leq M_d$;*

Assumption (C_4)**.** *$M_d\sqrt{log2d^2} \sim \mathcal{O}(\sqrt{n})$;*

Assumption (C_5)**.** *For $\forall a_0 > 0, \exists J < \infty,\ such\ \ that, a_{J-1} = a_0\, 2^J \geq 2\|\beta^*\|_1$;*

Assumption (C_6)**.** *$\| \beta^* \|_0 \ \leq \ s$;*

Assumption (C_7)**.** *$\delta\ L_n\, \rho_{(\,\Delta,\,\beta^*\,)}\ (\,W,\,Y\,) \geq \ \kappa \,\|\,\Delta\,\|_2$*

Notation:

$$\lambda^* \ \equiv \ \sqrt{2}\ M_d \sqrt{\frac{\log\,(\,2\,d\,) + \log n}{n}}\ ;$$

$$\gamma^* \equiv \ M_d^2 \sqrt{\frac{\log\,(\,2\,d^2\,) + \log n}{2n}}\ ,$$

$$\delta_1 \ \equiv \ \frac{2M_d}{n};$$

$$\delta_2 \ \equiv \ 2\,M_d \sqrt{\frac{2\,log\,2d}{n}}$$

$$\delta_0 = \delta_1 + \delta_2$$

B1. Proof of Theorem 3.1

Theorem 3.1 [Event B]
Under Assumption $C_1 - C_4$, when $\lambda > \lambda^*$, $\gamma > \gamma^*$,

$$P\,[\,\beta^* \;\in\; \mathcal{C}_{(\,\lambda,\,\gamma\,)}\,] \;>\; 1 - \frac{2}{n}.$$

proof of Theorem 3.1.

Since $\quad L_n\,\rho_\beta\,(\,W,\,Y\,) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\;\Big\{\,Y_i\,W_i\,\beta - \log\,[\,1 + \exp\,(\,W_i\,\beta\,)\,]\,\Big\}\,;$

it holds, $\quad \nabla_\beta\;L_n\,\rho_\beta\,(\,W, Y\,) = \dfrac{1}{n}\;\displaystyle\sum_{i=1}^{n}\;\Big\{\,W_i\,[\,Y_i - \mu\,(\,W_i\,\beta\,)\,]\,\Big\};$

Thus, (**??**) is equivalent to

$$\mathcal{C}_{(\,\lambda,\,\gamma\,)} = \Big\{\;\beta\;\in\;R^d:\;\frac{1}{n}\,\|\,W^T\,[\,Y - \mu\,(\,W\beta\,)\,]\,\|_\infty\;\le\;\lambda + \;\gamma\,\|\,\beta\,\|_1\;\Big\}, \tag{41}$$

where

$$\mu\,(\,W\beta\,) = \frac{W\,\exp\,(\,W\beta\,)}{1 + \exp\,(\,W\beta\,)}\;\;\in\;(\,0, 1\,).$$

By Assumption C_2,

$$W\beta = Z\beta + \Xi\beta$$

thus by Cauchy Remainder Theorem,

$$\mu\,(\,W\beta\,) = \mu\,(\,Z\beta\,) + \mu'\,(\,\xi\,\beta\,)\,(\,\Xi\beta\,)$$

where $\xi\beta$ lies in the segment between $W\beta$ and $Z\beta$.

Therefore,

$$P\left(\beta^* \in \mathcal{C}_{(\lambda,\gamma)}\right)$$

$$=P\left\{\frac{1}{n}\parallel W^T[\,Y-\mu\,(W\beta^*)\,]\parallel_\infty\ \leq\ \lambda+\gamma\parallel\beta^*\parallel_1\right\}$$

$$=P\left\{\frac{1}{n}\parallel W^T[\,Y-\mu\,(Z\beta^*)-\mu'\,(\xi\beta^*)\,(\Xi\beta^*)\,]\parallel_\infty\ \leq\ \lambda+\gamma\parallel\beta^*\parallel_1\right\}$$

By triangle inequality,

$$\frac{1}{n}\parallel W^T[\,Y-\mu\,(Z\beta^*)-\mu'\,(\xi\beta^*)\,(\Xi\beta^*)\,]\parallel_\infty$$

$$\leq\frac{1}{n}\parallel W^T[\,Y-\mu\,(Z\beta^*)\,]\parallel_\infty+\frac{1}{n}\parallel W^T\mu'\,(\xi)\,(\Xi\beta^*)\parallel_\infty$$

Define

$$Event\ \ B:=\{\ \beta^*\ \in\ \mathcal{C}\,(\lambda,\gamma)\ \}$$

$$:=\{\ \frac{1}{n}\parallel W^T[\,Y-\mu\,(Z\beta^*)-\mu'\,(\xi\,\beta^*)\,(\Xi\beta^*)\,]\parallel_\infty\leq\lambda\ +\ \gamma\parallel\beta^*\parallel_1\};$$

Define

$$Event\ \ B_1:=\{\ \frac{1}{n}\parallel W^T[\,Y-\mu\,(Z\beta^*)\,]\parallel_\infty\leq\lambda\,\};$$

$$Event\ \ B_2:=\{\ \frac{1}{n}\parallel W^T\mu'\,(\xi\,\beta^*)\,(\Xi\beta^*)\parallel_\infty\leq\gamma\parallel\beta^*\parallel_1\}.$$

Notice that, $Event\ \ B_1$ and $Event\ \ B_2$ implies $Event\ \ B$. Now we investigate the probability of each event respectively.

(i) $Event$ B_1 :

$$P\,(\,B_1\,) = P\,\{\,\frac{1}{n}\,\|\,W^T\,[\,Y - \mu\,(\,Z\beta^*\,)\,]\,\|_\infty \leq \lambda\,\}$$

let $\epsilon = Y - \mu\,(\,Z\beta^*\,)$, then since $E(\epsilon) = 0$ and $\|W\|_\infty < M_d$, analogous to Theorem 2.1, it holds:

$$\text{when }\ \lambda\ \geq\ \sqrt{2}\,M_d\sqrt{\frac{\log\,(\,2\,d\,) + \log n}{n}}\ \equiv \lambda^*,$$

$$P\,(\,B_1\,) \geq 1 - \frac{1}{n}.$$

(ii) $Event$ B_2

Notice that

$$\mu\,(\,\xi_i\,\beta\,) = \frac{\exp\,(\,\xi_i\,\beta\,)}{1 + \exp\,(\,\xi_i\,\beta\,)}\ \in\ (\,0,\ 1\,),$$

then

$$\mu'\,(\,\xi_i\,\beta\,) = \frac{\exp\,(\,\xi_i\,\beta\,)}{[\,1 + \exp\,(\,\xi_i\,\beta\,)\,]^2} = \mu\,(\,\xi_i\,\beta\,)\,[\,1 - \mu\,(\,\xi_i\,\beta\,)\,]\ \in\ (\,0,\ \frac{1}{4}\,);$$

Thus,

$$\frac{1}{n}\,\|\,W^T\,\mu'\,(\,\xi\,\beta\,)\,(\,\Xi\beta^*\,)\,\|_\infty \leq \frac{1}{4\,n}\,\|\,W^T\,(\,\Xi\beta^*\,)\,\|_\infty \leq \frac{1}{4\,n}\,\|\,\Xi^T\,W\,\|_\infty\|\,\beta^*\,\|_1.$$

Define

$$Event\ \ B_2' :=\ \{\,\frac{1}{4\,n}\,\|\,\Xi^T\,W\,\|_\infty \leq \gamma\,\};$$

then,

$$P\,(\,B_2'\,) = P\left(\,\frac{1}{4\,n}\,\|\,\Xi^T\,W\,\|_\infty \leq \gamma\,\right) = P\left(\,\max_k\,\max_j\,\frac{1}{n}\,\left|\,\sum_{i=1}^{n}\,\frac{1}{4}\,\Xi_{ik}\,W_{ij}\,\right| \leq \gamma\,\right).$$

Denote

$$T_{i,k,j} = \frac{1}{4}\,\Xi_{ik}\,W_{ij},$$

then

$$\sum_{i=1}^{n} T_{i,k,j} = \sum_{i=1}^{n} \frac{1}{4} \, \Xi_{ik} \, W_{ij};$$

Since $E\,(\,W_{ij}\,) = 0$,

$$E\left(\frac{1}{n} \sum_{i=1}^{n} T_{i,k,j} \right) = E\left(\frac{1}{4\,n} \sum_{i=1}^{n} \Xi_{ik} \, W_{ij} \right) = \frac{1}{4\,n} \, E\left[E\left(W_{ij} \right) \sum_{i=1}^{n} \Xi_{ik} \right] = 0$$

and

$$|\, T_{i,j,k} \,| = \frac{1}{4} \, |\, \Xi_{ik} \, W_{ij} \,| \leq \frac{1}{4} \, \|\, \Xi \,\|_{\infty} \, \|\, W \,\|_{\infty} \leq \frac{1}{2} \, M_d^2$$

Then apply Hoeffding's Inequality,

$$P\left(\left| \frac{1}{n} \sum_{i=1}^{n} T_{ijk} \right| > \gamma \right) \;\leq\; 2 \;\exp\; \left[- \frac{2\,n^2\,\gamma^2}{\sum_{i=1}^{n} \,[\,2\,(\frac{1}{2}\,M_d^2\,)\,]^2} \right]$$

$$=\; 2 \;\exp\; \left[- \frac{2\,n\,\gamma^2}{{M_d}^{4}} \right] \tag{42}$$

The union bounds can be achieved as following:

$$\begin{aligned}
P\,(B_2') \;&= P \left\{ \left\| \frac{1}{4n} \, \Xi^T \, W \right\|_{\infty} \leq \gamma \right\} \\
&= P\left(\max_k \; \max_j \; \frac{1}{n} \; |\sum_{i=1}^{n} T_{ijk} \,| \;\leq\; \gamma \right) \\
&= P\left(\cap_{k=1}^{d} \; \cap_{j=1}^{d} \; \left\{ \frac{1}{n} \; |\sum_{i=1}^{n} T_{ijk} \,| \;\leq\; \gamma \right\} \right) \\
&= 1 - P\left(\cup_{k=1}^{d} \; \cup_{j=1}^{d} \; \left\{ \frac{1}{n} \; |\sum_{i=1}^{n} T_{ijk} \,| \;>\; \gamma \right\} \right) \\
&\geq 1 - \sum_{k=1}^{d} \sum_{j=1}^{d} P\left(\frac{1}{n} \; |\sum_{i=1}^{n} T_{ijk} \,| \;>\; \gamma \right)
\end{aligned}$$

by the result of (42),

$$\geq 1 - \exp\Big[\, - \frac{2\,n\,\gamma^2}{M_d^2} \; + \log\,(\,2\,d^2\,)\,\Big] \tag{43}$$

With

$$\gamma \;\geq\; M_d^2 \sqrt{\frac{\log\,(\,2\,d^2\,) + \tau}{2n}}\,,$$

it holds,

$$P\;(\;{B_2}'\;)\;\geq 1 - e^{-\tau}.$$

let

$$\gamma \;\geq\; M_d^2 \sqrt{\frac{\log\,(\,2\,d^2\,) + \log n}{2n}} \;\equiv \gamma^*,$$

then,

$$P\;(\,B_2'\,)\;\geq\; 1 - \frac{1}{n}.$$

Therefore, when $\lambda > \lambda^*$, $\gamma > \gamma^*$,

$$P\;(\;B\;)\;>\;1 - \frac{2}{n}.$$

$\square$

B2. Proof of Theorem 3.2

[Theorem 3.2]

Under Assumption $C_1 - C_5$, when $\lambda > \lambda^*$, $\gamma > \gamma^*$, with probability at least $1 - \frac{2}{n}$, it holds:

$$\mathcal{E}(\hat{\beta}_{MHCS}) \leq \left(3\,\lambda + 2\,\gamma \parallel \beta^* \parallel_1 + \delta_0 \right) \parallel \beta^* - \hat{\beta}_{MHCS} \parallel_1 \, + \, \delta_0\, a_0\,.$$

Proof of Theorem 3.2.
Define

$$V_{\lambda,\gamma} = \sup_{\hat{\beta} \in \mathcal{B}\,(\,\lambda,\,\gamma\,)} \frac{(\,L_n - L\,)\,(\,\rho_{\beta^*} - \rho_{\hat{\beta}}\,)\,(\,Z,\,Y\,)}{a_0 + \parallel \hat{\beta} - \beta^* \parallel_1} \tag{44}$$

where

$$\mathcal{B}\,(\,\lambda,\,\gamma\,) := \left\{ \,\hat{\beta} \;\in\; R^d :\; \hat{\beta} = \underset{\hat{\beta} \in \mathcal{C}\,(\,\lambda\,,\gamma\,)}{\arg\min} \;\parallel \hat{\beta} \parallel_1 \right\} \tag{45}$$

Similar to Theorem 2.2, we patition $\mathcal{B}_{\,(\,\lambda,\,\gamma\,)}$ into $\{\mathcal{B}_j\}_{j=0}^J$:

$$\mathcal{B}_0 = \{\hat{\beta} : \hat{\beta} \in \mathcal{B}_{(\lambda,\gamma)}, \|\hat{\beta} - \beta^*\|_1 \leq a_0\}$$

$$\mathcal{B}_j = \{\hat{\beta} : \hat{\beta} \in \mathcal{B}_{(\lambda,\gamma)}, a_{j-1} < \|\hat{\beta} - \beta^*\|_1 \leq a_j\}; (1 \leq j \leq J-1)$$

$$\mathcal{B}_J = \{\hat{\beta} : \hat{\beta} \in \mathcal{B}_{(\lambda,\gamma)}, \|\hat{\beta} - \beta^*\|_1 > a_{J-1}\} \tag{46}$$

For $1 \leq j \leq J-1$:

$$a_j = 2a_{j-1};$$

by Assumption C_5, it holds:

$$a_{J-1} \geq 2\,\|\,\beta^*\|_1 \qquad and \qquad a_0 \geq \frac{\|\beta^*\|_1}{2^J};$$

Then, we can derive the bound according to this partition $\mathcal{B}_{(\lambda,\gamma)}$ as follow:

$$P\,(\,V_{\lambda,\gamma} > \delta_0\,) = P\,\Big(\sup_{\hat{\beta}\,\in\,\mathcal{B}_{(\lambda,\gamma)}} \frac{(\,L_n - L\,)\,(\,\rho_{\beta^*} - \rho_{\hat{\beta}}\,)}{a_0 + \|\hat{\beta} - \beta^*\|_1} > \delta_0\,\Big)$$

$$\leq \sum_{j=0}^{J} P\,\Big(\sup_{\hat{\beta}\,\in\,\mathcal{B}_j} \frac{(\,L_n - L\,)\,(\,\rho_{\beta^*} - \rho_{\hat{\beta}}\,)}{a_0 + \|\hat{\beta} - \beta^*\|_1} > \delta_0\,\Big). \tag{47}$$

to be simplified, let

$$V_j = \sup_{\hat{\beta}\,\in\,\mathcal{B}_j} \frac{(\,L_n - L\,)\,(\,\rho_{\beta^*} - \rho_{\hat{\beta}}\,)}{a_0 + \|\hat{\beta} - \beta^*\|_1}(Z, Y) \tag{48}$$

then (47) is equivalent to

$$P\,(\,V_{\lambda,\gamma} > \delta_0\,) \leq \sum_{j=0}^{J} P\,\Big(V_j > \delta_0\,\Big).$$

According to Lemma 5.7: For $0 \leq j \leq J - 1$

$$P\,\Big(V_j > \delta_0\,\Big) < 2\,e^{-2n}$$

By Theorem 3.1, when $\lambda > \lambda^*$ and $\gamma > \gamma^*$, it holds

$$P\,\Big(V_J > \delta_0\,\Big) \leq P\,\Big(\,\|\hat{\beta}{-}\beta^*\|_1 > a_{J-1}\Big) \leq P\,\Big(\,\|\hat{\beta}{-}\beta^*\|_1 > 2\,\|\beta^*\|_1\Big) \leq P\,(\,\beta^* \notin \mathcal{C}_{\lambda,\gamma}\,)\;\leq\;\frac{2}{n}$$

Thus,

$$P\,\Big(V_{\lambda,\gamma} > \delta_0\,\Big) < 2J\,e^{-2n} + \frac{2}{n}.$$

It comes to conclude that, with probability at least $1 - 2J\,e^{-2n} - \frac{2}{n}$, it holds:

$$V_{\lambda,\gamma} := \sup_{\hat{\beta}\in\mathcal{B}_{(\lambda,\gamma)}} \frac{|(\,L_n - L\,)\,(\,\rho_{\beta^*} - \rho_{\hat{\beta}}\,)(Z, Y)|}{a_0 + \|\beta^* - \hat{\beta}\|_1}\;<\delta_0.$$

Since our estimator $\hat{\beta}_{MHCS} \in \mathcal{B}_{(\lambda,\gamma)}$, it holds

$$\frac{|\,(L_n - L\,)\,(\,\rho_{\beta^*} - \rho_{\hat{\beta}_{MHCS}}\,)\,(\,Z,\,Y\,)\,|}{a_0 + \|\hat{\beta}_{MHCS} - \beta^*\|_1} \;\leq\; \sup_{\hat{\beta}\,\in\,\mathcal{B}(\lambda,\gamma)} \frac{|(L_n - L)\,(\,\rho_{\beta^*} - \rho_{\hat{\beta}}\,)\,(\,Z,\,Y\,)\,|}{a_0 + \|\hat{\beta} - \beta^*\|_1} \leq \delta_0$$

thus,

$$L_n\,(\,\rho_{\beta^*} - \rho_{\hat{\beta}_{MHCS}}\,) - L\,(\,\rho_{\beta^*} - \rho_{\hat{\beta}_{MHCS}}\,) \leq \;\delta_0\,a_0\; + \delta_0\,\|\,\beta^* - \hat{\beta}_{MHCS}\,\|_1\,.$$

rearrange the orders, then

$$\mathcal{E}(\hat{\beta}_{MHCS}) = L\,(\,\rho_{\hat{\beta}_{MHCS}} - \rho_{\beta^*}\,) \leq L_n\,(\,\rho_{\hat{\beta}_{MHCS}} - \rho_{\beta^*}\,) + \;\delta_0\,\|\,\beta^* - \hat{\beta}_{MHCS}\,\|_1\,+\,\delta_0\,a_0$$

since:

$$\Big|\,L_n\rho_{\hat{\beta}_{MHCS}}\,(\,Z,\,Y\,) - L_n\rho_{\beta^*}\,(\,Z,\,Y\,)\,\Big| \leq \;\|\,\nabla_\beta\,L_n\,\rho_{\hat{\beta}_{MHCS}}(\,Z,Y\,)\,\|_\infty\,\|\,\hat{\beta}_{MHCS} - \beta^*\,\|_1$$

by Lemma 5.8

$$\leq (\,3\,\lambda + 2\,\gamma\,\|\,\hat{\beta}_{MHCS}\,\|_1\,)\,\|\,\hat{\beta}_{MHCS} - \beta^*\,\|_1;$$

then,

$$\mathcal{E}(\hat{\beta}_{MHCS}) \;= L\rho_{\hat{\beta}_{MHCS}}(\,Z,\,Y\,) - L\rho_{\beta^*}(\,Z,\,Y\,)$$

$$\leq \big|L_n\rho_{\hat{\beta}_{MHCS}}(\,Z,\,Y\,) - L_n\rho_{\beta^*}(\,Z,\,Y\,)\big| + \delta_0\,\|\,\beta^* - \hat{\beta}_{MHCS}\,\|_1 + \delta_0\,a_0$$

$$\leq (\,3\,\lambda + 2\,\gamma\,\|\,\hat{\beta}_{MHCS}\,\|_1\,)\,\|\,\beta^* - \hat{\beta}_{MHCS}\,\|_1 + \delta_0\,\|\,\beta^* - \hat{\beta}_{MHCS}\,\|_1 + \delta_0\,a_0\,;$$

Under $\beta^* \in \mathcal{C}_{\lambda,\gamma},\; \|\,\hat{\beta}_{MHCS}\,\|_1 \leq \|\,\beta^*\,\|_1$;

$$\mathcal{E}(\hat{\beta}_{MHCS}) \leq \;\big(\,3\,\lambda + 2\,\gamma\,\|\,\beta^*\,\|_1 + \delta_0\,\big)\,\|\,\beta^* - \hat{\beta}_{MHCS}\,\|_1\;+\;\delta_0\,a_0\,.$$

□

Lemma 5.7. *As $\mathcal{B}_j$, V_j, defined in (46), (48), for $0 \leq j \leq J-1$, it holds:*

$$P\,(\,V_j > \delta_0\,) < 2\,e^{-2n}.$$

Proof. Analogous to Theorem 2, we have following results for V_j:

$$P(\ |\ V_j - E\,(\,V_j\,)\,|\ \ > \delta_1\ ) < 2\,e^{-2n};$$

and

$$E(\ V_j\ )\ \le\ \delta_2;$$

set $\delta_0 = \delta_1 + \delta_2$, then

$$P(\ V_j >\ \ \delta_0\ ) < 2\,e^{-2n}.$$

□

Lemma 5.8.

With same notations in Theorem 3.2, when $\lambda > \lambda^*$ and $\gamma > \gamma^*$, then with probability

at least $1 - \dfrac{2}{n}$, *it holds:*

$$\frac{1}{n}\left\|\ Z^T\,[\ Y - \mu\,(\,Z\hat{\beta}_{MHCS}\,)\ ]\ \right\|_\infty\ \le\ 3\,\lambda + 2\,\gamma\ \|\ \hat{\beta}_{MHCS}\ \|_1\ .$$

Proof.

By triangle inequality,

$$\frac{1}{n}\left\| Z^T [\, Y - \mu\,(\, Z\hat{\beta}_{MHCS}\,)\,] \right\|_\infty - \frac{1}{n}\left\| \Xi^T [\, Y - \mu\,(\, Z\hat{\beta}_{MHCS}\,)\,] \right\|_\infty$$

$$; \qquad -\frac{1}{n}\left\| W^T \mu'\,(\,\xi\,\hat{\beta}_{MHCS}\,)(\,\Xi\hat{\beta}_{MHCS}\,) \right\|_\infty$$

$$\leq \frac{1}{n}\left\| Z^T [\, Y - \mu\,(\, Z\hat{\beta}_{MHCS}\,)\,] \,+\, \Xi^T [\, Y - \mu\,(\, Z\hat{\beta}_{MHCS}\,)\,] -\, W^T \mu'\,(\,\xi\,\hat{\beta}_{MHCS}\,)\,(\,\Xi\hat{\beta}_{MHCS}\,) \right\|_\infty$$

$$= \frac{1}{n}\left\| W^T [\, Y - \mu\,(\, W\hat{\beta}_{MHCS}\,)\,] \right\|_\infty .$$

Thus, after rearranging orders, the gradient of target population through high confidence set estimation, would be bounded by the following three parts.

$$\frac{1}{n}\left\| Z^T [\, Y - \mu\,(\, Z\hat{\beta}_{MHCS}\,)\,] \right\|_\infty \leq \frac{1}{n}\left\| W^T [\, Y - \mu\,(\, W\hat{\beta}_{MHCS}\,)\,] \right\|_\infty$$

$$+ \frac{1}{n}\left\| \Xi^T [\, Y - \mu\,(\, Z\hat{\beta}_{MHCS}\,)\,] \right\|_\infty + \frac{1}{n}\left\| W^T \mu'\,(\,\xi\,\hat{\beta}_{MHCS}\,)\,(\,\Xi\hat{\beta}_{MHCS}\,) \right\|_\infty$$

First, since $\hat{\beta}_{MHCS} \in \mathcal{C}_{\,(\,\lambda,\,\gamma\,)}$, by definition of $C_{(\lambda,\gamma)}$, it holds,

$$\frac{1}{n}\, \|\, W^T [\, Y - \mu\,(\, W\hat{\beta}_{MHCS}\,)\,]\, \|_\infty \,\leq\, \lambda + \,\gamma\, \|\, \hat{\beta}_{MHCS}\, \|_1; \tag{49}$$

Similar to the process in proving previous theorem $Event\ B_1$ and $Event\ B_2$, under the condition $\lambda \geq \lambda^*$ and $\gamma \geq \gamma^*$, the second part and third part will be bounded with high probability:

$$P\left(\frac{1}{n}\left\|\,\Xi^T\,[\,Y-\mu\,(\,Z\hat{\beta}_{MHCS}\,)\,]\,\right\|_\infty \leq 2\,\lambda\right) > 1-\frac{1}{n}; \tag{50}$$

$$P\left(\frac{1}{n}\left\|\,W^T\,\mu'\,(\,\xi\,\hat{\beta}_{MHCS}\,)\,(\,\Xi\hat{\beta}_{MHCS}\,)\,\right\|_\infty \leq \gamma\,\|\,\hat{\beta}_{MHCS}\,\|_1\right) > 1-\frac{1}{n}; \tag{51}$$

Thus by De Morgan's Law again,

$$P\left(\frac{1}{n}\left\|\,Z^T\,[\,Y-\mu\,(\,Z\hat{\beta}_{MHCS}\,)\,]\,\right\|_\infty \leq 3\,\lambda+2\,\gamma\,\|\,\hat{\beta}_{MHCS}\,\|_1\right) \geq 1-\frac{2}{n}.$$

□

B3. Proof of Theorem 3.3

Theorem 3.3:

Under AssumptionC_1 - AssumptionC_7, when $\lambda \geq \lambda^*$ and $\gamma \geq \gamma^*$, with probability at least $1 - \dfrac{2}{n}$, it holds:

$$(i) \quad \| \hat{\beta}_{MHCS} - \beta^* \|_2 \leq \frac{4\,(\lambda + \gamma \| \beta^* \|_1)\sqrt{s}}{\kappa};$$

$$(ii) \quad \| \hat{\beta}_{MHCS} - \beta^* \|_1 \leq \frac{8\,(\lambda + \gamma \| \beta^* \|_1)\, s}{\kappa}.$$

proof of Theorem 3.3.

By the definition of $\mathcal{C}_{(\lambda,\gamma)}$, it holds:

$$\| \nabla_\beta L_n \rho_{\hat{\beta}_{MHCS}} (W, Y) \|_\infty \leq \lambda + \gamma \| \hat{\beta}_{MHCS} \|_1;$$

Condition on $\beta^* \in \mathcal{C}_{(\lambda,\gamma)}$,

$$\| \nabla_\beta L_n \rho_{\beta^*} (W, Y) \|_\infty \leq \lambda + \gamma \| \beta^* \|_1;$$

then by the triangle inequality,

$$\| \nabla_\beta L_n \rho_{\hat{\beta}_{MHCS}} (W, Y) - \nabla_\beta L_n \rho_{\beta^*} (W, Y) \|_\infty$$

$$\leq \| \nabla_\beta L_n \rho_{\hat{\beta}_{MHCS}} (W, Y) \|_\infty + \| \nabla_\beta L_n \rho_{\beta^*} (W, Y) \|_\infty$$

$$\leq 2\lambda + \gamma \| \hat{\beta}_{MHCS} \|_1 + \gamma \| \beta^* \|_1 \leq 2\lambda + 2\gamma \| \beta^* \|_1 \tag{52}$$

Thus, the corresponding result of (36) is:

$$\delta\, L_n\, \rho_{(\,\hat{\Delta},\beta^*\,)}(\,W,\,Y\,) \;:=\; L_n\, \rho_{(\,\beta^*+\hat{\Delta}\,)}(\,W,\,Y\,) - L_n\, \rho_{\beta^*}(\,W,\,Y\,) - \left\langle\, \nabla_\beta\, L_n\, \rho_{\beta^*}(\,W,\,Y\,),\; \hat{\Delta}\,\right\rangle$$

$$\leq \|\; \nabla_\beta\, L_n\, \rho_{\hat{\beta}_{MHCS}}\;(\,W,\,Y\,) \;-\; \nabla_\beta\, L_n\, \rho_{\beta^*}\;(\,W,\,Y\,)\,\|_\infty \;\;\|\,\hat{\Delta}\,\|_1$$

$$\leq\; 2\,(\,\lambda + \gamma\,\|\,\beta^*\,\|_1\,)\,\|\,\hat{\Delta}\,\|_1 \tag{53}$$

Under $Event\ B$, $\|\hat{\beta}_{MHCS}\|_1 \leq \|\beta^*\|_1$, thus, (37) and (38) still valid.

By $Assumption\ C_7$,

$$\delta\, L_n\, \rho_{(\,\hat{\Delta},\beta^*\,)}\,(\,W,\,Y\,) \;\geq\; \kappa\,\|\,\hat{\Delta}\,\|_2^2;$$

combine with (53) and (38) we have,

$$\kappa\,\|\,\hat{\Delta}\,\|_2^2 \;\leq\; 2\,(\,\lambda + \gamma\,\|\,\beta^*\,\|_1\,)\,\|\,\hat{\Delta}\,\|_1 \;\leq\; 4\,(\,\lambda + \gamma\,\|\,\beta^*\,\|_1\,)\,\sqrt{s}\;\|\,\hat{\Delta}\,\|_2;$$

therefore,

$$\|\,\hat{\Delta}\,\|_2 \;\leq\; \frac{4\,(\,\lambda + \gamma\,\|\,\beta^*\,\|_1\,)\,\sqrt{s}}{\kappa}; \tag{54}$$

plug (54) into (38), we have

$$\|\,\hat{\Delta}\,\|_1 \;\leq\; \frac{8\,(\,\lambda + \gamma\,\|\,\beta^*\,\|_1\,)\,s}{\kappa}. \tag{55}$$

$\square$

[Corollary] Under Assumption C_1-C_7, when $\lambda > \lambda^*$, $\gamma > \gamma^*$, with probability at least $1 - 2J\, e^{-2n} - \frac{2}{n}$, it holds that:

$$\mathcal{E}(\hat{\beta}_{MHCS}) \leq \frac{8s(\,\lambda + \gamma\|\beta^*\|_1\,)\,(3\,\lambda\, + 2\gamma\|\beta^*\|_1 + \delta_0)}{\kappa} + \delta_0\; a_0$$

Proof. Take the result of (55) into Theorem 3.2, the result can be achieved. □

C Code

https://github.com/firfre/high-confidence-set