Statistical Inference for High-Dimensional Regularized Huber Regression

by

Tagbo Innocent Aroh

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 6, 2023

Keywords: robust regression, selective inference, high dimension, lasso, outlier detection, randomization

Approved by

Asheber Abebe, Professor of Mathematics & Statistics
Elvan Ceyhan, Associate Professor of Mathematics & Statistics
Mark Carpenter, Professor of Mathematics & Statistics
Peng Zeng, Chair, Associate Professor of Mathematics & Statistics

Abstract

With the rapid advancement of technology, the amount of available data for extracting interesting insights and meaningful patterns has grown exponentially, resulting in a significant increase in the dimensionality of datasets. However, high-dimensional data can be easily contaminated by outliers or errors with heavy-tailed distributions, rendering many conventional methods inadequate for analysis. Consequently, there has been a growing interest in applying robust methods to analyze high-dimensional data, with Huber regression with regularization being a popular choice. Existing robust methods are primarily used for parameter estimation and variable selection, and there has been a lack of tools for statistical inference in high dimensions. To overcome this challenge, researchers have incorporated techniques such as lasso in statistical inference. Specifically, they have used such shrinkage penalty as a tool for variable selection and applied ordinary least squares on the selected variables to construct confidence intervals and p-values. However, this approach results in statistical inference that is not valid because it fails to account for all the variability in the selection process. The generalized lasso problem is one of the most commonly used convex optimization problems, therefore, in this dissertation, I will focus on developing conditional statistical inferential tools in high dimensions using Huber regression with a generalized lasso as the regularization term (gl-huber). To address this problem, I will follow a framework that characterizes the distribution of a post-selection estimator that is conditioned on the selection process.

Specifically, I will characterize the conditional distribution of the gl-huber post-selection estimator while conditioning on both variable selection and outlier identification events by first demonstrating that the event of variable selection and outlier detection can be represented as an affine constraint in the response variable y (a polyhedron). Using this approach, I will then show that the conditional distribution of a linear combination of responses is a univariate truncated normal distribution in cases where the random error is normal. In cases where the distribution of random error is not normal, I will show that the asymptotic distribution is still truncated normal under certain weak conditions. This will enable the development of valid post-selection conditional p-values and confidence intervals that account for the variability in the selection process and satisfy all necessary frequency properties. To further improve the procedure's performance, I propose incorporating randomized responses. To validate the efficacy of the proposed methods, both theoretical properties and computational algorithms are investigated, and their practical utility is demonstrated through a range of simulation and real-world examples.

Acknowledgments

I owe my successes to several individuals who have played an integral role in shaping my academic and personal growth. First and foremost, I am deeply grateful to the almighty God for blessing me with the gift of life and good health, which has enabled me to pursue my academic goals with passion and perseverance.

I want to express my sincere appreciation to my wonderful supervisor, Dr. Peng Zeng, whose guidance, support, and encouragement have been invaluable throughout my graduate studies. Dr. Zeng's unwavering belief in my potential, and his commitment to providing me with the necessary resources and mentorship, has been instrumental in helping me reach this milestone. Furthermore, I would like to thank the members of my committee, Dr. Asheber Abebe, Dr. Elvan Ceyhan, and Dr. Mark Carpenter, for their generous support and guidance. Their insightful feedback and academic expertise have been crucial in shaping my research and pushing me to grow as a scholar. I would like to extend my heartfelt gratitude to Dr. Yang Zhou, who served as my university reader.

To my beloved wife, Oluchi, and our daughter, Jachimma, I am immensely grateful for your constant love, understanding, and encouragement. Your presence in my life has been a source of strength, inspiration, and motivation, and I cannot overstate the role you have played in my success throughout this journey. To my parents, parents-in-law, siblings, siblings-in-law, and my friends, thank you for your prayers and support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we will introduce the problem discussed in this dissertation. We will discuss the motivation of my research, review some literature, and explain my contribution.

Consider a linear regression setup, with an outcome vector $y \in \mathbb{R}^n$ and a matrix of predictor variables $X \in \mathbb{R}^{n \times p}$ related by

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 \mathbb{I}_{n \times n}) \tag{1.1}$$

High-dimensional data analysis ($p > n$) often seeks to identify a subset of important features, and further assess the effects of these identified features on the outcome variable. Traditional statistical inference procedures based on standard regression methods often fail in the presence of high-dimensional features. Recently, regularization methods have quite shown to be promising tools for analyzing high-dimensional data, for example, the lasso (Tibshirani (1996)), the adaptive lasso (Zou (2006)), the Elastic Net (Zou and Hastie (2005), Zou and Zhang (2009)), one-step local linear approximation (Zou and Li (2008)), SCAD (Fan and Li (2001)), etc. These methods simultaneously identify informative variables and produce stable coefficient estimates for the selected variables to induce a model for prediction (variable selection and estimation). Although these regularization methods are very effective for variable selection and stable parameter estimation, but they yield estimators whose

sampling distribution is difficult to obtain, hence constructing interval estimators for the regression parameters will be difficult in finite samples. Take for instance the lasso, lasso-type estimators have a nonstandard limiting distribution that depends on which components of the coefficient vector are 0. Because the lasso estimator does not have explicit solution, the limiting distribution cannot be estimated directly. Furthermore, standard bootstrap methods fail when the true coefficient vector is sparse (Fu and Knight (2000)). Over the years, these regularization methods have been mainly studied under the following criteria;

- the correct recovery of the support set $S = \{j \in \{1, 2, \dots p\} : \beta_j \neq 0\}$ of the model coefficients β for a linear model (1.1)
- l_q estimation errors $||\hat{\beta} - \beta||_q^q$, especially l_1 & l_2 where $\hat{\beta}$ is the estimate of β
- prediction error $||X\hat{\beta} - X\beta||_2^2$

but there are never enough discussions on inferential tools (confidence intervals and p-values) for these regularization methods until recently. Although most of the theoretical work on high-dimensional linear models focuses on consistency, recently, Lee et al. (2016) and Tibshirani et al. (2016) described a general scheme to perform valid inference after any selection event (like the lasso, stepwise selection, etc) that can be characterized as the response y falling into a polyhedral set. This framework can be used to conduct post selection inference while conditioning on the variable selection event. Another class of approaches are by Zhang and Zhang (2014), Van de Geer et al. (2014), and Javanmard and Montanari (2014b), these approaches are based on debiasing or denoising a regularized regression estimator, like the lasso. We will dive deep later in this chapter via an extensive literature review to compare

and contrast these various frameworks.

In this dissertation, we will discuss robust techniques like Huber loss for handling outliers in high dimension, and we seek to establish inferential tools (confidence intervals) for the Huber loss function with a generalized regularization term while conditioning on the outlier identification event and the variable selection event. The idea is to condition any inferences on the components of the data used to generate the hypotheses, thus preventing information in those components from being used again. We will now give more context to the terminologies and ideas mentioned above by discussing the problem of regression in high dimension in section 1.1, some regularization methods and their theorietical results in section 1.2, some robust regression techniques for errors with heavy tails (outliers) in section 1.4, then we discuss the two broad frameworks of inference in high dimensional regression models in section 1.3, before we narrow down to the problem we are trying to consider in section 1.5, alongside the philosophical discussions of our approach.

1.1 High Dimensionality Problem

In the past couple of decades, data acquisition technologies have rapidly evolved, enabling devices to gather vast amounts of data simultaneously. As a result, high-dimensional data has emerged, where the number of features can surpass the number of observations. For example, gene expression studies using micro-arrays can contain hundreds of samples, each with tens of thousands of genes, leading to millions of possible gene combinations for one individual. Similarly, other fields such as finance, high-resolution imaging, and website analysis generate

large-scale datasets. While having access to massive data may seem like an advantage, analyzing high-dimensional data presents significant challenges. Distinguishing meaningful signals from noise is often a formidable task, and traditional statistical techniques developed in the 20^{th} century may not be suitable for such scenarios. Classical statistical inference, for instance, is often not efficient for high-dimensional problems, and standard methods like ordinary least-squares fitting of a linear model with more predictors than observations can be ill-posed, i.e., let $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$, the least squares coefficients can be defined as the solution of the optimization problem

$$\min_{\beta} \frac{1}{2} ||y - X\beta||_2^2$$

If rank$(X) = p$, i.e, the predictors are linearly independent, and the null space of X, $null(X)$ contains only the zero vector, and $X^T X$ is positive definite which implies invertibility, then the above least squares problem has a unique solution, which is $\hat{\beta} = (X^T X)^{-1} X^T y$. Now, when $p > n$, which implies that rank $(X) < p$, the least squares problem will have infinitely many solutions, i.e., given one solution $\hat{\beta}$, the quantity $\hat{\beta} + \gamma$ is also a solution for any $\gamma \in null(X)$, since the $dim(null(X)) = p - n$, therefore interpretation will almost be impossible. Also, supposing OLS estimates are obtained in the above case, they will be poor (low bias & high variance). Most of the classical statistical theory provides results for the asymptotic setting where the number of parameters is fixed and the sample size goes to infinity. This asymptotic theory is very useful for analyzing data for large n while p is small and fixed, but it can give misleading results for modern high-dimensional data, this can clearly be seen in Portnoy (1986). Analyzing "large p" data therefore requires some new inferential

tools. In a way of summary, let's say that traditional methods and theory are not applicable or computationally infeasible to handle these high dimensional data, hence there is a great appeal for novel statistical models and methods and it has given rise to a huge effort from the statistical and data analyst community for developing new tools. Assuming certain notions of **sparsity**, there has been a revolution of methodological, computational and mathematical advances which allow for high-dimensional statistical inference. For example, the sparsity assumption that the health status of a person is depending only on a few among several thousands of biomarkers appears much more realistic than considering a model where all the thousands of variables would contribute in a smooth way to the state of health, see Bühlmann and Van De Geer (2011). Else where, other researchers have tried to reduce the dimension of the data matrix using some of popular dimension reduction techniques, prior to the regression modelling. It is safe to discuss that high dimension data for supervised learning (regression & classification methods), or unsupervised learning (clustering, etc), and without any form of further assumptions (like penalization from the field of machine learning which have proven to be more flexible), will still be ill posed.

1.2 Regularized Regression Methods

Irrespective of the linear model assumption, linear regression has some short comings like, *predictive ability* - the fit often has low bias but high variance when there are too many predictors, etc. In a high dimensional setting where the number of predictors p exceeds the number of observations n, these short comings become a major problem. As a matter of fact, in such settings, the linear regression estimate is actually not well-defined. However, penalized regression models are seemingly more adaptable to high-dimensional data compared

to traditional statistical regression approaches for estimating linear regression parameters. Also, an important topic in linear regression analysis is variable selection, variable selection is particularly important when the true underlying model has sparse representation. By shrinking estimates to zero, regularization can reduce the variability in estimates of regression coefficients, thereby improving the predictive error. Some classical examples are:

Consider $||\beta||_0 = \sum_{j=1}^{p} 1\{\beta_j \neq 0\}, \quad ||\beta||_1 = \sum_{j=1}^{p} |\beta_j|, \quad ||\beta||_2 = (\sum_{j=1}^{p} \beta_j^2)^{1/2}$

We put the methods in constrained form as below, where $k,\ t \geq 0$ are tuning parameters.

$$\min_{\beta \in \mathbb{R}^p} ||y - X\beta||_2^2 \quad subject\ to \quad ||\beta||_0 \leq k \qquad (Best\ subset\ selection) \qquad (1.2)$$

$$\min_{\beta \in \mathbb{R}^p} ||y - X\beta||_2^2 \quad subject\ to \quad ||\beta||_1 \leq t \qquad (lasso\ regression) \qquad (1.3)$$

$$\min_{\beta \in \mathbb{R}^p} ||y - X\beta||_2^2 \quad subject\ to \quad ||\beta||_2 \leq t \qquad (Ridge\ regression) \qquad (1.4)$$

The above problems can be formulated in penalized form as follows:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_0 \qquad (Best\ subset\ selection) \qquad (1.5)$$

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}||y - X\beta|_2^2 + \lambda||\beta||_1 \qquad (lasso\ regression) \qquad (1.6)$$

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_2 \qquad (Ridge\ regression) \qquad (1.7)$$

where $\lambda \geq 0$ is the regularization parameter. (1.4) and (1.7) are equivalent, i.e., for any $\lambda \geq 0$ and a solution $\hat{\beta}$ in (1.7), there is a corresponding value $t \geq 0$ such that $\hat{\beta}$ solves (1.4) and vice versa, likewise (1.2) and (1.5), (1.3) and (1.6). Because of the wonderful theory of convex duality and optimality, any local mininmizer is a global minimizer. Amongst other

things, convex optimization problems appear to be more interesting. The lasso and ridge regression problems are convex optimization problems. Best subset selection is no way near being convex. The lasso regression and best subset selection induces sparsity, ridge regression doesn't. The ridge regression is always strictly convex, and hence will always have a unique solution, while the lasso regression is not always strictly convex, therefore it need not have a unique solution, although this can be mitigated by using the elastic net Zou and Hastie (2005). Elastic net applies a penalty of the form $\lambda_1||\beta||_1 + \lambda_2||\beta||_2^2$.

1.2.0.1 lasso

As in (1.6), because of the penalty term $||\beta||_1$, the lasso solution is usually sparse when a large λ is used, the lasso method is a shrinkage approach that avoids over-fitting, identifies the true signals (informative predictors) from a pool of candidate variables, with low variability and an increase in bias, the lasso estimates model parameters, i.e., it does both parameter estimation and variable selection simultaneously. λ can be chosen using cross validation. If X has columns in general position, then we'll have uniqueness of the lasso solution, (Tibshirani, 2013). Since this is a convex optimization problem, it can easily be solved using optimization packages like **CVXR**, **Gurobi**, etc. However, Efron et al. (2004) introduced an efficient algorithm called *Least Angle Regression (Lars)*, and this package is available in the R package for computing the entire path solution at a fair computational cost. lasso has some limitations:

- Because of the nature of convex optimization problem lasso tries to minimize, the lasso selects at most n variables in the case where the number of predictors is larger than the number of observations ($n << p$). This is called *sparsity limitation*.

- For most real world datasets, usually there is a group of variables among which the pairwise correlations are very high , then lasso tends to arbitrarily select only one variable from the group. This situation is not ideal, especially in gene selection problems, for example; the ideal gene selection method in gene expression is *group selection* which is eliminating the trivial genes and automatically include whole groups into the model once one gene among them is selected. Elastic net is better suited here.

- Though ridge regression won't help in feature selection and model interpretability is low, if there is high correlation between the predictors especially in high dimensional data, ridge regression has better prediction power than the lasso.

1.2.0.2 Theoretical Results for lasso

There is a large body of theoretical work on the behavior of the lasso (1.6). It is largely focused on consistency of the parameter estimates in l_2 or some other norm, prediction error consistency, and recovery of the nonzero support set of the true regression parameters, sometimes called sparsistency. For MSE consistency, if β and $\hat{\beta}$ are the true and lasso estimated parameters respectively, it can be shown that as $p, n \to \infty$

$$||X(\hat{\beta} - \beta)||_2^2/n \leq C||\beta||_1\sqrt{\log(p)/n}$$

with high probability (Bühlmann and Van De Geer (2011), Chapter 6). Hence the lasso is consistent for prediction. The result only assumes that the design X is fixed and has no other conditions on X. However, estimation error and correct support recovery requires

more stringent assumptions.

Bounds on lasso l_2-Error

We mention some conditions on the model matrix X that are needed to establish bounds on l_2-error for the lasso. The intuition behind these conditions can be explicitly found in Hastie et al. (2015) and the references therein. The condition we need here is strong convexity of the least-square loss, and the least-squares loss is strongly convex if and only if the eigenvalues of the $p \times p$ positive semidefinite matrix $X^T X$ are uniformly bounded away from zero. However, it is easy to see that any matrix of the form $X^T X$ has rank at most $\min\{n, p\}$, so it is always rank-deficient—and hence not strongly convex whenever $n < p$. So, we relax the strong convexity condition and require restricted strong convexity (see Bühlmann and Van De Geer (2011)), which is in turn equivalent to lower bounding the restricted eigenvalues of the model matrix for the case of linear regression, see Bühlmann and Van De Geer (2011). Suppose that the model matrix X satisfies the restricted eigen value bound with parameter $\gamma > 0$ and the regularization parameter is chosen as $\lambda_n \geq 2||X^T \varepsilon||_\infty / n > 0$, then an estimate $\hat{\beta}$ from the regularized lasso (1.6) satisfies the bound

$$||\hat{\beta} - \beta||_2 \leq \frac{3}{\gamma} \sqrt{s_0} \lambda_n$$

where $s_0 = |S_0|, \quad S_0 = \{j : \beta_j \neq 0,\ j = 1, ; \ldots, p\}$, see Hastie et al. (2015). For the case of classical linear Gaussian model, for which the observation noise $\varepsilon \in \mathbb{R}^n$ is Gaussian with i.i.d $\mathcal{N}(0, \sigma)$ entries, and the design matrix X is fixed. Then, the above bound reduces to

$$||\hat{\beta} - \beta||_2 \leq c\frac{3}{\gamma}\sqrt{\frac{\tau s_0 \log p}{n}}$$

with probability at least $1 - 2e^{-\frac{1}{2}(\tau-2)\log p}$, for some $\tau > 2$ and an appropriately chosen constant c. Negahban et al. (2012) made use of restricted strong convexity of the cost function and decomposability of the regularizer, and the authors provided a general framework for analyzing the estimation error $||\hat{\beta} - \beta||_2$ for the family of M-estimators, which includes lasso as a special case. For more theoretical results addressing error bounds for lasso estimates, the reader may consult any of the following (Bickel et al. (2009), Bunea et al. (2007a), Bunea et al. (2007b), Candes and Tao (2007), Meinshausen and Yu (2009), Van De Geer and Bühlmann (2009), Zhang and Huang (2008)).

More theoretical results on prediction error consistency for lasso can be found here (Bunea et al. (2007a), Greenshtein and Ritov (2004), Van De Geer and Bühlmann (2009), Zhang and Huang (2008)), and theoretical results for exactly recovery can be found here (Meinshausen and Bühlmann (2006), Wainwright (2009), Zhao and Yu (2006)).

1.2.0.3 Adaptive lasso

Zou (2006), and some references therein stated that the lasso does not have oracle properties (i.e., doesn't identify the right subset of true variables and doesn't have optimal estimation rate). They claimed that there are cases where a given λ that leads to optimal estimation rate ends up with inconsistent selection of variables. Also, there are cases with the right selection of variables but showing biased estimates for large coefficients, thereby leading to

sub-optimal prediction rates. For a suitable choice of λ, adaptive lasso is a refinement of the lasso that has the oracle properties. It has the same advantage as the lasso, i.e., it shrinks some of the coefficients to zero, thereby performing selection of variables with the regularization. Also, as regularization technique, adaptive lasso avoids over-fitting penalizing large coefficients. Adaptive lasso replaces the l_1 penalty by a re-weighted version.

$$\hat{\beta}_{adapt}(\lambda) = \underset{\beta}{\text{argmin}}(\frac{1}{n}||y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{init,j}|})$$

where $\hat{\beta}_{init}$ is an initial estimator and if $\hat{\beta}_{init,j} = 0 \implies \hat{\beta}_{adapt,j} = 0$. The lasso estimator can be used as an initial estimator.

The design matrix satisfying the neighborhood stability or irrepresentable condition is a necessary and sufficient condition for the lasso to achieve a consistent variable selection in a linear model. These condition(s) might be unrealistic in practice. For the adaptive lasso, assuming compatibility conditions on the design matrix are sufficient to achieve consistent variable selection, and these conditions are weaker than that of the lasso, hence the adaptive lasso selects the true set of nonzero coefficients with probability tending to one. See Bühlmann and Van De Geer (2011), Zou (2006) for an in-depth treatment of these terminologies.

1.2.0.4 Other Non-Convex Regularization Methods

It has been well known that convex penalties introduce non negligible estimation biases. To eliminate the estimation bias, a family of folded-concave penalties was introduced, which includes the smooth clipped absolute deviation (SCAD) Fan and Li (2001), minimax concave

penalty (MCP) Zhang (2010), etc. Compared to their convex counterparts, these non-convex penalties eliminate the estimation bias and attain more refined statistical rates of convergence. However, it is more challenging to analyze the theoretical properties of the resulting estimators due to non-convexity of the penalty functions.

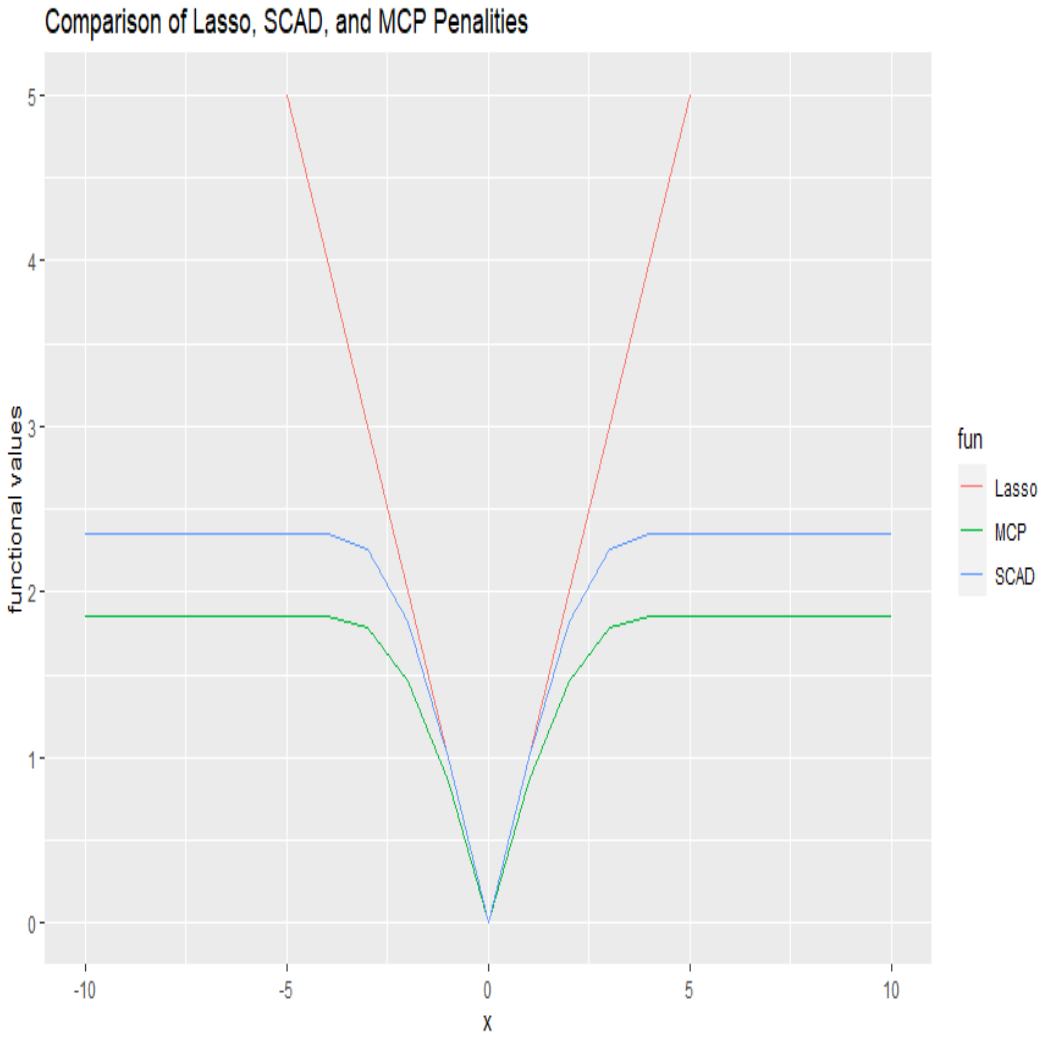


Figure 1.1: Plots of various Penalties

1.3 Inference for High-Dimensional Regression Models

It is widely recognized that modern statistical problems are increasingly high-dimensional, i.e., require estimation of more parameters than the number of observations/samples. Examples abound from signal processing to genomics, collaborative filtering and so on. A number of successful estimation techniques have been developed over the last the years to tackle these problems, but fitting high-dimensional statistical models often necessitates more complex computational algorithms, and explicit solutions for the estimate may not exist. As a consequence, it is generally impossible to obtain an exact characterization of the probability distribution of the parameter estimates. This in turn implies that it is extremely challenging to quantify the uncertainty associated with a certain parameter estimate. Concretely, there hasn't been enough literature on computing classical measures of uncertainty and statistical significance as confidence intervals or p-values for these high dimensional models. We now discuss two broad frameworks for statistical inference for high dimensional linear regression under Gaussian noise. The first framework is "Post - Selection Inference Methods" or "Selective Inference methods" as championed by Lee et al. (2016) and Tibshirani et al. (2016), the R package *selectiveInference* provides their implementation alongside other similar methods, see `https://cran.r-project.org/web/packages/selectiveInference/index.html`. The second framework is "High Dimensional Inference" as championed by Zhang and Zhang (2014) and Van de Geer et al. (2014), the R package *hdi* provides their implementation alongside other similar methods, see `https://cran.r-project.org/web/packages/hdi/`.

1.3.1 Post - Selection Inference

In classical statistics, a model and a corresponding set of parameters are assumed to be chosen independently of the data that is subsequently used for statistical inference. However, in practice, data analysts often examine the data to inform their model and parameter choices. Ignoring this adaptivity can lead to flawed conclusions and loss of inferential guarantees. While instructing analysts to avoid such exploration is not practical or recommended, it has led to an effort in the statistical community to develop tools for selective inference. This effort has been driven by a realization of the replication problem in science and a desire to address it. These tools aim to perform inference while accounting for the effect of data-dependent model selection and/or target parameter selection. The field of post selection inference provides a formal framework for addressing such issues by constructing valid statistical procedures that account for the adaptive nature of the inference process. One such approach is to condition on the event of selection, which can be described by a set of inequalities that constrain the selection event. In this way, one can adjust the p-values and confidence intervals to ensure that the resulting inferences are valid, even in the presence of data-dependent selection. The field of of post selection inference is still evolving, and there are many open questions and challenges in constructing and implementing valid selective inference procedures. However, the development of such methods is crucial to ensuring the reliability and reproducibility of scientific findings.

In recent years, there have been several studies on post-selection inference using various approaches and frameworks, including Berk et al. (2013), Fithian et al. (2014), and Lee and

Taylor (2014). Here, we will focus on a general approach to performing reliable inference after model selection, which was originally developed by Lee et al. (2016) and Tibshirani et al. (2016). Their approach characterizes the distribution of a post-selection estimator conditional on the variable selection event, where the selection event is hypothesized to partition the sample space into at least convex or polyhedral sets. This methodology can be applied to many widely used automatic model selection procedures, such as marginal screening, lasso, and forward-selection. In practical terms, let us consider a linear regression scenario with unknown coefficients to be estimated, represented by $\beta \in \mathbb{R}^p$. The proposed framework can accommodate any procedure for which the selection events can be characterized by a set of affine inequalities in the response variable, denoted by y. The selection event can be expressed as $Ay \leq b$, where A is a matrix and b is a vector that can be computed from the data. Then, the distribution of any estimator $\hat{\beta}$ that is selected based on this event can be represented as a truncated normal distribution with a mean and estimable covariance matrix. Explicit formulas for the matrix A and vector b can be found in Lee et al. (2016). This framework is also applicable to successive steps of the LAR algorithm and can provide a finite sample form of the covariance test. This methodology holds considerable practical significance due to its ability to address post-selection inference challenges as it can yield exact p-values and confidence intervals in the Gaussian case, that account for the uncertainty introduced by model selection. This method can also be applied to forward stepwise regression, and to the lasso at a fixed choice of the regularization parameter λ. Please refer to Chapter two for more details.

Post-selection inference has the advantage of not assuming that any of the candidate models is correct. Subsequently, Tian and Taylor (2017), Tibshirani et al. (2018), and Tibshirani et al. (2016) addressed issues related to post-selection inference: In these papers, confidence sets for the selected variables are considered to have a guaranteed coverage probability conditionally on the event that a particular model has been selected by the model selection procedure, hence False Coverage rate (FCR) is controlled at a nominal level of α.

1.3.2 High Dimensional Inference

For the lasso, in particular, a de-sparsifying method has recently been developed by Belloni et al. (2014), Van de Geer et al. (2014), and Zhang and Zhang (2014), these papers and the references therein are about confidence intervals and p-values for coefficients in high dimensional linear models based on the lasso estimator while controlling the resulting type I error, they do not address post-selection inference; their target is β, the coefficients in the true model (1.1), rather than $\hat{\beta}^E$, the coefficients in the selected model where $\beta^{\hat{E}} = \text{argmin}_\beta \, \mathbb{E}||y - X_E\beta||^2$ and E is the model selected by the lasso estimator. Although inference for β is appealing, it requires a lot of assumptions like, correctness of the linear model assumption, compatibility or incoherence-type assumption, etc, which for the most part might be unrealistic in practice.

Let us describe the method by Zhang and Zhang (2014), for high dimensional case in (1.1) with $p > n$, the idea is to pursue a regularized projection. Instead of ordinary least squares regression, we use lasso regression of the j^{th} column $X^{(j)}$ versus the remaining

columns $X^{(-j)}$, with corresponding residual vector $Z^{(j)}$: such penalized regression involves a regularization parameter λ_j for the lasso, and hence $Z^{(j)} = Z^{(j)}(\lambda_j)$. For any vector $Z^{(j)}$, we immediately obtain

$$\frac{Y^{\mathrm{T}} Z^{(j)}}{(X^{(j)})^{\mathrm{T}} Z^{(j)}} = \beta_j + \sum_{k \neq j} P_{jk}\beta_k + \frac{\varepsilon^{\mathrm{T}} Z^{(j)}}{(X^{(j)})^{\mathrm{T}} Z^{(j)}}, \qquad P_{jk} = (X^{(k)})^{\mathrm{T}} Z^{(j)}/(X^{(j)})^{\mathrm{T}} Z^{(j)} \tag{1.8}$$

Note that in the low-dimensional case with $Z^{(j)}$ being the residuals from ordinary least squares, due to orthogonality, $P_{jk} = 0$. When using the lasso-residuals for $Z^{(j)}$, we do not have exact orthogonality and a bias arises. Thus, we make a bias correction by plugging in the lasso estimator $\hat{\beta}$ (of the regression y versus X): the bias-corrected estimator is

$$\hat{b}_j = \frac{Y^{\mathrm{T}} Z^{(j)}}{(X^{(j)})^{\mathrm{T}} Z^{(j)}} - \sum_{k \neq j} P_{jk}\hat{\beta}_k$$

Using (1.8), we have

$$\sqrt{n}(\hat{b}_j - \beta_j) = \frac{n^{-1/2}\varepsilon^{\mathrm{T}} Z^{(j)}}{n^{-1}(X^{(j)})^{\mathrm{T}} Z^{(j)}} + \sum_{k \neq j} \sqrt{n} P_{jk}(\beta_k - \hat{\beta}_k)$$

The second term on the right is negligible under the following assumptions:

1. The design matrix X has compatibility constant (see Bühlmann and Van De Geer (2011), page 106) bounded away from zero, and the sparsity is $s_0 = \mathrm{o}(\sqrt{n}/\log(p))$. Where $s_0 = |S_0|, \quad S_0 = \{j : \beta_j \neq 0,\ j = 1, ; \ldots, p\}$

2. The rows of $\mathbf{X}$ are fixed realizations of i.i.d random vectors $\sim \mathcal{N}_p(0, \Sigma)$, and the minimal eigen value of Σ is bounded away from zero. Σ is the variance covariance matrix.

3. The inverse Σ^{-1} is row-sparse with $s_j = \sum_{k \neq j} \mathcal{I}((\Sigma^{-1})_{jk} \neq 0) = \mathrm{o}(n/\log(p))$

Van de Geer et al. (2014), and Zhang and Zhang (2014) then showed that for a linear model as in 1.1 with fixed design and Gaussian errors, assumming 1,2 & 3 above, we have;

$$\sqrt{n}\sigma_\varepsilon^{-1}(\hat{b} - \beta) = W + \Delta, \quad W \sim \mathcal{N}_p(0, \Omega), \quad \Omega_{jk} = \frac{n(Z^{(j)})^{\mathrm{T}} Z^{(k)}}{[(X^{(j)})^{\mathrm{T}} Z^{(j)}][(X^{(k)})^{\mathrm{T}} Z^{(k)}]}$$

$$||\Delta||_\infty = \mathrm{o}_p(1)$$

which asymptotically imples

$$\sigma_\varepsilon^{-1} \, \Omega_{jj}^{-1/2} \, \sqrt{n} \, (\hat{b}_j - \beta_j) \to \mathcal{N}(0, 1),$$

from which we can immediately construct a confidence interval or hypothesis test by plugging in any consistent estimator $\hat{\sigma_\varepsilon}$ of σ_ε.

Javanmard and Montanari (2014a) claimed that the methods of Zhang and Zhang (2014) can be sub optimal, because it requires the design to be generated from a population distribution whose inverse covariance matrix is sparse. The authors went ahead and constructed a de-biased estimator from the lasso solution and their approach applies to general covariance structures. Their de-biased estimator was given by the simple formula $\hat{\beta}^D = \hat{\beta} + (1/n) \, M \, X^T \, (Y - X\hat{\beta})$, where M is an estimator of Σ^{-1}. Their basic intuition was that $X^T \, (Y - X\hat{\beta})/(n\lambda)$ is a subgradient of the l_1 norm at the lasso $\hat{\beta}$. By adding a term proportional to the above subgradient, the procedure compensates the bias introduced by l_1 penalty in the lasso. The authors proved that $\hat{\beta}^D$ is approximately Gaussian, with mean β

and covariance $\sigma^2(M\hat{\Sigma}M)/n$, where $\hat{\Sigma} = (X^TX/n)$ is the empirical covariance of the feature vectors, this result allows to construct confidence and p-values in complete analogy with classical statistics procedures, i.e., letting $Q \equiv M\hat{\Sigma}M$, $[\hat{\beta}_i^D - 1.96\sigma\sqrt{Q_{ii}/n},\ \hat{\beta}_i^D + 1.96\sigma\sqrt{Q_{ii}/n}]$ is 95% confidence interval. The noise standard deviation can be replaced by an consistent estimator $\hat{\sigma}$. The matrix M was to primarily decorrelate the columns of X, the authors constructed M by solving a convex program that aims at optimizing two objectives, i.e., control the non-Gaussianity and bias of $\hat{\beta}^D$ by controlling $|M\hat{\Sigma} - I|_\infty$, and also control the variance of $\hat{\beta}_i^D$ by minimizing $[M\hat{\Sigma}M]_{ii}$ for each i.

Other methods for high dimensional inference here includes "Multi sample-splitting" by Meinshausen et al. (2009), "Ridge projection and bias correction" by Bühlmann (2013), etc. See Dezeure et al. (2015) and the references therein for an in-depth summary of these methods.

Summary Table for the Two Approaches		
*	High Dimensional Inference (hdi)	Exact Post Selection Inference
Main Idea	confidence intervals and p-values for β, i.e., full model inference	confidence intervals and p-values for $\hat{\beta}^E$, the coefficients in the selected model
Mathematical Tools	correctness of the linear model, compatibility condition on X and beta-min condition, sparsity assumption of β	*Affine selection procedure*, σ^2 is known, $y \sim \mathcal{N}(\mu(X), \sigma^2 I_n)$.

Pros	Computes robust unconditional full model inference (confidence intervals and p-values).	Does not require correctness of the linear model which might be unrealistic in practice.
Cons	Computationally expensive for large p, Ridge projection does not reach the asymptotic Cramér–Rao efficiency bound, desparsified lasso requires inverse covariance matrix to be sparse, and beta-min assumption is not ideal.	The strength of the signal affects the width of confidence intervals. Conditioning on both the model and coefficient signs reduces power, while conditioning only on the model increases statistical efficiency but reduces computational efficiency.

1.4 Robust Regression Methods

The ordinary least squares estimate for linear regression is sensitive to errors with large variance. It is not robust to heavy-tailed errors or outliers, which are commonly encountered in applications, especially in this era of big data, hence the OLS performance will be poor. Rousseeuw and Leroy (2005) explicitly defined some three types of outliers as;

- Vertical outliers: observations that are not outlying in the explanatory variables but have outlying values in the response variable. Their presence affects both OLS estimate, especially the estimate of the intercept.

- Good leverage points: These are observations that are located fairly close to the regression line and they are outlying in the explanatory variables. Their presence affects only the standard errors, but not the estimates.

- Bad leverage points: these are observations that are far from the true regression line, and are outlying in both explanatory variables and the response variable. They tremendously affect the OLS estimates (both the slope and the intercept).

We know that the breakdown point of the sample mean is almost 0 which is very low, hence in linear regression, the breakdown of the OLS estimator is analogous to the breakdown of the sample mean: a few extreme observations (vertical outliers) can largely determine the value of the OLS estimator, therefore researchers have considered many different loss functions to replace the squared error loss in least squares estimate to achieve robustness. Some of the proposed loss functions are, Huber's M-estimators Huber (1964), MM-estimators Yohai (1987), Least Median of Squares estimators and Least Trimmed Squares estimators Rousseeuw (1984), S-estimators Rousseeuw and Yohai (1984) and quantile regression methods Koenker and Bassett Jr (1978). Amongst all of those, the most common general method of robust regression is **M-estimation**, which is regarded as a generalization of the maximum likelihood estimation, hence the name **M**. Huber defined the 'plain vanilla' regression M-estimates as :

$$\hat{\beta} = \min_{\beta} \sum_{j=1}^{p} \rho(y_i - \beta^T x_i)$$

or after taking derivatives:

$$\sum_{j=1}^{p} \psi(y_i - \hat{\beta}^T x_i) x_i = 0$$

with $\rho'(.) = \psi(.)$. If $\rho(.)$ is convex, the two approaches are essentially equivalent. It appears that M-estimates offer enough flexibility and are by far the easiest to cope with, simultaneously, with regard to computation, asymptotic theory, and intuitive interpretation, also it is the only robust regression estimate whose asymptotic behavior are believed to be understood in fair detail. The objective function $\rho(.)$ is an outlier resistant function with some known properties;

- Always non negative, $\rho(u) \geq 0$
- Symmetric, $\rho(-u) = \rho(u)$
- $\rho(0) = 0$
- Monotone in $|u_i|$, $\quad \rho(u_i) \geq \rho(u_i')$ for $|u_i| \geq |u_i'|$

Some of the import cases of $\rho(.)$ are;

1. $\rho(u) = u^2$, which gives the OLS estimator.

2. Huber estimator Huber (1981)

$$\rho_M(u) = \begin{cases} \frac{1}{2}u^2 & |u| \leq M \\ M|u| - \frac{1}{2}M^2 & |u| > M \end{cases} \tag{1.9}$$

for a user-specified constant M. Here, $\rho(.) = \rho_M(.)$ which is clearly convex and differentiable. This function is a hybrid of a quadratic function for small values and a linear function for large values. The constant M is viewed as a shape parameter that controls the level of trade-off between efficiency and robustness, where a smaller value of M leads to better robustness and a larger value of M corresponds to better efficiency. The shape parameter is always set

to be M = 1.35 following the recommendation in Owen (2007), Ronchetti and Huber (2009), Lambert-Lacroix and Zwald (2011).

3.

$$\rho_M(u) = \begin{cases} 1 - [1 - (\frac{u}{M})^2]^3 & |u| \leq M \\ 1 & |u| > M \end{cases}$$

This gives the Tukey Biweight estimator, where M is usually chosen to be 4.685 Bai (2014)

4.

$$\rho_\theta(u) = \begin{cases} \theta u & u \geq 0 \\ -(1-\theta)u & u < 0 \end{cases}$$

This corresponds to quantile regression Koenker and Bassett Jr (1978) where $0 < \theta < 1$. Clearly, you can see that when $\theta = 0.5$, $\rho_{0.5}(u) = |u|$ gives least absolute deviations regression (LAD) or the median regression. In some other literature, a weight matrix was introduced to reduce the influence of outliers and iterative algorithm was used to solve problem, and this is the called Iteratively Reweighted Least-Squares (IRLS) algorithm.

The asymptotic properties (normality) of M estimators have been investigated both theoretically and empirically, see Huber (1964), Anscombe (1967), Relles (1968), Huber (1972), Andrews and Hampel (2015), Huber (1973), Yohai (1974), Bickel (1975), Bickel (1984), Portnoy (1984), Portnoy (1985), Portnoy (1987), Mammen (1989), Ronchetti and Huber (2009). The asymptotic properties are derived when assuming p is fixed and n diverges to infinity. However, in practice, p and n tend to become large simultaneously; in crystallography, where some of the largest least squares problems occur (with hundreds or thousands of parameters),

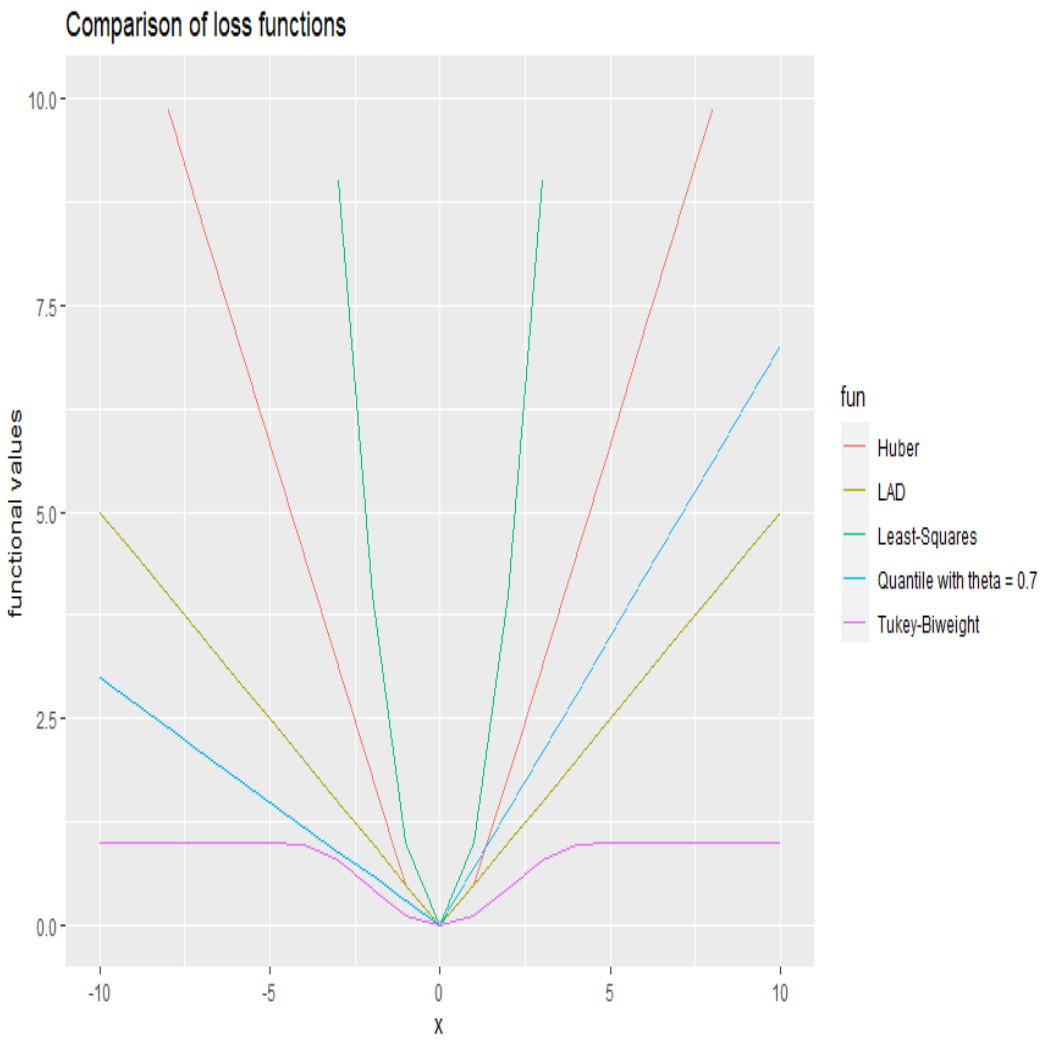


Figure 1.2: Plots of various loss functions, observe that LAD loss is the quantile loss when $\theta = 0.5$

according to Huber (1973) we find the explicit recommendation that there should be at least five observations per parameter, this suggests that a meaningful asymptotic theory should be in terms of $p/n \to 0$ or in terms of $h \to 0$, where h is the maximal diagonal element of the hat matrix. Hence we'll talk about the asymptotic theory of these estimators for the cases of p fixed and $p \to \infty$. The proof for the case of fixed p is a consequence of Bickel (1975). In the case of $p \to \infty$, it is assumed ψ has a bounded derivative, Yohai and Maronna (1979) required that $p^{\frac{3}{2}}h \to 0$, which improves an analogous result by Huber (1973) who required

$p^2h \to 0$. An overview of articles connected with asymptotics of M-estimators in linear models with increasing dimension is contained in Portnoy (1984). Portnoy (1985) assumed that the dimension of p grows with n in such a way that: $p^{\frac{3}{2}}(\log n)^{\frac{3}{2}}/n \to 0 \quad as\ n \to \infty$, and this condition was relaxed by Mammen (1989). Overall, for $p \to \infty$, p tends to infinity slower than n (hence, n will be sufficiently larger). In addition, since biases caused by asymmetric error distributions exist and can cause havoc within the asymptotic theory, although for most practical purposes, they will be so small that they can be neglected, Huber (1973) gave a procedure for computing the estimated covariance matrix with a correction factor. Boos (1980) proposed a method of constructing approximate confidence intervals for M-estimates for the special case of monotone non-decreasing right continuous ψ functions. El Karoui et al. (2013) recently discussed M-estimators for the case when p/n does not go to zero. It is worth noting that robust regression and outlier detection (using residuals from a robust regression estimate to identify outliers, so as to counter the issue of masking & swamping) are two closely related but not quite identical problems. There is a connection between the so called mean shift model and the Huber's M-estimates, see She and Owen (2011).

1.4.0.1 Penalized Huber Regression

In recent years, there has been a growing interest in applying robust methods to analyze high-dimensional data. One such method is regularized Huber-loss regression. Owen (2007) studied Huber-loss regression with a reversed Huber penalty, while Lambert-Lacroix and Zwald (2011) examined Huber-loss regression with an adaptive lasso penalty and obtained the estimate's asymptotic properties. These methods obtain estimators which are robust

against outliers and also enjoys a sparse representation. Here we consider the methods discussed above but replacing the least square loss with the huber loss $\rho_M(.)$ as in (1.1);

- Huber loss with l_1 penalty: $\min_\beta \sum_{i=1}^p \rho_M(y_j - \beta^T x_i) + \lambda||\beta||_1$
- Adaptive lasso penalty: $\min_\beta \sum_{i=1}^p \rho_M(y_j - \beta^T x_i) + \lambda(\sum_{j=1}^p \frac{|\beta_j|}{|\hat{\beta}_{init,j}|})$
- SCAD penalty: $\min_\beta \sum_{i=1}^p \rho_M(y_j - \beta^T x_i) + \sum_{j=1}^p pen_{\lambda,a}(\beta_j)$
- MCP penalty: $\min_\beta \sum_{i=1}^p \rho_M(y_j - \beta^T x_i) + \sum_{j=1}^p pen_{MCP,\lambda,a}(\beta_j)$

1.4.0.2 Some Simulation Study

Next, we do some simulation studies to demonstrate the performance of model selection of these methods. In all computations, the problems were solved using Iterative Local Adaptive Majorize-Minimization (I-LAMM) Algorithm for Nonconvex Regularized Robust Regression as introduced by Pan et al. (2021). The software is available online: `https://github.com/XiaoouPan`. We set up the experiment as follows, randomly generate a design matrix of dimension $n \times p$, the rows of X are independently sampled from a multivariate normal distribution with mean 0 and covariance matrix with entries $0.5^{|i-j|}$ for $i, j = 1, 2, \ldots, p$. Take $n = 400, p = 100$ and calculate the response as $y = X\beta + \varepsilon$ where $\beta = (3,\ 1,\ -1,\ 2,\ -0.5,\ 0,\ 0,\ 0,\ 0,\ 0,\ \ldots,\ 0)^T$ with only the first five components taken to be non zero, i.e., $s = 5$, hence β is sparse. Three different distributions will be considered for the random error ε, namely, $\mathcal{N}(0, 0.5)$, the students's t with degress of freedom 1.5, and mixed-normal distribution $\mathcal{N}(0, 0.5) + \mathcal{N}(0, 2)$. Both t and mixed-normal distributions are heavy-tailed, and produce outliers with high chance. The tuning parameter λ and the

robustification parameter M were all chosen implicitly in the algorithm using cross validation, as Pan et al. (2021) pointed out that with a properly chosen robustification parameter, calibrated by the noise level, sample size and parametric dimension, the effects of the heavy-tailed noise can be removed or dampened, see also Sun et al. (2020). We ran the simulation repeatedly (100 replications) and took the average of the summary statistics. The performance of the methods above were summarized with a table including True Positive (TP), which is the number of signal variables that are selected; False positive (FP), which is the number of noise variables that are selected; True Positive Rate(TPR), and False Positive Rate (FPR), are defined, respectively, as the ratio of true positive to s and the ratio of false positive to $p-s$. l_1 error and l_2 estimation error was also reported for more context. For the heavy tailed distribution, we can see that Huber-MCP & Huber-SCAD outperformed the rest, with fewer spurious discoveries (false positives), and smaller estimation errors. The performance of these methods are fairly similar for normal error as expected. Overall, methods with Huber loss outperformed methods with least square loss as it's evident in the estimation errors. This agrees with the fact that the use of Huber loss is particularly suited for heavy-tailed problems in both low and high dimensions, see (1.6)

1.5 Contribution / Thesis Outline

In recent years, there has been a growing interest in developing post-selection inference methods for high-dimensional data, particularly for linear models with sub-Gaussian errors. A lot of effort has been put into studying post-selection inference based on least squares estimation for these models. Specifically, Lee et al. (2016) derived closed-form confidence intervals and p-values by fitting the lasso with a fixed value of the regularization parameter. Similarly,

Taylor et al. (2014) provided similar results for forward stepwise regression and least angle regression (LAR), also see Tibshirani (2013) and Tibshirani et al. (2016). It is noteworthy that these results are non-asymptotic and hence applicable for any sample size n. However, one limitation of these methods is that they utilize the squared loss function, which are not robust to outliers and heavy-tailed errors. In addition, the sub-Gaussian assumption, which is made for technical convenience, may not be realistic in many practical situations, especially for data with heavy-tailed errors that are commonly observed in finance and economics.

In this dissertation, I will propose to use a Huber loss function with a generalized lasso penalty (gl-huber), and establish a finite sample conditional post-selection inferential tools for gl-huber while simultaneously conditioning on the outlier identification event and the variable selection event. Chapters 2 and 3 will present a comprehensive overview of the developed methodology, including its intricate details. Mainly, there three contributions of this dissertation:

- First, the proposed methodology employs a generalized lasso penalty that permits a broad range of penalization techniques, encompassing the usual lasso, adaptive lasso, and fused lasso as special cases. This approach offers superior flexibility and adaptability compared to other penalization methods.

- Secondly, the developed methodology characterizes the conditional distribution of the post-selection gl-huber estimator, while accounting for both variable selection and outlier identification events. This novel approach enables the development of valid conditional post-selection confidence intervals and p-values in high dimension that take

into account the variability in the selection process and satisfies all necessary frequency properties in the presence of heavy tailed error distribution / outliers. This contributes significantly to the statistical inference literature, particularly in the context of robust high-dimensional data analysis.

- Thirdly, in Chapter 4, this dissertation investigates the integration of a randomization technique from differential privacy with the developed methodology. The incorporation of this technique leads to a significant boost in the power of post-selection inferences made using the model.

The justification for using this conditional approach stems from the fact that researchers typically employ data to produce compelling hypotheses, and then conduct statistical inference on those hypotheses. In order to properly account for this exploratory data analysis, p-values and confidence intervals must be adjusted. One effective method for accomplishing this is to condition any inferences on the specific data components that were used to generate the hypotheses. This prevents the information contained within those components from being utilized again.

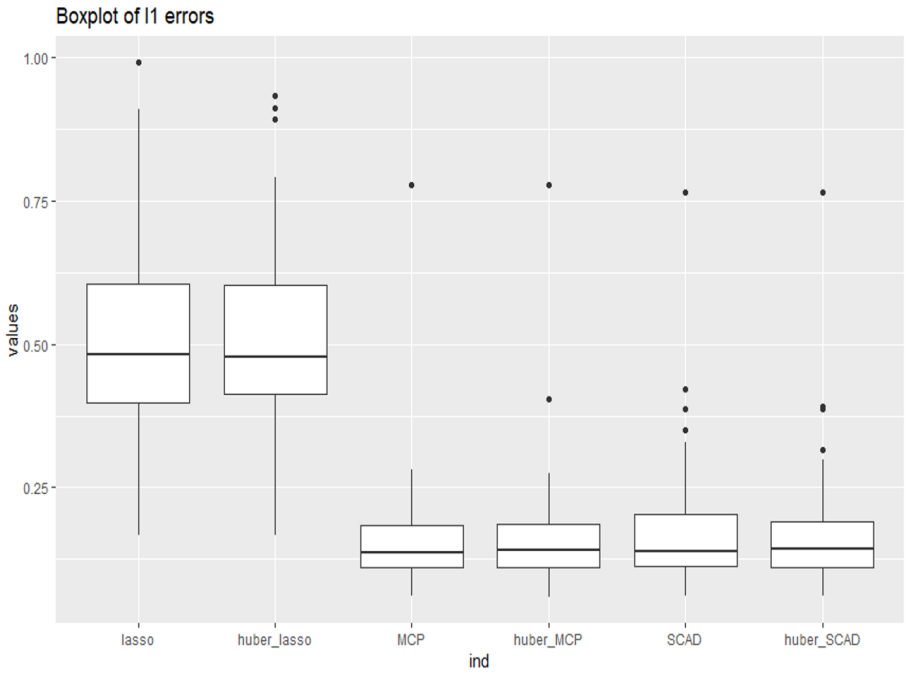
The dissertation is organized into five chapters. In Chapter 2, the focus is on post-selection inference, discussing methods for valid inference after model selection, with a particular emphasis on the general framework of post-selection inference by Lee et al. (2016) and Tibshirani et al. (2016). Chapter 3 addresses the problem of post-selection inference for robust regression techniques using Huber-loss regression with a generalized lasso penalty and heavy-tailed error distribution. Chapter 4 discusses post-selection inference after randomization.

Finally, Chapter 5 provides a summary of the conclusions drawn from the previous chapters and discusses the future work and importance of the topic.

1.6 Tables for Chapter 1 Simulations

$\epsilon \sim N(0, 0.5),\ sim = 100,\ n = 400,\ p = 100$						
Method	lasso	MCP	SCAD	Huber-lasso	Huber-MCP	Huber-SCAD
TP	5	5	5	5	5	5
TPR	1	1	1	1	1	1
FP	12.39	0.75	1.44	12.10	0.71	1.33
FPR	0.130	0.007	0.015	0.127	0.007	0.014
$l1$ - error	0.507	0.156	0.164	0.506	0.157	0.164
$l2$ - error	0.168	0.077	0.077	0.169	0.077	0.078

Table 1.2: Summary table for $\epsilon \sim N(0, 0.5),\ sim = 100,\ n = 400,\ p = 100$

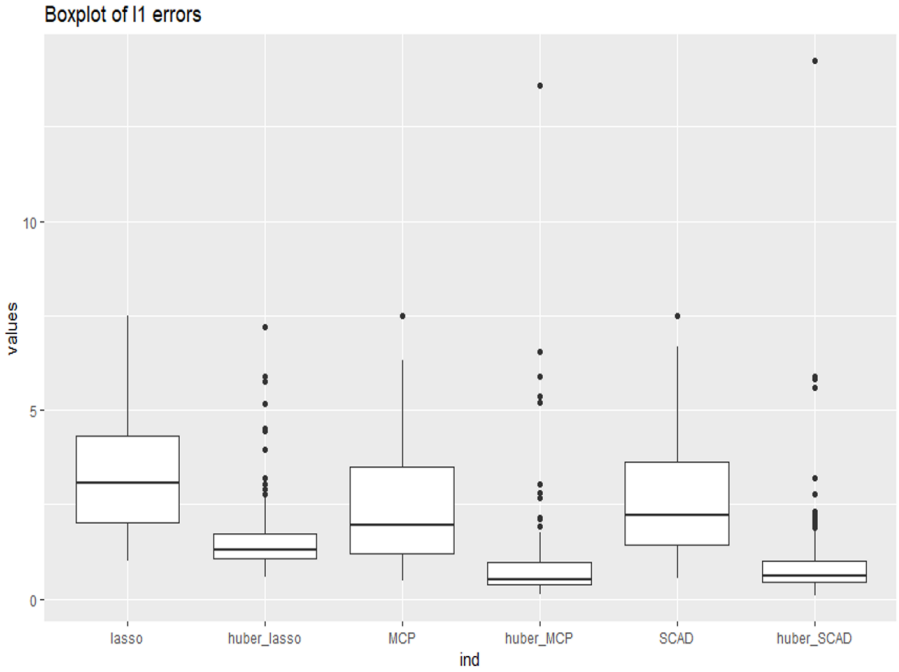


(a) boxplot

Figure 1.3: l_1 error plot for $\epsilon \sim N(0, 0.5), \quad sim = 100, \quad n = 400, \quad p = 100$

$\epsilon \sim t_{1.5},\ sim = 100,\ n = 400,\ p = 100$						
Method	lasso	MCP	SCAD	Huber-lasso	Huber-MCP	Huber-SCAD
TP	4.07	3.71	3.93	4.92	4.81	4.87
TPR	0.814	0.742	0.786	0.984	0.962	0.974
FP	12.84	3.57	6.11	7.30	1.51	4.23
FPR	0.135	0.037	0.064	0.076	0.015	0.044
$l1$ - error	3.429	2.624	2.815	1.676	1.042	1.117
$l2$ - error	1.411	1.272	1.277	0.556	0.408	0.421

Table 1.3: Summary table for $\epsilon \sim t_{1.5},\ sim = 100,\ n = 400,\ p = 100$

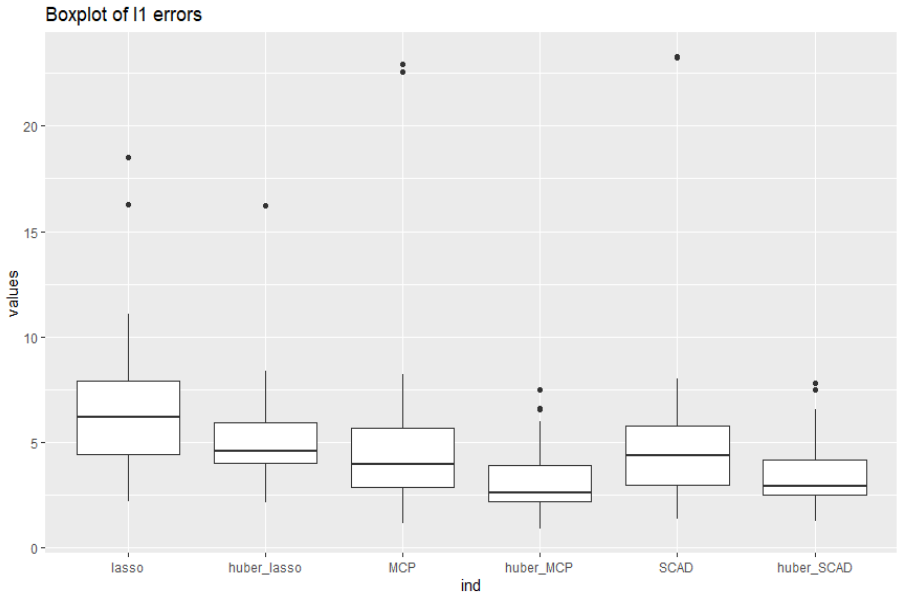


(a) boxplot

Figure 1.4: l_1 error plot for $\epsilon \sim t_{1.5}, \quad sim = 100, \quad n = 400, \quad p = 100$

$\epsilon \sim \mathcal{N}(0, 0.5) + \mathcal{N}(0, 2),\ sim = 100,\ n = 400,\ p = 100$						
Method	lasso	MCP	SCAD	Huber-lasso	Huber-MCP	Huber-SCAD
TP	3.34	2.66	3.14	3.94	3.18	3.44
TPR	0.668	0.532	0.628	0.788	0.636	0.688
FP	12.02	2.68	6.56	11.40	1.26	3.76
FPR	0.127	0.028	0.069	0.120	0.013	0.039
$l1$ - error	6.645	4.868	5.243	5.056	3.088	3.404
$l2$ - error	2.211	2.045	1.998	1.724	1.4669	1.522

Table 1.4: Summary table for $\epsilon \sim \mathcal{N}(0, 0.5) + \mathcal{N}(0, 2),\ sim = 100,\ n = 400,\ p = 100$



(a) boxplot

Figure 1.5: l_1 error plot for $\epsilon \sim \mathcal{N}(0, 0.5) + \mathcal{N}(0, 2), \quad sim = 100, \quad n = 100, \quad p = 100$

Chapter 2

Post Selection Inference

Conducting a valid statistical inference after model selection is currently a very active area in research. Here, we discuss a general approach to valid inference after model selection. For convenience of presentation, we restate the model as follows. The vector of responses $y \in \mathbb{R}^n$ is related to the design matrix $X \in \mathbb{R}^{n \times p}$ via a linear model.

$$y = X\beta + \varepsilon,$$

Many researchers incorporate lasso in statistical inference by a two-step procedure.

1. Use lasso as a tool for variable selection, that is, fit lasso to y and X to get the lasso solution $\hat{\beta}$. Define the active set as $\hat{\mathcal{A}} = \{k : \hat{\beta}_k \neq 0\}$, which contains the indexes of variables with nonzero coefficients.

2. Apply the ordinary least squares to y and $X_{\hat{\mathcal{A}}}$, where $X_{\hat{\mathcal{A}}}$ contains the columns corresponding to the indexes in $\hat{\mathcal{A}}$. The usual statistical inference, such as confidence interval and p-value, can be conducted based on the selected variables.

However, this two-step procedure ignores the fact that $\hat{\mathcal{A}}$ is computed from the data using lasso and is therefore subject to randomness. As a result, the statistical inference based on the selected variables may not be valid because it fails to account for all the variability in the

variable selection process. The selective inference formulates this problem as a conditional inference, conditional on the variables selected in the first step. The justification for using this conditional inference approach stems from the fact that researchers typically employ data to produce compelling hypotheses, and then conduct statistical inference on those hypotheses. In order to properly account for this exploratory data analysis, p-values and confidence intervals must be adjusted. One effective method for accomplishing this is to condition any inferences on the specific data components that were used to generate the hypotheses. This prevents the information contained within those components from being utilized again. We formally state the problem following the notations introduced in Tian and Taylor (2018). Let us assume that $y_i|x_i \sim G(\mu(x_i), \sigma^2(x_i))$, where x_i is the i^{th} row of the matrix X. We then apply a feature selection procedure to choose a subset $E \subset \{1, \ldots, p\}$, and the goal of the analysis is to infer the subset $\{\beta_j^E, j \in E\}$, where β_j^E is the coefficient for feature j in the linear regression model using only the features in E, i.e.,

$$\beta^E = \arg\min_b E\|y - X_E b\|^2$$

Supposing we are interested in inference for the first component of β^E, that is, $e_1^T\beta^E$. Its estimate is $e_1^T\hat{\beta}^E = \eta^T y$, where $\eta = e_1^T(X_E^T X_E)^{-1}X_E^T$ and e_1 is a vector with 1 at the first coordinate and 0 elsewhere. Standard theory tells us what to do when the sampling distribution of the data is Gaussian ($y \sim N(\mu, \sigma^2\mathbb{I}_{n\times n})$) and η is pre-specified: the Z-statistic

$$Z = \frac{\eta^\mathsf{T} y - \eta^\mathsf{T}\mu}{\sigma||\eta||_2} \sim \mathcal{N}(0,1) \tag{2.1}$$

is a pivot from which we can derive p-values and confidence intervals that will satisfy the desired frequency conditions. However, this theory does not hold because our η depends on the data through E, i.e., $\eta = \eta(E)$ and upon substituting this into the above expression, the pivotal relationship no longer holds since the OLS estimator $\eta^{\mathsf{T}} y = e_1^T (X_E^T X_E)^{-1} X_E^T y$ is no longer univariate Gaussian. Hence, the key is to derive the conditional distribution of $\eta^T y$ for some vector η, conditionally on the variables selected by lasso.

2.1 Polyhedral Method

Definition 2.1.1 *Define an affine selection procedure:*

$\mathcal{E}^* : \mathbb{R}^n \times \mathbb{R}^{n \times p} \to \mathcal{S}$, *where* $\mathcal{S}$ *is a finite set of models,* $\mathcal{S} = \{\mathcal{E}_1, \ldots, \mathcal{E}_{|\mathcal{S}|}\}$ *and for each potential model to be selected* $\mathcal{E} \in \mathcal{S}$,

$\{\mathcal{E}^*(z, X) = \mathcal{E}\} = \{A(\mathcal{E}, X)z \leq b(\mathcal{E}, X)\},\ (z, X) \in \mathbb{R}^n \times \mathbb{R}^{n \times p},\ A \in \mathbb{R}^{k \times n}\ b \in \mathbb{R}^k,\ k \in \mathbb{N}.$

There are several affine selection procedures that use different algorithms to select a set of variables, denoted by E, based on the data and other information. Some examples of these procedures include selecting E as the active set of the lasso solution at a fixed λ, selecting E as the first variable to enter the lasso or LARS path, or selecting E as the k variables included at the k^{th} step of forward stepwise selection. After solving the lasso objective function, variable selection output can be described as E, which is the set of indexes corresponding to non-zero components of $\hat{\beta}$, $s_E \in \{\pm 1\}^{|E|}$ are their signs. The works of Lee et al. (2016) aim to derive the conditional distribution of $\eta^T y$ by conditioning on the active set E and s_E, also known as the selection event. This selection event $\{E, s_E\}$ is an affine selection procedure as defined above, so Lee et al. (2016) and Tibshirani et al. (2016) showed

a simple characterization of the selection event as follows

$$\{\mathcal{E}^*(y, X) = \mathcal{E}\} = \{A(\mathcal{E}, X)y \leq b(\mathcal{E}, X)\},$$

where $\mathcal{E}^*(y, X) = \hat{\beta}_j^E, j \in E$. The RHS above is a polyhedral region in outcome space (the exact form for the matrix $A(\mathcal{E}, X)$ and vector $b(\mathcal{E}, X)$ can be derived from the KKT conditions at the solution).

Lemma 2.1.2 (Lee et al. (2016), Tibshirani et al. (2016))

$$\begin{aligned} \{A(\mathcal{E}, X)y \leq b(\mathcal{E}, X)\} &= \{L_{\mathcal{E}}(z) \leq \eta^T y \leq U_{\mathcal{E}}(z)\} \\ L_{\mathcal{E}}(z) &\equiv \max_{j:(A(\mathcal{E},X)c)_j<0} \frac{b_j(\mathcal{E}, X) - (A(\mathcal{E}, X)z)_j}{(A(\mathcal{E}, X)c)_j} \\ U_{\mathcal{E}}(z) &\equiv \min_{j:(A(\mathcal{E},X)c)_j>0} \frac{b_j(\mathcal{E}, X) - (A(\mathcal{E}, X)z)_j}{(A(\mathcal{E}, X)c)_j} \\ z \equiv (\mathbb{I}_n - c\eta^T)y, &\quad c \equiv \Sigma\eta(\eta^T\Sigma\eta)^{-1}, \quad \Sigma = \sigma^2\mathbb{I}_n \end{aligned} \tag{2.2}$$

$\eta^T y$ and $(L_{\mathcal{E}}(z), U_{\mathcal{E}}(z))$ are all statistically independent since $L_{\mathcal{E}}(z), U_{\mathcal{E}}(z)$ are all functions of z only, z can be defined as a nuisance statistics which corresponds to nuisance parameters, i.e., all directions orthogonal to the our direction of interest η. From lemma 2.1.2, we can see that the selection event $\{\mathcal{E}^*(y, X) = \mathcal{E}\} = \{A(\mathcal{E}, X)y \leq b(\mathcal{E}, X)\}$ is equivalent to the event that $\eta^T y$ falls into a certain range $L_{\mathcal{E}}(z), U_{\mathcal{E}}(z)$, a range that depends on $A(\mathcal{E}, X)$ and $b(\mathcal{E}, X)$.

2.2 Normal case

Lee et al. (2016) obtained the conditional distribution of $\eta^T y$ when the errors are normally distributed, that is, $y|X \sim G(\mu(X), \sigma^2(X)) \equiv \mathcal{N}(0, \sigma^2 I_{n\times n})$. The following summary presents their result, which is non-asymptotic and applicable to any n. From (2.2), Lee et al. (2016) established the conditional distribution of $\eta^\mathsf{T} y$ given $\{A(\mathcal{E}, X)y \leq b(\mathcal{E}, X)\}$ to be equivalent to

$$[\eta^T y \,|\, \{A(\mathcal{E}, X)y \leq b(\mathcal{E}, X)\}] \stackrel{d}{=} [\eta^T y \,|\, \{L_{\mathcal{E}}(z) \leq \eta^T y \leq U_{\mathcal{E}}(z)\}]$$

Lemma 2.2.1 (Lee et al. (2016), Tibshirani et al. (2016))

$$[\eta^T y | A(\mathcal{E}, X)y \leq b(\mathcal{E}, X), \wp_\eta^\perp y = z] \sim \mathcal{TN}(\eta(\mathcal{E}, X)^T \mu, \sigma^2 ||\eta(\mathcal{E}, X)||^2, L_{\mathcal{E}}(z), U_{\mathcal{E}}(z))$$

where $\mathcal{TN}(\cdot, \cdot, \cdot, \cdot)$ is the truncated normal distribution on the interval $[L_{\mathcal{E}}(z), U_{\mathcal{E}}(z)]$. Therefore we have that, the distribution of $\eta^T y$ conditioned on the selection event and on the nausiance statistics ($\wp_\eta^\perp y$ is the piece of y orthogonal to η), follows a univariate truncated normal, with $L_{\mathcal{E}}(z)$, $U_{\mathcal{E}}(z)$ explicitly computed from data. Using the probability integral transform, we can get a pivot as follows:

Theorem 2.2.2 (Lee et al. (2016), Tibshirani et al. (2016)) *Let $F_{\mu,\sigma^2}^{[a,b]}$ denote the CDF of a $\mathcal{N}(\mu, \sigma^2)$ random variable truncated to the interval $[a, b]$, that is,*

$$F_{\mu,\sigma^2}^{[a,b]}(x) = \frac{\Phi((x-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}$$

and Φ *represents the cumulative distribution function of a standard normal distribution. Then, marginalizing over the selection procedure* $\mathcal{E}^*$*, we have the following*

$$F^{[L_{\mathcal{E}^*}(z),U_{\mathcal{E}^*}(z)]}_{\eta(\mathcal{E}^*)^T\mu,\ \sigma^2||\eta(\mathcal{E}^*)||^2}(\eta(\mathcal{E}^*)^T y) \mid \{A(\mathcal{E}^*, X)y \leq b(\mathcal{E}^*, X)\} \sim \mathcal{U}(0,1) \tag{2.3}$$

where $L_{\mathcal{E}^*}$ *and* $U_{\mathcal{E}^*}$ *are defined in 2.1.2.*

We obtain confidence intervals by inverting the pivotal quantity (2.3). For a $1-\alpha$ interval, we find the largest and the smallest $\eta^T\mu$ such that the value of pivotal quantity remains in the interval $\left[\frac{\alpha}{2},\ 1-\frac{\alpha}{2}\right]$. It is worth noting that we can condition on either $\{E, s_E\}$ or E, and according to Lee et al. (2016), there are advantages and disadvantages to conditioning on either E, s_E or just E. It's important to consider the tradeoffs associated with each option.

2.3 Non-normal case

If we remove the assumption that the error is Gaussian, then (2.3) will be false, and subsequently, the conclusion of Theorem 2.2.2 does not hold anymore. The best we can hope for is a weak convergence result that the same pivotal quantities (2.3) would converge to $\mathcal{U}(0,1)$ (as $n \to \infty$). Tian and Taylor (2017) relaxed the Gaussian assumption and showed that the conclusion of theorem 2.2.2 is true asymptotically when the error is not necessarily Gaussian. Their approach was to compare the distribution of the pivots (2.3) under a non-Gaussian error distribution denoted as $\mathcal{L}(y|X)$ with under Gaussian distribution denoted as $\mathcal{L}(\mathcal{Y}|X)$, and this requires some conditions on both the distribution $\mathcal{L}(y|X)$ and the selection procedure $\mathcal{E}^*$ as established by Tian and Taylor (2017).

Assumption 1 (Tian and Taylor (2017)) *Suppose we have $X_n \in \mathbb{R}^{n\times p_n}$, and $y_n \in \mathbb{R}^n$ (conditional on X_n) has some distribution, and $\mathcal{Y}_n$ is generated independently (conditional on X_n) from $N(\mu(X_n), \Sigma(X_n))$ a Gaussian distribution with the same means and variances. We have affine selection procedures such that $\mathcal{E}^* = \mathcal{E}^*_n$. Here we assume there exists $\gamma_n \to 0$*

$$\mathbb{P}(U_{\mathcal{E}^*}(y_n) - L_{\mathcal{E}^*}(y_n) < \gamma_n) \to 0,$$

$$\mathbb{P}(U_{\mathcal{E}^*}(\mathcal{Y}_n) - L_{\mathcal{E}^*}(\mathcal{Y}_n) < \gamma_n) \to 0,$$

$$\mathbb{P}(\min(|U_{\mathcal{E}^*}(y_n)|, |L_{\mathcal{E}^*}(y_n)|) > 1/\gamma_n) \to 0,$$

$$\mathbb{P}(\min(|U_{\mathcal{E}^*}(\mathcal{Y}_n)|, |L_{\mathcal{E}^*}(\mathcal{Y}_n)|) > 1/\gamma_n) \to 0,$$

Tian and Taylor (2017) imposed conditions on $(\gamma_n, M(\mathcal{E}^*_n, \eta_n), r(\mathcal{E}^*_n), |\mathcal{S}_n|)$ to ensure the convergence of the pivot (2.3). Where $\mathcal{S} = \{\mathcal{E}_1, \ldots, \mathcal{E}_{|\mathcal{S}|}\}$ is a finite set of models as in 2.1.1,

$$M(\mathcal{E}^*, \eta) = \max_{\mathcal{E}\in\mathcal{S}} M(\mathcal{E}, \eta), \text{ where}$$
$$M(\mathcal{E}, \eta) = \max_{\substack{1\le i\le \mathrm{nrow}(A(\mathcal{E}))\\ 1\le j\le n}} \left|\frac{A(\mathcal{E})_{ij}}{(A(\mathcal{E})\Sigma\eta(\mathcal{E}))_i}\right| + ||\eta(\mathcal{E})||_\infty,$$
$$r(\mathcal{E}) = \mathrm{nrow}(A(\mathcal{E})), \quad r(\mathcal{E}^*) = \max_{\mathcal{E}\in\mathcal{S}} \mathrm{nrow}(A(\mathcal{E})).$$

The conditions are choosing an appropriate γ_n for Assumption 1, bounding $|\mathcal{S}_n|$, $M(\mathcal{E}^*_n, \eta_n)$ and $r(\mathcal{E}^*_n)$.

Theorem 2.3.1 (Tian and Taylor (2017)) *(Convergence of pivot)*

Suppose we have a sequence of y_n generated with means $\mu_n = \mu(X_n)$, and variances $\Sigma_n =$

$\Sigma(X_n)$ *and have finite third moments. We also assume Assumption 1 is satisfied with a sequence of* γ_n*. Furthermore, let* $\mathcal{E}_n^*$ *be a sequence of affine selection procedures,* $\eta_n = \eta(\mathcal{E}_n^*)$*, and the corresponding* $M(\mathcal{E}_n^*, \eta_n), r(\mathcal{E}_n^*)$ *and* $\mathcal{S}_n$ *are properly defined. Then if*

$$\frac{1}{\gamma_n^6} \cdot M(\mathcal{E}_n^*, \eta_n)^3 \cdot n[\ \log(r(\mathcal{E}_n^*)) + \log(|\mathcal{S}_n|)]\ ^4 \to 0,\ \ as\ n \to \infty$$

we have

$$P(\eta_n^T y_n;\ \eta_n^T \Sigma_n \eta_n,\ \eta_n^T \mu_n,\ L_{\mathcal{E}_n^*},\ U_{\mathcal{E}_n^*}) \xrightarrow{d} \mathcal{U}(0,1),\ \ n \to \infty \tag{2.4}$$

where $P(x; \sigma^2, m, a, b) = 2\min(F^{[a,b]}_{m,\sigma^2}(x), 1 - F^{[a,b]}_{m,\sigma^2}(x))$ *is the two sided pivot.*

2.4 Limitations

The method presented in Lee et al. (2016) and Tibshirani et al. (2016) enables precise conclusions to be drawn following the use of the lasso (or any other method whose selection event is polyhedral) to formulate hypotheses. Nevertheless, it is not flawless:

- The intervals condition on (E, s_E) rather than just E. This would be appropriate if we also used information in the signs when forming our hypotheses. For example, we might test against a one sided alternative, $H_1 : \beta_j^{(E)} > 0$, if the sign of the j^{th} variable is positive. However, in most applications, the signs are not used in hypothesis generation, hence we have thrown away unused information by conditioning on s_E. As a result, the intervals tend to be quite long.

- Lee et al. (2016) made mention of the point above and showed how to condition only on E by an enumeration of all possible sign vectors. Their enumeration method is intractable when $|E|$ is large as there are $2^{|E|}$ possible sign vectors.

- Kivaranovic and Leeb (2020) proved that the expected length of the above intervals are infinite. They showed that if the truncation limits are bounded either from above or from below, i.e., $\forall\ z,\ -\infty < L_{\mathcal{E}^*}(z)$ or $U_{\mathcal{E}^*}(z) < \infty$ or both, then the expected length will be infinite. This intuitively stems from the fact that one expects confidence intervals to be wide if you condition on a bounded set because extreme values cannot be observed on a bounded set and the confidence intervals have to take this into account, see Kivaranovic and Leeb (2020).

But these are not quite surprising since Lee et al. (2016) mentioned that if the observed statistic is too close to either end of the truncation interval V^- and V^+, then one or possibly both endpoints of the interval of desired coverage cannot be computed, and default to +/- ∞. The fairly lengthy confidence intervals which in turn implies slight loss of power is a necessary price to be paid for better justification of statistical inference in the context of the pre-inferential liberties taken in today's data-analytic practice. There are many related researches in the literature, which may inspire ideas for further improvement in power, Tian and Taylor (2018) introduced a randomization scheme in linear regression that involves additive noise.

2.5 Randomization

As discussed earlier in this chapter, the approach developed by Lee et al. (2016) for post selection inference is based on truncating the generative law of the data to realizations that lead to a selection event. However, Kivaranovic and Leeb (2020) showed that the expected length of confidence intervals based on this method is typically infinite. To address this issue, there have been ongoing developments in post selection inference, which can be roughly divided into two main categories: (1) post selection inference based on a minimal conditioning set and (2) introducing randomized procedures to improve inferential power. In the second category, the authors found that randomized response result in significantly shorter intervals than those based on the polyhedral method alone. The randomized response involves adding a noise term to the response variable in the model, and it yields more powerful statistical tests while only incurring a small cost in terms of the quality of the selected models. The inclusion of a small amount of randomization has a minimal impact on the model selection process but results in a significant increase in the power of inferences made using the model, and this is due to the concept of leftover Fisher information. This section provides an in-depth discussion of post-selection inference after randomization.

2.5.1 A natural Gaussian Randomization Scheme

If you consider the response variable $y \in \mathbb{R}^n$ as in (1.1), partition y into selection and inference data sets y_1 and y_2, containing n_1 and $n_2 = n - n_1$ data points respectively.

Data splitting procedure uses the lasso on y_1 to select the model and y_2 for inference. Post-selection inference for data splitting relies only on the data y_2, and fails to utilize any left-over information from the selection data y_1. A solution for valid post-selection inference, namely *carving* Fithian et al. (2017), is an efficient alternative to data splitting: because carving eliminates the information used in selection instead of discarding the selection data all at once, i.e., data carving uses the lasso on y_1 to select the model, and y_2 and whatever is left over of y_1 for inference. The polyhedral lemma by Lee et al. (2016) applied to carving immediately yields a carved pivot which is then inverted to produce confidence intervals. The length of resulting $100(1-\alpha)\%$ confidence intervals are shorter than those by Lee et al. (2016) and they aim to control a false coverage proportion post-selection. Carving can be perceived as the best of the two worlds, i.e., the best of data splitting and method by Lee et al. (2016), and in this section, we intend to show that data carving is actually asymptotically equivalent to randomizing with Gaussian Noise.

As in (1.1), $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, then the mathematical formulation of the randomization scheme with Gaussian noise for the case of lasso is given by adding a noise term drawn from Gaussian distribution which is linear in the parameter β, i.e.,

$$\min_{\beta} \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1 - \omega^T\beta \quad \omega \sim \mathcal{N}(0, \tau^2 I_p) \tag{2.5}$$

where ω is regarded as the noise term, randomization term or perturbation term. In the differential privacy literature, the objective function above (2.5) is termed 'objective perturbation', and it also holds true when you replace the squared loss with any other loss function and the l_1 penalty with any other penalty. Observe that τ^2 is independent of the

data distribution, and we can fix the scale of τ^2 however we want to depending on the extent we want to randomize our objective. To show the equivalence of carving and randomization, let $y_{(1)}^{n_1}$, $X_{(1)}^{n_1 \times p}$ be n_1 randomly chosen samples, the lasso objective for data carving becomes,

$$\text{lasso}: \quad \min_{\beta} \frac{1}{2\rho}||y_{(1)} - X_{(1)}\beta||_2^2 + \lambda||\beta||_1$$

where ρ is the fraction of the data we are using for selection and can be explicitly determined, here $\rho = \frac{n_1}{n}$, and if we are using 50% of the data for selection, then $\rho = \frac{1}{2}$. We can write the lasso objective as an objective on the entire data, i.e.,

$$\min_{\beta} \frac{1}{2}||y - X\beta||_2^2 - \left\{\frac{1}{2}||y - X\beta||_2^2 - \frac{1}{2\rho}||y_{(1)} - X_{(1)}\beta||_2^2\right\} + \lambda||\beta||_1 \tag{2.6}$$

now, the gradient of the second term with respect to β can be taken as the randomization term, which can be shown to be asymptotically Gaussian,

$$\omega = \nabla\left\{\frac{1}{2}||y - X\beta|_2^2 - \frac{1}{2\rho}||y_{(1)} - X_{(1)}\beta||_2^2\right\} \approx \mathcal{N}(0, \Sigma) \tag{2.7}$$

where Σ is some covariance matrix that depends on the splitting fraction ρ. We have rewritten the lasso objective function for data carving as a randomization perturbed objective, and the randomness here is coming from the split, the fact that we have randomly chosen a split of the data in which we have conducted the selection. The noise ω in (2.7) having asymptotic Gaussian distribution is general in nature and does not depend on the lasso procedure, i.e., it holds true when you replace the squared loss with any other loss function and the l_1 penalty with any other penalty. There is a direct relationship between ρ in (2.6) and τ^2 in (2.5), both

are independent of the data distribution and be explicitly determined. Also, there is a huge trade-off between model selection and inference after selection, for (2.5), we control the scale of τ^2 because we don't want to randomize our objective to a great extent that we lose all the signals and report completely different findings that we would have gotten without the noise, hence we don't have to add too much noise that we get a completely noisy selection. Similarly, for (2.6), we don't want to carve the data in a way that we will fail to detect the true signals during selection. In a way of summary, the Gaussian randomization scheme is very natural and it comes from the data carving idea.

2.5.2 Post selection Inference with Randomization Responses

We formally state the problem as follows, consider solving the lasso at a fixed λ

$$\min_{\beta} \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1 \tag{2.8}$$

we incorporate randomness into (2.8), in particular, we consider solving

$$\min_{\beta} \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1 - \omega^T\beta + \frac{\epsilon}{2}||\beta||_2 \tag{2.9}$$

where $\omega \sim G$ is a random vector independent of (X, y), whose distribution is chosen by the data analyst, hence known. Here, we will assume that G is supported on all of $\mathbb{R}^p$ with density g. The ridge term with small parameter ϵ ensures the problem above has a solution, and this is the ridge term in the elastic net as established by Zou and Hastie (2005). Randomized convex programs are a type of optimization method that have been examined

in Tian and Taylor (2018). To conduct post selection inference after randomization, suppose we solve the randomized lasso (2.9) to obtain a selection (E, s_E), where $E \subset \{1, \ldots, p\}$ is the candidate set of variables and $s_E \in \{\pm 1\}^{|E|}$ are their signs. The selection event consists of all data and randomization pairs such that solving the randomized lasso above for that pair gives the same (E, s_E):

$$S_{(E,s_E)} = \{(X', y', \omega') : \hat{\beta}_{-E}(X', y', \omega') = 0, \text{sign}(\hat{\beta}_E(X', y', \omega')) = s_E\}$$

The selection $S_{(E,s_E)}$ is an affine selection procedure, hence it's equivalent to the selection region $S_{(E,s_E)} \equiv \{y : Ay + \omega \leq b\}$, which is again a set of polyhedral constraints except that selection is now defined in both the response y and randomization ω, where A and b are functions of the selection of E, i.e., functions of the coefficient of the active set. In order to conduct inference, the underling question is how do we calibrate the right law for the OLS estimator $\eta^{\intercal} y = \left(X_E^T X_E\right)^{-1} X_E^T y$ given that $\eta = \left(X_E^T X_E\right)^{-1} X_E^T$ depends on the selection event E. Without randomization (that is setting the scale of randomization to zero), Lee et al. (2016) showed that

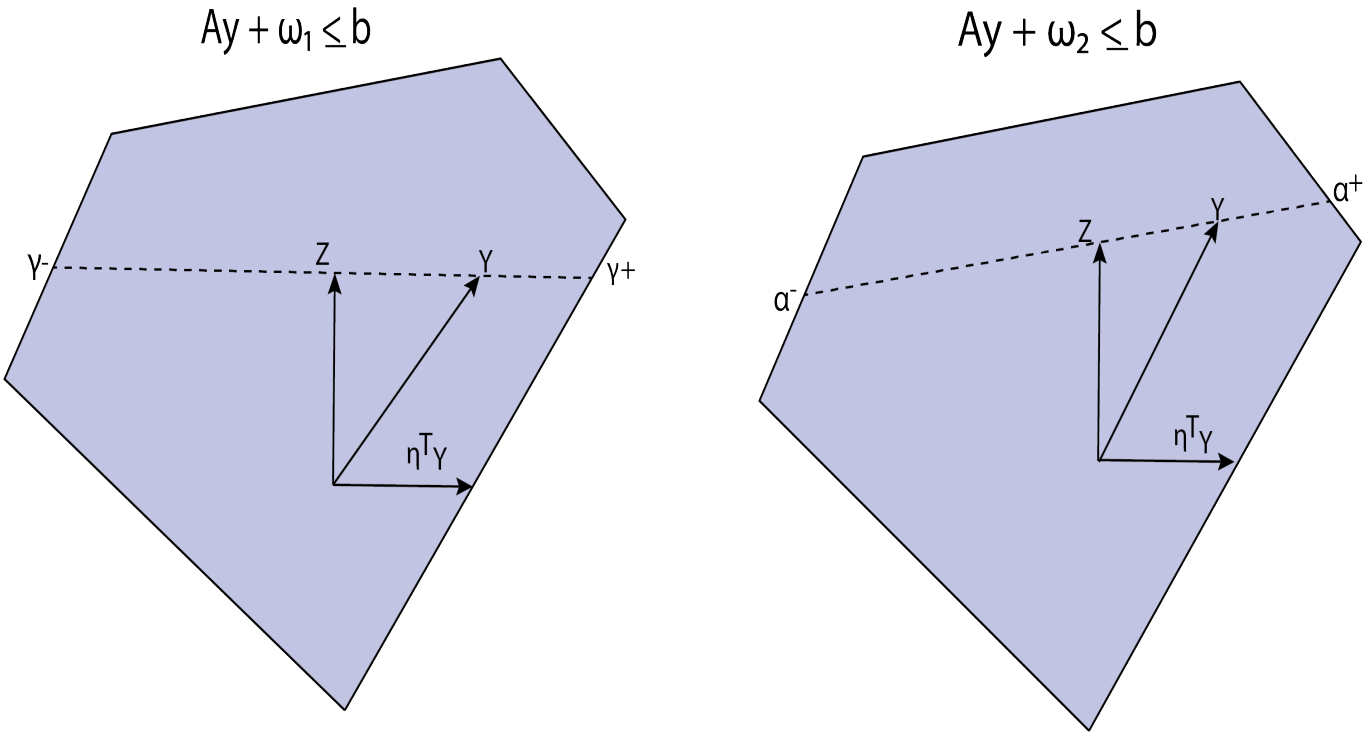
$$\mathcal{L}(\eta^T y | Ay \leq b, \wp_\eta^\perp y = z) \sim \mathcal{TN}(\eta^T \mu, \sigma^2 ||\eta||^2, L(z), U(z))$$

i.e., the distribution of $\eta^T y$ conditioned on the selection event and on the nausiance statistics, follows a univariate truncated Gaussian, with $L(z),\ U(z)$ explicitly computed from data, see (2.2.1). Lee et al. (2016) calculated the pivot by applying the CDF transform of a univariate truncated Gaussian law to the target statistic, and the confidence intervals are then obtained

by inverting a pivotal statistic based on the truncated law. However, with randomization, the same law of our naive statistics $\eta^T y$ in a particular direction (say the first coordinate of the OLS estimator we obtained by refitting our data our selected subspace X_E) conditioned on the polyhedra event $Ay + \omega \leq b$ and the nausiance statistic $\wp_\eta^\perp y$ is no longer Guassian, i.e.,

$$\mathcal{L}(\eta^T y | Ay + \omega \leq b, \wp_\eta^\perp y) \neq \textbf{TRUNCATED GUASSIAN}$$

making the pivots to be intractable. To see why it is not so, if we have this polyhedral



selection at a fixed realization of ω say ω_1 (such as the one shown in the left of the figure above), then by also conditioning on the nausiance statistics we can see that $\eta^T y$ would be restricted between γ^- and γ^+. However, the selection is not a fixed realization of ω instead it is marginalizing over these ω's, i.e., the selection probability that we are computing here is randomizing both over the data as well as the ω's, so if we compute at a different realisation

of ω say ω_2, we'll get $\eta^T y$ to be restricted between a different ray α^- and α^+ as in the figure on the right. Therefore, $\eta^T y | Ay + \omega \leq b$ is no longer restricted to a single ray γ^- and γ^+ as we had in the polyhedral lemma of Lee et al. (2016), since γ^- and γ^+ are now random quantities which vary as we vary different realizations of ω which is a random quantity, i.e.,

$$\{\eta^T y | Ay + \omega \leq b\} \neq \{\gamma^- \leq \eta^T y \leq \gamma^+\}$$

which in turn implies that $\mathcal{L}(\eta^T y | Ay + \omega \leq b, \wp_\eta^\perp y) \neq$ TRUNCATED GUASSIAN, hence we cannot proceed with polyhedral lemma by Lee et al. (2016). Tian and Taylor (2018) also pointed out that the exact forms of $\eta^T y | Ay + \omega \leq b$ cannot be computed. Here, γ^-, α^- denote different lower limits $L(z)$'s, while γ^+, α^+ denote different upper limits $U(z)$'s.

Now, we don't have an exact pivotal quantity anymore since the polyherdral lemma of Lee et al. (2016) no longer apply, what might seem natural is to run a sampler from the joint law $\mathcal{L}(\eta^T y, \omega)$ of data and randomization truncated to the selection region which is basically a polyhedral region.

2.5.3 MCMC Approach

First, let us try to understand $\mathcal{L}(\eta^T y, \omega)$, as in (1.1), let $y = \mu + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and consider an adaptive target $b_E = \left(X_E^T X_E\right)^{-1} X_E^T \mu$ which is the population least square coefficient (the projection of the true mean μ on to the selected subspace of X), with first coordinate $b = e_1^T b_E$ which is the target we want to infer about, and a target statistic which

is the OLS estimator $T = e_1^T \left(X_E^T X_E\right)^{-1} X_E^T y$.

If we take $\omega \sim \mathcal{N}(0, \tau^2 I_p)$, the pre-selection joint density of (T, ω) at (t, w) is proportional to

$$exp(-(t-b)^2/2\sigma^2) \times exp(-||w||^2/2\tau^2)$$

and we know that selection constraints both data and randomization to lie in the polyhedral region $S_{(E,s_E)}$ defined above, which is described in terms of data and randomization. Now, the selection adjusted joint density of (T, ω) at (t, w) is proportional to the same density truncated to the selection region $S_{(E,s_E)}$, and the selection region $S_{(E,s_E)}$ is determined by the selection we carried out on the randomized version of our data. Hence, the joint density becomes,

$$exp(-(t-b)^2/2\sigma^2) \times exp(-||w||^2/2\tau^2) \times \mathbb{I}_{(t,w)\in S_{(E,s_E)}} \tag{2.10}$$

If we run a sampler, say a Monte Carlo sampler, what a sampler will do is to simply sample t and w from density (2.10), and discard the samples of randomization w since they are not used for inference, and hence can use the samples of the data t to proceed in conducting inference for b. The approach outlined in references Fithian et al. (2014) and Tian and Taylor (2018) involves selecting samples from a restricted area within the sample space as in (2.10). Both studies employ the hit-and-run algorithm to generate distributions within this subset of the space. The constrained subsets outlined in references Fithian et al. (2014), Tian and Taylor (2018), Lee et al. (2016), and Taylor and Tibshirani (2018) can be intricate and specific to the loss function in use. While algorithms that don't use MCMC, like those in references Lee et al. (2016) and Taylor and Tibshirani (2018), are less affected

by this problem as they only require calculating the boundary once, methods in references Fithian et al. (2014) and Tian and Taylor (2018) require recomputing the boundary at each step of the simulation, resulting in a significantly higher computational cost. The complexity and specificity of the constrained region $S_{(E,s_E)}$ to the loss function makes it a challenging aspect for sampling. However, by using a reparametrization technique, we can express the constrained region as a simple set that is unrelated to the loss function. To do that, define the optimization variables $O = \begin{pmatrix} \hat{\beta}_E \\ \\ \mu_{-E} \end{pmatrix}$ where μ_{-E} being the subgradient vector for the penalty corresponding to inactive variables. Denote the observed data vector $D = \begin{pmatrix} D_E \\ \\ D_{-E} \end{pmatrix} = \begin{pmatrix} \bar{\beta}_E \\ \\ X_{-E}^T(y - X_E\bar{\beta}_E) \end{pmatrix}$, where $\bar{\beta}_E = \left(X_E^T X_E\right)^{-1} X_E^T y$. By applying the KKT conditions on our objective function (2.9), Harris et al. (2016) defined a linear map between randomization and the augmented vector (D, O), called randomization reconstruction as

$$\omega = \omega(D, O) = A_0 D + BO + \gamma \tag{2.11}$$

where $A_0 = \begin{pmatrix} X_E^T X_E & 0 \\ \\ X_{-E}^T X_E & I_{p-|E|} \end{pmatrix}$, $B = \begin{pmatrix} X_E^T X_E + \epsilon I_{|E|} & 0 \\ \\ X_{-E}^T X_E & \lambda I_{p-|E|} \end{pmatrix}$ are fixed matrices, and $\gamma = \begin{pmatrix} s_E \\ \\ 0 \end{pmatrix}$ is a fixed vector. The authors showed that the selection event (E, s_E) from the solver in (2.9) is now described by the map $\omega(D, O)$ where optimization variables O are constrained to the region

$$\mathcal{K} = \{o \in \mathbb{R}^p : \text{sign}(o_E) = s_E, ||o_{-E}||_\infty \leq \lambda\} \tag{2.12}$$

hence, selective inference will now be based on the joint law of data and randomization (D, ω), conditional on the event that constrains the optimization variables O to lie in $\mathcal{K}$, and a change of measure formula from the space (D, ω) to that of (D, O) will enable sampling to be done from a density supported on the much simpler constraint region $\mathcal{K}$. Using the change of measure trick of Harris et al. (2016), the truncated joint of (D, O) at (d, o) becomes

$$f_D(d) \times g(w(d, o)) \times \mathbb{I}_{o \in \mathcal{K}} \tag{2.13}$$

where $f_D(d)$ is the pre-selection density of D, $g(.)$ is the density of the randomization ω. Sampling from (2.10) and (2.13) are comparable, but (2.13) is much simpler and nicer set than (2.10) as it only requires o_E to be in a certain quadrant and $||o_{-E}||_\infty \leq \lambda$. Another related literature is Markovic and Taylor (2016), where the authors constructed confidence intervals via Monte Carlo sampling in the randomized setting.

2.6 Recent Advancements

The exact post-selection inference discussed above, and also exact post-selection inference for sequential regression have been further studied. For the existing frameworks, we have two main assumptions, and they are (i) the variance of the responses, σ^2, is known, and (ii) the response y follows a Gaussian distribution. Tian and Taylor (2018) addressed first assumptions by applying the exact post-selection inference to square-root lasso for inference on selected submodel after model selection, where an estimate for σ^2 is derived based on the square-root lasso. While alternatively, Tibshirani et al. (2018) constructed a computationally efficient bootstrapped version of the truncated normal statistics in 2.2.1 which does

not depend on σ^2. For the second assumption, if we remove the assumption that y follows a Gaussian distribution, the conclusion of 2.2.2 will be false. Tian and Taylor (2017) and Tibshirani et al. (2018) examined large sample conditional framework of exact post-selection inference, both of their works show that under certain conditions on the distribution of y, selection procedures and unknown regression coefficients, the pivotal quantity in (2.3) converge ($n \to \infty$, p constant) to the uniform distribution, hence subsequent construction of post-selection confidence interval can be conducted in the same fashion. Further exploration into the asymptotic aspects of exact post-selection inference can be found in Taylor and Tibshirani (2018) and the references therein, Taylor and Tibshirani (2018) showed that the exact post-selection inference framework can be generalized for statistical inference of a large class of l_1-penalized regression models, including generalized linear models, Cox's proportional hazards model, and the graphical lasso. Zhao et al. (2022) applied the exact post-selection inference approach in a two-stage proposal for solving effect modification problem, where they showed that this method is asymptotically valid, both theoretically and via simulations. Hyun et al. (2021) studied post-detection inference of change point problems, the authors first characterised the change point detection as polyhedral selection events, and applied the exact post-selection inference framework to obtain p-values for the hypothesis testing of interest. Jewell et al. (2019), Mehrizi and Chenouri (2021), and the references therein, studied different variants of point change problem and applied the exact post-selection inference framework to obtain p-values and confidence intervals. To test the significance of a difference in the means of two connected components obtained from the graph fused lasso, Chen et al. (2022) applied the exact post-selection inference framework, their method conditions on less and was shown to be powerful. Gao et al. (2022) considered

testing for a difference in means of a pair of clusters identified via a data-dependent procedure. The authors first showed that sample splitting is not a valid procedure in this case, and they proposed an exact post-selection inference method that conditions on the selected clusters and a truncated chi-squared counterpart is obtained in the computation of p-values. Their R package *clusterpval* computes valid p-values for a difference in means by correcting for double dipping (generating a hypothesis based on your data, and then testing the hypothesis on that same data). Neufeld et al. (2021) applied the exact post-selection inference framework on inference associated with the Classification and Regression Tree (CART) algorithm. Specifically, the authors obtain an exact post-selection inference based p-values for testing a difference in the mean response between a pair of terminal nodes and confidence interval for the mean response within a single terminal.

Chapter 3

Post Outlier & Variable Selection Confidence Intervals for Regularized Regression with Huber Loss

In (1.4), it was discussed that heavy-tailed errors or outliers are frequently encountered in applications, whether in response variables or predictors. The objective of this chapter is to utilize penalized Huber regression as a robust regression estimate to detect outliers and subsequently condition on both the outlier identification and variable selection events during inference. As far as we know, no previous research has addressed the issue of building conditional confidence intervals for generalized penalized Huber regression estimates that account for both outlier identification and variable selection events simultaneously. To formally define the problem, let us consider the Huber-loss regression model with a generalized lasso penalty, where we seek to minimize the objective function given by:

$$\min_{\beta} \sum_{i=1}^{p} \rho_M(y_i - \beta^T x_i) + \lambda ||D\beta||_1 \tag{3.1}$$

where $\rho_M(.)$ is the Huber function defined earlier, $\lambda \geq 0$ is a tuning parameter, $||.||_d$ is the l_d - norm of a vector, $D \in \mathbb{R}^{m \times p}$. We may obtain adaptive lasso penalty, fused lasso penalty, or other penalty functions by appropriately choosing the matrix D. Let $\mathcal{V} = \{i : \hat{v}_i \neq 0\}$, $\mathcal{A} =$

$\{k : D_k\hat{\beta} \neq 0\}$ be the solutions obtained by solving (3.1) and

$$\beta^{\mathcal{A}} = \arg\min_{b} \mathbb{E}(||y_{-\mathcal{V}} - X_{-\mathcal{V},\mathcal{A}}b||^2)$$

suppose that we are interested in a component of $\beta^{\mathcal{A}}$, that is, $e_k^T\beta^{\mathcal{A}}$. Its estimate is $e_k^T\hat{\beta}^{\mathcal{A}} = e_k^T(X_{-\mathcal{V},\mathcal{A}}^T X_{-\mathcal{V},\mathcal{A}})^{-1}X_{-\mathcal{V},\mathcal{A}}^T y_{-\mathcal{V}}$. Hence, the key is to derive the conditional distribution of $\eta^T y_{-\mathcal{V}}$ for some vector η, conditionally on the variables selected by generalized lasso $\mathcal{A}$, and on the selected non-outlying observations $\mathcal{V}^c$.

The outline of this chapter is as follows. Section 1 derives the KKT conditions of the Huber gl-lasso and introduces two active sets $\mathcal{V}, \mathcal{A}$. Section 2 contains our core results, lemma 3.2.1 and Theorems 3.2.2 - 3.2.3 entails our main findings when the error distribution is Gaussian. Section 3 discussed heavy tailed distribution in the context of our problem (non-Gaussian case). Simulation studies are reported in Section 4 and a real data example is discussed in Section 5. Section (3.6) is the appendix

3.1 Preliminaries

Generally, D_k means the k^{th} row of matrix D. Given an index set $\mathcal{A}$, $D_{\mathcal{A}}$ is the matrix obtained by selecting the rows of D corresponding to the indexes in $\mathcal{A}$. Similarly, $D_{-\mathcal{A}}$ is the matrix obtained by selecting the rows of D corresponding to the indexes not in $\mathcal{A}$. The same rule applies to other matrices and index sets, unless otherwise stated. For a matrix A, col(A) denotes the column space of A and null(A) means the null space of A. For the sake of

simplicity in notation, the symbols $-\mathcal{V}$ and $\mathcal{V}^c$ are utilized interchangeably to indicate the complement of $\mathcal{V}$.

KKT Conditions

Here we derive the subgradient conditions that characterize the solution of the generalized Huber - lasso. The Moreau-Yosida regularization of the absolute value function gives us the Huber function Lambert-Lacroix and Zwald (2011). Mathematically, the Huber function is

$$\rho_M(u) = \min_{z\in\mathbb{R}} \frac{1}{2}(u-z)^2 + M|z|,$$

where the minimizer is $z = 0$ if $|u| \leq M$ and $z = u - sign(u)M$ if $|u| > M$. Hence, problem (3.1) is equivalent to

$$\min_{\beta\in\mathbb{R}^p} \min_{v\in\mathbb{R}^n} \sum_{i=1}^{n} \{\frac{1}{2}(y_i - \beta^T x_i - v_i)^2 + M|v_i|\} + \lambda||D\beta||_1$$

where $v = (v_1\,, \ldots, v_n)^T$. Writing in a compact form, we have

$$\min_{\beta,v} \frac{1}{2}||y - X\beta - v||_2^2 + M||v||_1 + \lambda||D\beta||_1 \tag{3.2}$$

To derive it's Karush-Kuhn-Tucker (KKT) conditions, (3.2) was formulated as a quadratic programming problem. Refer to Liu et al. (2021) for some details. Notice that Liu et al. (2021) considered the problem with linear constraints. Here, for simplicity, we ignore the linear constraints. Use the same notations as in Liu et al. (2021). Let $\hat{\beta}$ and $\hat{v}$ be the

minimizer to the above optimization problem (2.3). Define index sets

$$\mathcal{V} = \{i : \hat{v}_i \neq 0\}, \quad \mathcal{A} = \{k : D_k\hat{\beta} \neq 0\}.$$

$$\hat{v}_i = \begin{cases} 0, & i \in -\mathcal{V} \\ r_i - sign(\hat{v}_i)M & i \in \mathcal{V} \end{cases}$$

where $\mathcal{V}$ is the set of indexes corresponding to non - zero components of $\hat{v}$, $\mathcal{A}$ is the set of indexes corresponding to non-zero components of $D\hat{\beta}$. Since $\mathcal{V}$ contains the indexes of observations with large residuals $r_i = y_i - \hat{\beta}^T x_i$ satisfying $|r_i| > M$, i.e., the indexes of observations with large residuals. Liu et al. (2021) established that $y_{-\mathcal{V}}$ explicitly contributes to the estimate $\hat{\beta}$, but $y_{\mathcal{V}}$ does not, and the dependence of $\hat{\beta}$ on $y_{\mathcal{V}}$ is implicit via the sign of $\hat{v}_{\mathcal{V}}$, which overall implies that the estimate $\hat{\beta}$ does not directly depend on $y_{\mathcal{V}}$ to some extent. This explains why the Huber - loss regression is robust to heavy-tailed errors and outliers, see Liu et al. (2021) for more technical details.

3.2 Main Results

Here we propose a method to construct valid confidence intervals for (3.1) while conditioning on the outlier-identification event and the variable selection event. The robustification parameter M is chosen adaptively as recommended by Sun et al. (2020), Pan et al. (2021), and the references therein.

3.2.1 Affine selection procedure

We will show in theorem 3.2.2 that the joint selection event $\{\mathcal{A}, \mathcal{V}^c\}$ alongside their corresponding signs $\{s_\mathcal{A}, s_{-\mathcal{V}}\}$ is an affine selection procedure as defined in definition 2.1.1. To see this, from KKT conditions for (3.1), we have

$$X_{-\mathcal{V}}^T X_{-\mathcal{V}}\hat{\beta} + X_{\mathcal{V}}^T X_{\mathcal{V}}\hat{\beta} + X_{\mathcal{V}}^T\hat{v}_\mathcal{V} + D_{-\mathcal{A}}^T\hat{u}_{-\mathcal{A}} = X_{-\mathcal{V}}^T y_{-\mathcal{V}} + X_{\mathcal{V}}^T y_\mathcal{V} - \lambda D_\mathcal{A}^T s_\mathcal{A}$$

$$X_\mathcal{V}\hat{\beta} + \hat{v}_\mathcal{V} = y_\mathcal{V} - M s_\mathcal{V}$$

$$D_{-\mathcal{A}}\hat{\beta} = 0.$$

where

$$s_\mathcal{A} = \text{sign}(D_\mathcal{A}\hat{\beta}),$$

$$s_\mathcal{V} = \text{sign}(\hat{v}_\mathcal{V}) = \text{sign}(y_\mathcal{V} - X_\mathcal{V}\hat{\beta}),$$

$$\hat{v}_{-\mathcal{V}} = 0,$$

$$\hat{v}_\mathcal{V} = r_\mathcal{V} - s_\mathcal{V} M,$$

$$\hat{u}_\mathcal{A} = \lambda s_\mathcal{A},$$

$$||\hat{u}_{-\mathcal{A}}||_\infty \leq \lambda.$$

We can eliminate $\hat{v}_\mathcal{V}$ and get the following, which is exactly equation (9) in Liu et al. (2021).

$$X_{-\mathcal{V}}^T X_{-\mathcal{V}}\hat{\beta} + D_{-\mathcal{A}}^T\hat{u}_{-\mathcal{A}} = X_{-\mathcal{V}}^T y_{-\mathcal{V}} - \lambda D_\mathcal{A}^T s_\mathcal{A} + M X_\mathcal{V}^T s_\mathcal{V} \tag{3.3}$$

We need to get an explicit expression for $\hat{\beta}$. Notice that

$$\begin{pmatrix} X_{-\mathcal{V}}^T X_{-\mathcal{V}} & D_{-\mathcal{A}}^T \\ D_{-\mathcal{A}} & 0 \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{u}_{-\mathcal{A}} \end{pmatrix} = \begin{pmatrix} X_{-\mathcal{V}}^T y_{-\mathcal{V}} - \lambda D_{\mathcal{A}}^T s_{\mathcal{A}} + M X_{\mathcal{V}}^T s_{\mathcal{V}} \\ 0 \end{pmatrix}.$$

Applying the formula for inverse of a block matrix, we know that

$$\begin{aligned} \hat{\beta} = &((X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} - (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} D_{-\mathcal{A}}^T (D_{-\mathcal{A}} (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} D_{-\mathcal{A}}^T)^{-1} D_{-\mathcal{A}} (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1}) \\ &\cdot (X_{-\mathcal{V}}^T y_{-\mathcal{V}} - \lambda D_{\mathcal{A}}^T s_{\mathcal{A}} + M X_{\mathcal{V}}^T s_{\mathcal{V}}) \end{aligned}$$

and

$$\hat{u}_{-\mathcal{A}} = (D_{-\mathcal{A}} (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} D_{-\mathcal{A}}^T)^{-1} D_{-\mathcal{A}} (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} (X_{-\mathcal{V}}^T y_{-\mathcal{V}} - \lambda D_{\mathcal{A}}^T s_{\mathcal{A}} + M X_{\mathcal{V}}^T s_{\mathcal{V}}).$$

To ease the notation, introduce

$$\begin{aligned} H_{-\mathcal{V},-\mathcal{A}} &= (D_{-\mathcal{A}} (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} D_{-\mathcal{A}}^T)^{-1} D_{-\mathcal{A}} (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} \\ P_{-\mathcal{V},-\mathcal{A}} &= (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} - (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} D_{-\mathcal{A}}^T (D_{-\mathcal{A}} (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} D_{-\mathcal{A}}^T)^{-1} D_{-\mathcal{A}} (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} \\ &= (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} - (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} D_{-\mathcal{A}}^T H_{-\mathcal{V},-\mathcal{A}} \end{aligned}$$

and write $\hat{\beta}$ and $\hat{u}_{-\mathcal{A}}$ as

$$\hat{\beta} = P_{-\mathcal{V},-\mathcal{A}} (X_{-\mathcal{V}}^T y_{-\mathcal{V}} - \lambda D_{\mathcal{A}}^T s_{\mathcal{A}} + M X_{\mathcal{V}}^T s_{\mathcal{V}})$$

$$\hat{u}_{-\mathcal{A}} = H_{-\mathcal{V},-\mathcal{A}} (X_{-\mathcal{V}}^T y_{-\mathcal{V}} - \lambda D_{\mathcal{A}}^T s_{\mathcal{A}} + M X_{\mathcal{V}}^T s_{\mathcal{V}}).$$

Notice that $P_{-\mathcal{V},-\mathcal{A}}$ and $H_{-\mathcal{V},-\mathcal{A}}$ are matrices depending on X, D, $\mathcal{V}$, and $\mathcal{A}$. Right now, we assume that all matrix inversions in calculating $P_{-\mathcal{V},-\mathcal{A}}$ and $H_{-\mathcal{V},-\mathcal{A}}$ are well defined and thus $P_{-\mathcal{V},-\mathcal{A}}$ and $H_{-\mathcal{V},-\mathcal{A}}$ are unique.

We want to derive the confidence intervals conditionally on $(\mathcal{A}, \mathcal{V}^c, \hat{s}_{\mathcal{A}}, \hat{s}_{-\mathcal{V}})$. This event includes the collection of y such that the corresponding $\hat{\beta}$ and $\hat{u}$ leads to eligible solutions.

Lemma 3.2.1 *Let $\mathcal{A}, \mathcal{V}^c, s_{\mathcal{A}}, s_{-\mathcal{V}}$ be properly defined as above, then we have*

$$\{(\mathcal{A}, \mathcal{V}^c, \hat{s}_{\mathcal{A}}, \hat{s}_{-\mathcal{V}}) = (\mathcal{A}, \mathcal{V}^c, s_{\mathcal{A}}, s_{-\mathcal{V}})\} = \{\|\hat{u}_{-\mathcal{A}}\|_\infty \leq \lambda,\ \|y_{-\mathcal{V}} - X_{-\mathcal{V}}\hat{\beta}\|_\infty \leq M,\ sign(D_{\mathcal{A}}\hat{\beta}) = s_{\mathcal{A}},$$
$$sign(y_{\mathcal{V}} - X_{\mathcal{V}}\hat{\beta}) = s_{\mathcal{V}}\}$$

PROOF. Follows from KKT conditions.

Theorem 3.2.2 *Let $A(\mathcal{A}, \mathcal{V}^c, s_{\mathcal{A}}, s_{-\mathcal{V}})$ and $b(\mathcal{A}, \mathcal{V}^c, s_{\mathcal{A}}, s_{-\mathcal{V}})$ be the matrix and vector resulting from the subgradient inequalites in Lemma 3.2.1, then*

$$\{(\mathcal{A}, \mathcal{V}^c, \hat{s}_{\mathcal{A}}, \hat{s}_{-\mathcal{V}}) = (\mathcal{A}, \mathcal{V}^c, s_{\mathcal{A}}, s_{-\mathcal{V}})\} = \{A(\mathcal{A}, \mathcal{V}^c, s_{\mathcal{A}}, s_{-\mathcal{V}})y \leq b(\mathcal{A}, -\mathcal{V}, s_{\mathcal{A}}, s_{-\mathcal{V}})\}$$

i.e.,

$$\begin{pmatrix} 0 & I - X_{-\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}X_{-\mathcal{V}}^T \\ 0 & -I + X_{-\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}X_{-\mathcal{V}}^T \\ 0 & H_{-\mathcal{V},-\mathcal{A}}X_{-\mathcal{V}}^T \\ 0 & -H_{-\mathcal{V},-\mathcal{A}}X_{-\mathcal{V}}^T \\ 0 & -diag(s_{\mathcal{A}})D_{\mathcal{A}}P_{-\mathcal{V},-\mathcal{A}}X_{-\mathcal{V}}^T \\ -diag(s_{\mathcal{V}}) & diag(s_{\mathcal{V}})X_{\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}X_{-\mathcal{V}}^T \end{pmatrix} \begin{pmatrix} y_{\mathcal{V}} \\ y_{-\mathcal{V}} \end{pmatrix} \leq \begin{pmatrix} M1 - X_{-\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}(\lambda D_{\mathcal{A}}^T s_{\mathcal{A}} - MX_{\mathcal{V}}^T s_{\mathcal{V}}) \\ M1 + X_{-\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}(\lambda D_{\mathcal{A}}^T s_{\mathcal{A}} - MX_{\mathcal{V}}^T s_{\mathcal{V}}) \\ \lambda 1 + H_{-\mathcal{V},-\mathcal{A}}(\lambda D_{\mathcal{A}}^T s_{\mathcal{A}} - MX_{\mathcal{V}}^T s_{\mathcal{V}}) \\ \lambda 1 - H_{-\mathcal{V},-\mathcal{A}}(\lambda D_{\mathcal{A}}^T s_{\mathcal{A}} - MX_{\mathcal{V}}^T s_{\mathcal{V}}) \\ -diag(s_{\mathcal{A}})D_{\mathcal{A}}P_{-\mathcal{V},-\mathcal{A}}(\lambda D_{\mathcal{A}}^T s_{\mathcal{A}} - MX_{\mathcal{V}}^T s_{\mathcal{V}}) \\ -diag(s_{\mathcal{V}})X_{\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}(\lambda D_{\mathcal{A}}^T s_{\mathcal{A}} - MX_{\mathcal{V}}^T s_{\mathcal{V}}) - M \end{pmatrix}$$

PROOF. Our approach is to work out each term in lemma 3.2.1 separately.

- $\|\hat{u}_{-\mathcal{A}}\|_\infty \leq \lambda$

  This immediately implies $-\lambda \leq H_{-\mathcal{V},-\mathcal{A}}(X_{-\mathcal{V}}^T y_{-\mathcal{V}} - \lambda D_{\mathcal{A}}^T s_{\mathcal{A}} + MX_{\mathcal{V}}^T s_{\mathcal{V}}) \leq \lambda$ holds componentwise, or

$$H_{-\mathcal{V},-\mathcal{A}}X_{-\mathcal{V}}^T y_{-\mathcal{V}} \leq \lambda 1 + \lambda H_{-\mathcal{V},-\mathcal{A}}D_{\mathcal{A}}^T s_{\mathcal{A}} - MH_{-\mathcal{V},-\mathcal{A}}X_{\mathcal{V}}^T s_{\mathcal{V}}$$

$$-H_{-\mathcal{V},-\mathcal{A}}X_{-\mathcal{V}}^T y_{-\mathcal{V}} \leq \lambda 1 - \lambda H_{-\mathcal{V},-\mathcal{A}}D_{\mathcal{A}}^T s_{\mathcal{A}} + MH_{-\mathcal{V},-\mathcal{A}}X_{\mathcal{V}}^T s_{\mathcal{V}}$$

- $\|y_{-\mathcal{V}} - X_{-\mathcal{V}}\hat{\beta}\|_\infty \leq M$

  For the residuals not in $\mathcal{V}$, this immediately implies $-M \leq y_{-\mathcal{V}} - X_{-\mathcal{V}}\hat{\beta} \leq M$ holds componentwise or by substituting for $\hat{\beta}$ and simplifying to have

$$y_{-\mathcal{V}} - X_{-\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}X_{-\mathcal{V}}^T y_{-\mathcal{V}} \leq M1 - \lambda X_{-\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}D_{\mathcal{A}}^T s_{\mathcal{A}} + MX_{-\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}X_{\mathcal{V}}^T s_{\mathcal{V}}$$

$$-y_{-\mathcal{V}} + X_{-\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}X_{-\mathcal{V}}^T y_{-\mathcal{V}} \leq M1 + \lambda X_{-\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}D_{\mathcal{A}}^T s_{\mathcal{A}} - MX_{-\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}X_{\mathcal{V}}^T s_{\mathcal{V}}$$

- $\text{sign}(D_{\mathcal{A}}\hat{\beta}) = s_{\mathcal{A}}$

  This immediately implies $\text{diag}(s_{\mathcal{A}})D_{\mathcal{A}}\hat{\beta} \geq 0$ or by substituting for $\hat{\beta}$ and simplifying to have

$$\text{diag}(s_{\mathcal{A}})D_{\mathcal{A}}P_{-\mathcal{V},-\mathcal{A}}(X_{-\mathcal{V}}^T y_{-\mathcal{V}} - \lambda D_{\mathcal{A}}^T s_{\mathcal{A}} + M X_{\mathcal{V}}^T s_{\mathcal{V}}) \geq 0$$

  which in turn implies

$$-\text{diag}(s_{\mathcal{A}})D_{\mathcal{A}}P_{-\mathcal{V},-\mathcal{A}}X_{-\mathcal{V}}^T y_{-\mathcal{V}} \leq -\text{diag}(s_{\mathcal{A}})D_{\mathcal{A}}P_{-\mathcal{V},-\mathcal{A}}(\lambda D_{\mathcal{A}}^T s_{\mathcal{A}} - M X_{\mathcal{V}}^T s_{\mathcal{V}})$$

- $\text{sign}(y_{\mathcal{V}} - X_{\mathcal{V}}\hat{\beta}) = s_{\mathcal{V}}$

  This immediately implies $\text{diag}(s_{\mathcal{V}})(y_{\mathcal{V}} - X_{\mathcal{V}}\hat{\beta}) \geq 0$ or by substituting for $\hat{\beta}$ and simplifying to have

$$\text{diag}(s_{\mathcal{V}})y_{\mathcal{V}} - \text{diag}(s_{\mathcal{V}})X_{\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}(X_{-\mathcal{V}}^T y_{-\mathcal{V}} - \lambda D_{\mathcal{A}}^T s_{\mathcal{A}} + M X_{\mathcal{V}}^T s_{\mathcal{V}}) \geq 0$$

  which in turn implies

$$-\text{diag}(s_{\mathcal{V}})y_{\mathcal{V}} + \text{diag}(s_{\mathcal{V}})X_{\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}X_{-\mathcal{V}}^T y_{-\mathcal{V}} \leq -\text{diag}(s_{\mathcal{V}})X_{\mathcal{V}}P_{-\mathcal{V},-\mathcal{A}}(\lambda D_{\mathcal{A}}^T s_{\mathcal{A}} - M X_{\mathcal{V}}^T s_{\mathcal{V}})$$

3.2.2 Gaussian errors

To compute the desired confidence intervals, we'll first try to find the distribution for $\eta^T y_{-\mathcal{V}}$ conditional on $(\mathcal{A}, \mathcal{V}^c, \hat{s}_{\mathcal{A}}, \hat{s}_{-\mathcal{V}})$ when the error is Gaussian, where $\eta \in \mathbb{R}^n$ is some direction of

interest. From theorem 3.2.2 above, we have written our selection event $\{(\mathcal{A}, \mathcal{V}^c, \hat{s}_{\mathcal{A}}, \hat{s}_{-\mathcal{V}}) = (\mathcal{A}, \mathcal{V}^c, s_{\mathcal{A}}, s_{-\mathcal{V}})\}$ as affine inequality in y, therefore we have from lemma 2.1.2 that

$$\{A(\mathcal{A}, \mathcal{V}^c, s_{\mathcal{A}}, s_{-\mathcal{V}})y \leq b(\mathcal{A}, \mathcal{V}^c, s_{\mathcal{A}}, s_{-\mathcal{V}})\} = \{L(z) \leq \eta^T y \leq U(z)\}$$

and, from lemma 2.2.1, the distribution of $\eta^T y_{-\mathcal{V}}$ conditioned on the selection event $A(\mathcal{A}, \mathcal{V}^c, s_{\mathcal{A}}, s_{-\mathcal{V}})$, follows a univariate truncated gaussian, with $L(z),\ U(z)$ as truncation limits. Therefore, we can now make conditional inference on the population regression coefficients of $y_{-\mathcal{V}}$ on $X_{-\mathcal{V},\mathcal{A}}$, i.e., we can construct exact and valid confidence intervals for the parameters of the active set in the Huber gl-lasso solution at fixed λ, with correct coverage conditional on the active sets and their signs. Our result is summarized in the next Theorem.

Theorem 3.2.3 (Main result1) *Let $\mathcal{A}$, $\mathcal{V}^c$ be active sets for (3.1) with signs $s_{\mathcal{A}}$, $s_{-\mathcal{V}}$ when the error is Gaussian, where λ is fixed and M is chosen adaptively. Let $\eta = e_j X_{\mathcal{A},-\mathcal{V}}(X^T_{\mathcal{A},-\mathcal{V}} X_{\mathcal{A},-\mathcal{V}})^{-1}$, then $[L^*,\ U^*]$ is a $(1-\alpha)$ confidence interval for $\beta_j = \eta^T y_{-\mathcal{V}}$ conditional on $\{(\mathcal{A}, \mathcal{V}^c, \hat{s}_{\mathcal{A}}, \hat{s}_{-\mathcal{V}}) = (\mathcal{A}, \mathcal{V}^c, s_{\mathcal{A}}, s_{-\mathcal{V}})\}$, i.e.,*

$$P(\beta_j \in [L^*, U^*] \,|\, \hat{\mathcal{A}} = \mathcal{A}, \hat{\mathcal{V}}^c = \mathcal{V}^c, \hat{s_{\mathcal{A}}} = s_{\mathcal{A}}, \hat{s}_{-\mathcal{V}} = s_{-\mathcal{V}}\ ) \geq 1 - \alpha$$

where L^* & U^* are unique values satisfying

$$F^{[L(z),U(z)]}_{L^*,\ \sigma^2||\eta||^2}(\eta^T y_{-\mathcal{V}}) = 1 - \frac{\alpha}{2}, \quad F^{[L(z),U(z)]}_{U^*,\ \sigma^2||\eta||^2}(\eta^T y_{-\mathcal{V}}) = \frac{\alpha}{2}$$

3.3 Generalizing to heavier tailed distributions (non-Gaussian errors)

As we mentioned in 2.3, if we remove the assumption that the error is Gaussian, then (2.3) will be false, and subsequently, the conclusion of Theorem 2.2.2 does not hold anymore. The best we can hope for is a weak convergence result that the same pivotal quantities (2.3) would converge to $\mathcal{U}(0,1)$ (as $n \to \infty$). Our main goal in this section is to show that given $\mathcal{L}(y|X) \equiv$ errors that are heavy tailed distributed with bounded $(1+\delta)$-th moment for any $\delta > 0$, that the conclusion of Theorem 3.2.3 still holds true.

We want to apply Theorem 2.3.1 to perform post selection inference after solving the huber-lasso (3.2) with heavy tailed error distribution, take $D = \mathcal{I}_{p\times p}$ in (3.2) for simplicity. To use the framework of Tian and Taylor (2017), we will explain why the selection procedure is affine, state the distribution $\mathcal{L}(y_n|X_n)$ and the quantities $(\gamma_n, M(\mathcal{E}_n^*, \eta_n), r(\mathcal{E}_n^*), |\mathcal{S}_n|)$. We suppress the dependencies on n when possible to help ease notations.

3.3.1 Affine selection precedure

Consider (3.2), where

$$\mathbb{E}(\varepsilon_i|x_i) = 0, \quad v_{i,\delta} = \mathbb{E}(|\varepsilon_i|^{1+\delta}) < \infty \tag{3.4}$$

i.e., $\mathcal{L}(y|X) \equiv$ errors that are heavy tailed distributed with bounded $(1+\delta)$-th moment for any $\delta > 0$, we solve to obtain the active sets $\mathcal{V}, \mathcal{A}$. Define $E = \{\mathcal{V}^c, \mathcal{A}\}$, and the signs $z_E = \{s_\mathcal{A}, s_{-\mathcal{V}}\}$, where X_E implies rows of X not in $\mathcal{V}$ and columns of X in $\mathcal{A}$, and $\hat{\beta}_E$ is $\hat{\beta}$ restricted to the set E, which can be taken as $\hat{\beta}_\mathcal{A}$, i.e., $\hat{\beta}$ restricted to the set $\mathcal{A}$. We

showed in section (3.2.1) that such selection procedure is equivalent to the affine constraints $A(E, z_E)y \leq b(E, z_E)$, where $A(E, z_E)$ and $b(E, z_E)$ are explicitly given in Theorem 3.2.2.

3.3.2 Bounding $|\mathcal{S}|$

The collection of possible interesting questions alongside the collection of possible non-outlying observations are

$$\mathcal{S} = \{(\mathcal{V}^c, \mathcal{A}) : \mathcal{V}^c \subset \{1, \ldots, n\}, \mathcal{A} \subset \{1, \ldots, p\}\}$$

We want to ensure $|\mathcal{S}|$ is polynomial in p with high probability. We begin by rewriting (3.2) as

$$\min_{\tilde{\beta}} \frac{1}{2}||y - \tilde{X}\tilde{\beta}||_2^2 + \tilde{\lambda}||\tilde{\beta}||_1 \tag{3.5}$$

where

$$\tilde{X} = \begin{pmatrix} X \\ I \end{pmatrix}, \quad \tilde{\beta} = \begin{pmatrix} \beta \\ v \end{pmatrix}, \quad \tilde{\lambda} = \begin{pmatrix} \lambda \\ M \end{pmatrix}, \text{ and define } \tilde{\varepsilon} = \begin{pmatrix} \varepsilon \\ 0 \end{pmatrix}.$$

Definition 3.3.1 *(Restricted strong convexity Negahban et al. (2012), Compatibility condition Bühlmann and Van De Geer (2011)). We say $X \in \mathbb{R}^{n \times p}$ satisfies the restricted strong convexity condition or compatibility condition for the set A with constant $m > 0$ if*

$$||Xv||_2^2 \geq m||v||_2^2,$$

for all $v \in \{\nabla \in \mathbb{R}^p : ||\Delta_{A^c}||_1 \leq 3||\Delta_A||_1\}$

To establish a bound for all the possible subsets of $\mathcal{S}$, we make the following assumptions,

Assumption 2 *$\tilde{X}$ satisfies the restricted strong convexity condition or compatibility condition for $A = supp(\tilde{\beta})$ with constant m, where $supp(\tilde{\beta}) = supp(\beta) \cup supp(v)$, and ϕ_{max}, the biggest eigenvalue of $\tilde{X}^T\tilde{X}$ is bounded by a constant Q*

Assumption 3 *The signal is sparse, i.e., $k = |supp(\tilde{\beta})|$ is bounded by a constant K, where $k = s_1 + s_2$, and $s_1 = |supp(\beta)|$, $s_2 = |supp(v)|$*

Assumption 4 *ε_i is as defined in (3.4)*

Following Negahban et al. (2012) and Tian and Taylor (2017), we have the following Lemmas

Lemma 3.3.2 *With Assumptions 2 - 4, if we solve (3.5) with the appropriate $\tilde{\lambda}$ and get an active set E_1, then,*

$$|E_1| \leq Q^2 c$$

Observe that $E_1 \equiv \{\mathcal{V}, \mathcal{A}\}$ which has a one-to-one correspondence with $\{\mathcal{V}^c, \mathcal{A}\}$, therefore, there is a bound for all the possible subsets of $\{\mathcal{V}^c, \mathcal{A}\}$, hence

$$|\mathcal{S}| \leq \mathcal{K}, \quad for\ some\ \mathcal{K}.$$

3.3.3 Bounding $M(\mathcal{E}^*, \eta)$

Lemma 3.3.3 *Consider the matrix $A(\mathcal{A}, \mathcal{V}^c, s_{\mathcal{A}}, s_{-\mathcal{V}})$ in (3.2.2), then for some constant $\mathcal{C}$*

$$\max_{i,j} |[A(\mathcal{A}, \mathcal{V}^c, s_{\mathcal{A}}, s_{-\mathcal{V}})]_{i,j}| \leq \mathcal{C} \cdot \max_{i,j} |X_{i,j}|$$

3.3.4 Choice of γ_n in Assumption 1

Suppose we normalize the columns of X to have norm 1 and, take $t = (1+c)\log p$ for $c > 0$, $\lambda = O((\log p)^{\min\{\frac{1}{1+\delta},\frac{1}{2}\}})$. Then we assume that Assumption 1 is satisfied with $\gamma_n = O(((\log p_n)^{\min\{\frac{1}{1+\delta},\frac{1}{2}\}})^{-1-\omega})$, for any small $\omega > 0$.

Lemma 3.3.4 *We assume that $z_E = 1$ and the matrix $(X_E^T X_E)^{-1}$ is equicorrelated, i.e.,*

$$((X_E^T X_E)^{-1})_{ii} = ((X_E^T X_E)^{-1})_{jj} = \tau_1 > 0,$$

$$\rho = \frac{((X_E^T X_E)^{-1})_{ii}}{((X_E^T X_E)^{-1})_{jj}} > 0, \forall i,\, j \in E,\, i \neq j.$$

For any $\omega > 0$, Assumption 1 is satisfied with $\gamma_n = O(\lambda_n^{-1-\omega})$, provided $||\tilde{\beta}_n||_\infty = O(\lambda_n)$

Theorem 3.3.5 (Main Result2) *Suppose we solve the Huber - lasso problem (3.2) with $M = C(\frac{\log p}{n})^{\min\{\delta/(1+\delta),1/2\}}$, $\lambda \geq C(\frac{\log p}{n})^{\min\{\delta/(1+\delta),1/2\}}$ and $C > 0$, Assumptions 1 - 4 are satisfied, and γ_n in Assumption 1 is chosen as $((\log p_n)^{\min\{\frac{1}{1+\delta},\frac{1}{2}\}})^{-1-\omega}$, and if we also assume $\max|X_{ij}| = O(n^{-1/2})$, $||\beta||_\infty = O((\log p)^{\min\{\frac{1}{1+\delta},\frac{1}{2}\}})$, there exists $\omega > 0$ such that*

$$n^{-1/2}[\log(2p_n) + \log p_n^{\mathcal{K}}]^4[(\log p_n)^{\min\{\frac{1}{1+\delta},\frac{1}{2}\}}]^{6+6\omega} \to 0$$

then the conclusion of theorem 3.2.3 holds true.

3.4 Simulation Study

Design of Experiment

To conduct our experiment, we will perform the following steps:

The first step of our experiment will involve randomly generating a design matrix of dimension $n \times p$. This design matrix will be used to generate a response vector y, which we will use to evaluate the performance of our proposed method. To generate the design matrix, we will use a multivariate normal distribution, where each row of the matrix is independently sampled with a mean of zero and covariance matrix with entries $0.5^{|i-j|}$ for $i, j = 1, 2, \ldots, p$. Once we have generated the design matrix, we will use it to calculate the response vector y as $y = X\beta + \varepsilon$. Here, β is a vector with only five non-zero values at the $1^{st}, 2^{nd}, 6^{th}, 7^{th}$, and 8^{th} components, with values (1, 1.5, 0, 0, 0, -1, -1.5, 2, 0, ..., 0). The random error term ε will be generated using three different distributions, namely Normal distribution, Student's t distribution with a small degrees of freedom, and Mixed normal distribution. By using different error terms, we will test the robustness of our proposed method and compare its performance with the existing method. The next step will involve choosing the values of λ and M according to the method outlined in Sun et al. (2020). To evaluate the performance of our proposed method, we will track a metric that best demonstrates its superiority. After extensive literature review, we have chosen to track the Average Coverage Probability. This metric provides an overall measure of how well the method performs in constructing confidence intervals for the coefficients. We will compare this metric for our proposed method and the existing method presented in Lee et al. (2016) to evaluate the performance of both methods. We will also report the average interval length accompanied by its standard deviation, highlighted in orange.

Based on the tables presented above, our method, Huber-Cond, outperformed the method proposed by Lee et al. (2016) in terms of coverage probability, particularly for heavy-tailed errors such as $t_{1.5}$ and Mix-Normal. This improvement can be attributed to the fact that

Table 3.1: (n, p) = (400, 10)

Huber-Cond			**lasso-Cond**	
Errors	Coverage Prob	Ave Length	Coverage Prob	Ave Length
$t_{1.5}$	95.8%	5.89 1.15	92%	3.23 1.06
Mix-Normal	95.2%	4.78 1.11	89.1%	2.90 1.10
Normal	95.01%	2.347 0.76	95%	2.233 0.73

Table 3.2: (n, p) = (4000, 10)

Huber-Cond			**lasso-Cond**	
Errors	Coverage Prob	Ave Length	Coverage Prob	Ave Length
$t_{1.5}$	94.92%	4.02 0.83	94.4%	3.67 0.81
Mix-Normal	95.4%	2.066 0.86	93.8%	1.226 0.62
Normal	95.02%	1.0324 0.54	95.01%	1.0323 0.48

Huber-Cond conditions on both the variable selection event and the outlier selection event. The Huber-Cond method has a slightly larger average interval length than the lasso-Cond method. This difference in length is due to the fact that the former also conditions on the non-outlying observations and their signs, resulting in a slight reduction in its power. However, increasing the sample size, n, led to a substantial enhancement in the performance of our method, especially in terms of the average interval length. On the other hand, for large values of p in table 3.3, the methods were distorted, particularly for heavier tailed errors. Nevertheless, our method, Huber-Cond, still performed better in terms of generating confidence intervals.

Table 3.3: (n, p) = (100, 400)

Huber-Cond			**lasso-Cond**	
Errors	Coverage Prob	Ave Length	Coverage Prob	Ave Length
$t_{1.5}$	91.7%	6.01 1.92	89.3%	4.31 1.97
Mix-Normal	93.4%	5.23 1.69	92.1%	3.64 1.77
Normal	94%	1.76 0.92	94.7%	1.52 0.82

3.5 Real data analysis

In this section, we apply the proposed conditional adaptive Huber regression method to analyze data on the relationship between Information and Communication Technology (ICT) and bank performance. To demonstrate the robustness of our proposed method in the presence of outliers, we compare it with the conditional lasso and naive OLS methods. The banking industry plays a crucial role in driving economic growth and development as it serves as a hub for resource pooling in many societies. Given the fast-changing business environments, ICT has offered tremendous opportunities for the banking industry to scale up, innovate, and respond to these changes, thereby improving service delivery and increasing accessibility to financial services. However, compared to other regions of the world, the banking industry in Sub-Saharan Africa (SSA) has one of the least ICT penetration indices, possibly due to factors such as corruption and under-development. In this study, we aim to investigate the relationship between ICT and bank performance in SSA. We collected our data from the Financial Development and Structure Dataset (FDSD), Financial Access Survey (FAS), and the World Development Indicators (WDI) published by the World Bank and the International Monetary Fund (IMF). The dataset spans 15 years, from 2004 to 2018, with

a total of $n = 525$ samples from 35 Sub-Saharan African countries. The response variable Y is the Return on Assets (ROA), and the explanatory variables include a set of proxies for measuring ICT: Return on Equity (X_1), No of ATM (X_2), Net Interest Margin (X_3), Capital Ratio (X_4), Liquid Asset Ratio (X_5), ATMs per 100,000 adults (X_6), ATMs per $1000, , km^2$ (X_7), Inflation (X_8), GDP (X_9), log of GDP (X_{10}), and log of Number of ATM (X_{11}). To demonstrate the performance of our proposed method, we artificially increased the values of some large observations of Y and decreased the values of some small observations of Y by a factor. We then compared our proposed method with the conditional lasso and naive OLS methods in terms of their ability to handle outliers in the response variable.

In comparing different variable selection techniques, we found that the lasso method selects a model with 11 variables, while the adaptive Huber technique selects a model with all variables except for X_{10}. Our proposed method involves constructing confidence intervals that condition on both the outlier selection event and the variable selection event, and it identifies X_2, X_3, X_4, and X_{11} as significant based on their corresponding confidence intervals. In contrast, the lasso method conditioned on the variable selection event alone, as introduced in Lee et al. (2016), identified X_2, X_3, X_4, X_{11}, and X_5 as significant. To assess the performance of these methods, we refitted our data Y to the selected subspace of X, which corresponds to $X_{\mathcal{A}}$, where $\mathcal{A}$ is the active set obtained from fitting the adaptive Huber regression. Naive confidence intervals declared X_2, X_3, X_4, X_{11}, X_5, and X_1 as significant. However, our proposed method, which accounts for both the outlier selection event and the variable selection event, performed the best, given the presence of outliers in our data. Lee's method, which only accounted for the variable selection event, performed

better than the naive confidence intervals, but still resulted in costly false positives, as X_5 was incorrectly identified as significant. Without accounting for both the outlier selection event and the variable selection event, X_5 and X_1 would have been incorrectly identified as significant, resulting in costly false positives in this context. To provide additional context, the association between the number of ATMs per 100,000 adults (X_5) and Return on Assets in sub-Saharan Africa may be highly dependent on a range of contextual factors. These could include the regulatory environment, consumer preferences, and infrastructure availability, all of which could play a significant role in determining the impact of ATM accessibility on Return on Assets in the region. For example, the cost of establishing and maintaining ATMs could be a significant financial burden for financial institutions, particularly in areas with limited infrastructure and high operating costs. If banks and other financial institutions are required to make substantial investments in ATM infrastructure to meet regulatory requirements or meet customer demand, they may be compelled to divert resources away from other areas of the business, such as marketing, customer service, or new product development. Likewise, with regards to Return on Equity (X_1), banks and financial institutions in sub-Saharan Africa may prioritize maximizing returns for their shareholders, potentially leading to less investment in areas such as research and development or customer service, which could ultimately result in higher costs and reduced profitability. This could be especially pertinent in sub-Saharan Africa, where operating costs and infrastructure costs can be considerable, and customers may be less loyal to a particular institution. Thus, it is important to recognize that the significance of X_5 and X_1 in relation to Return on Assets cannot be confidently declared.

3.6 Appendix

Definitions

1. **Compatibility condition** (Bühlmann and Van De Geer (2011), page 106). Consider a fixed design matrix X. We define the following:

The compatibility condition holds if for some $\phi_0 > 0$ and all β satisfying $||\beta_{s_0^c}||_1 \leq 3||\beta_{s_0}||_1$,

$$||\beta_{s_0}||_1^2 \leq \beta^T \hat{\Sigma} \beta s_0 / \phi_0^2, \quad \hat{\Sigma} = n^{-1} X^T X. \tag{3.6}$$

The number ϕ_0 is called the compatibility constant.

2. A function f satisfies **strong convexity** at β_1 with respect to $\mathcal{C}$ if there is a constant $\gamma > 0$ such that

$$\frac{\nu^T \nabla^2 f(\beta) \nu}{||\nu||_2^2} \geq \gamma \text{ for all nonzero } \nu \in \mathcal{C} \tag{3.7}$$

and for all $\beta \in \mathbb{R}^p$ in a neighborhood of β_1. In the case of linear regression, this reduces to lower bounding the restricted eigenvalues of the model matrix, i.e.,

$$\frac{\frac{1}{n}\nu^T X^T X \nu}{||\nu||_2^2} \geq \gamma \text{ for all nonzero } \nu \in \mathcal{C} \tag{3.8}$$

where $\mathcal{C}(S, \alpha) := \{\nu \in \mathbb{R}^p \mid ||\nu_{S^c}||_1 \leq \alpha ||\nu_S||_1\}$ for some $\alpha \geq 1$

3.6.1 Proof of Theorem 3.2.3

The proof is by directly applying polyhedral lemma from Lee et al. (2016)

3.6.2 Proof of Lemma 3.3.2

Assume $\tilde{\lambda} \geq 2||\tilde{X}\tilde{\varepsilon}||_\infty$ and according to KKT conditions,

$$\begin{cases} \tilde{x}_j^T(\tilde{y} - \tilde{X}\hat{\tilde{\beta}}) = \tilde{\lambda}\text{sign}(\hat{\tilde{\beta}}) & \text{if } j \in E_1 \\ |\tilde{x}_j^T(\tilde{y} - \tilde{X}\hat{\tilde{\beta}})| \leq \tilde{\lambda} & \text{if } j \notin E_1 \end{cases}$$

For any j,

$$\begin{aligned} \tilde{x}_j^T(\tilde{y} - \tilde{X}\hat{\tilde{\beta}}) &= \tilde{x}_j^T(\tilde{X}\tilde{\beta} - \tilde{X}\hat{\tilde{\beta}} + \tilde{\varepsilon}) \\ &= \tilde{x}_j^T\tilde{X}(\tilde{\beta} - \hat{\tilde{\beta}}) + \tilde{x}_j^T\tilde{\varepsilon} \\ &\geq \tilde{x}_j^T\tilde{X}(\tilde{\beta} - \hat{\tilde{\beta}}) - \frac{\tilde{\lambda}}{2} \end{aligned}$$

From the proof of lemma 1 in Tian and Taylor (2017), thus for $j \in E_1$, we have

$$||\tilde{X}^T\tilde{X}(\hat{\tilde{\beta}} - \tilde{\beta})||_2^2 \geq \frac{\tilde{\lambda}^2}{4}|\text{supp}(\hat{\tilde{\beta}})|$$

Also,

$$\begin{aligned} ||\tilde{X}^T\tilde{X}(\tilde{\beta} - \hat{\tilde{\beta}})||_2^2 &\leq ||\tilde{X}^T\tilde{X}||_2^2||\hat{\tilde{\beta}} - \tilde{\beta}||_2^2 \\ &\leq \frac{1}{n^2}||\tilde{X}||_2^2 \cdot O_{\tilde{\lambda}} \\ &\leq \phi_{max}^2 \cdot O_{\tilde{\lambda}} \end{aligned}$$

where $||\hat{\tilde{\beta}} - \tilde{\beta}||_2^2 \leq O_{\tilde{\lambda}}$, and the bound $O_{\tilde{\lambda}}$ depends on $\tilde{\lambda} \geq 2||\tilde{X}\tilde{\varepsilon}||_\infty$

Combining the two inequalities, we have that

$$|\text{supp}(\hat{\tilde{\beta}})| \leq \frac{\phi_{max}^2 \cdot O_{\tilde{\lambda}}}{\tilde{\lambda}^2}$$

3.6.3 Proof of Lemma 3.3.3

Recall that

$$H_{-\mathcal{V},-\mathcal{A}} = (D_{-\mathcal{A}}(X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} D_{-\mathcal{A}}^T)^{-1} D_{-\mathcal{A}}(X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1}$$

$$P_{-\mathcal{V},-\mathcal{A}} = (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} - (X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} D_{-\mathcal{A}}^T H_{-\mathcal{V},-\mathcal{A}}$$

Then, we have

$$D_{-\mathcal{A}}(X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} D_{-\mathcal{A}}^T = D_{-\mathcal{A}} \begin{pmatrix} X_{-\mathcal{V},-\mathcal{A}}^T X_{-\mathcal{V},-\mathcal{A}} & X_{-\mathcal{V},-\mathcal{A}}^T X_{-\mathcal{V},\mathcal{A}} \\ X_{-\mathcal{V},\mathcal{A}}^T X_{-\mathcal{V},-\mathcal{A}} & X_{-\mathcal{V},\mathcal{A}}^T X_{-\mathcal{V},\mathcal{A}} \end{pmatrix}^{-1} D_{-\mathcal{A}}^T$$

Applying the formula for inverse of a block matrix and rearranging the matrix $D_{-\mathcal{A}}$ as an appropriate block matrix, we have

$$(D_{-\mathcal{A}}(X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1} D_{-\mathcal{A}}^T)^{-1} = (X_{-\mathcal{V},-\mathcal{A}}^T X_{-\mathcal{V},-\mathcal{A}}) - X_{-\mathcal{V},-\mathcal{A}}^T X_{-\mathcal{V},\mathcal{A}} (X_{-\mathcal{V},\mathcal{A}}^T X_{-\mathcal{V},\mathcal{A}})^{-1} X_{-\mathcal{V},\mathcal{A}}^T X_{-\mathcal{V},-\mathcal{A}}$$

So, therefore $H_{-\mathcal{V},-\mathcal{A}}$ becomes

$$H_{-\mathcal{V},-\mathcal{A}} = \left[(X_{-\mathcal{V},-\mathcal{A}}^T X_{-\mathcal{V},-\mathcal{A}}) - (X_{-\mathcal{V},-\mathcal{A}}^T X_{-\mathcal{V},\mathcal{A}})(X_{-\mathcal{V},\mathcal{A}}^T X_{-\mathcal{V},\mathcal{A}})^{-1} (X_{-\mathcal{V},\mathcal{A}}^T X_{-\mathcal{V},-\mathcal{A}}) \right] D_{-\mathcal{A}}(X_{-\mathcal{V}}^T X_{-\mathcal{V}})^{-1}$$

To ease notations, we suppress indexes where possible, keep in mind that here, $H = H_{-\mathcal{V},-\mathcal{A}}$, $P = P_{-\mathcal{V},-\mathcal{A}}$, and $A = A(\mathcal{A}, \mathcal{V}, s_{\mathcal{A}}, s_{\mathcal{V}})$. Therefore,

$$
\begin{aligned}
||H||_{max} &= || \left[(X^T_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V},-\mathcal{A}}) - (X^T_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V},\mathcal{A}})(X^T_{-\mathcal{V},\mathcal{A}} X_{-\mathcal{V},\mathcal{A}})^{-1}(X^T_{-\mathcal{V},\mathcal{A}} X_{-\mathcal{V},-\mathcal{A}}) \right] D_{-\mathcal{A}}(X^T_{-\mathcal{V}} X_{-\mathcal{V}})^{-1} ||_{max} \\
&\leq || \left[(X^T_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V},-\mathcal{A}}) - (X^T_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V},\mathcal{A}})(X^T_{-\mathcal{V},\mathcal{A}} X_{-\mathcal{V},\mathcal{A}})^{-1}(X^T_{-\mathcal{V},\mathcal{A}} X_{-\mathcal{V},-\mathcal{A}}) \right] D_{-\mathcal{A}}(X^T_{-\mathcal{V}} X_{-\mathcal{V}})^{-1} ||_2 \\
&\leq ||(X^T_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V},-\mathcal{A}}) - (X^T_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V},\mathcal{A}})(X^T_{-\mathcal{V},\mathcal{A}} X_{-\mathcal{V},\mathcal{A}})^{-1}(X^T_{-\mathcal{V},\mathcal{A}} X_{-\mathcal{V},-\mathcal{A}})||_2 ||D_{-\mathcal{A}}(X^T_{-\mathcal{V}} X_{-\mathcal{V}})^{-1}||_2
\end{aligned}
$$

Applying triangular inequality, we have

$$
\leq \left[||(X^T_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V},-\mathcal{A}})||_2 + ||(X^T_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V},\mathcal{A}})(X^T_{-\mathcal{V},\mathcal{A}} X_{-\mathcal{V},\mathcal{A}})^{-1}(X^T_{-\mathcal{V},\mathcal{A}} X_{-\mathcal{V},-\mathcal{A}})||_2 \right] ||D_{-\mathcal{A}}(X^T_{-\mathcal{V}} X_{-\mathcal{V}})^{-1}||_2
$$

which implies

$$
\max_{i,j} |H_{i,j}| = ||H||_{max} \leq \mathcal{N}_1
$$

where $\mathcal{N}_1$ comprises of values from the euclidean norm of the individual terms.

Similarly, we have

$$
\max_{i,j} |P_{i,j}| = ||P||_{max} \leq \mathcal{N}_2
$$

where $\mathcal{N}_2$ comprises of norm of terms in P.

Now, we try to bound the leading terms in A, we have

$$\begin{aligned}
||I - X_{-\mathcal{V}} P_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V}}^T||_{max} &\leq ||I||_{max} + ||X_{-\mathcal{V}} P_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V}}^T||_{max} \\
&\leq 1 + ||X_{-\mathcal{V}}||_2 ||P_{-\mathcal{V},-\mathcal{A}}||_{max} ||X_{-\mathcal{V}}^T||_{max} \\
&= \frac{\min_{i,j} |X_{i,j}|}{\min_{i,j} |X_{i,j}|} + ||X_{-\mathcal{V}}||_2 ||P_{-\mathcal{V},-\mathcal{A}}||_{max} ||X_{-\mathcal{V}}^T||_{max} \\
&\leq \frac{\max_{i,j} |X_{i,j}|}{k} + \mathcal{B}_1 \mathcal{N}_2 \max_{i,j} |X_{i,j}| \\
&= \left(\frac{1}{k} + \mathcal{B}_1 \mathcal{N}_2 \right) \max_{i,j} |X_{i,j}|
\end{aligned}$$

where $\min_{i,j} |X_{i,j}| \geq k > 0$, $||X_{-\mathcal{V}}||_2 = \mathcal{B}_1$, and $||X_{-\mathcal{V}}^T||_{max} \leq \max_{i,j} |X_{i,j}|$.

$$||H_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V}}^T||_{max} \leq \mathcal{N}_1 \max_{i,j} |X_{i,j}|$$

$$\begin{aligned}
||D_{\mathcal{A}} P_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V}}^T||_{max} &\leq ||D_{\mathcal{A}}||_{max} ||P_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V}}^T||_{max} \\
&\leq \mathcal{N}_2 \max_{i,j} |X_{i,j}|
\end{aligned}$$

$$\begin{aligned}
||X_{\mathcal{V}} P_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V}}^T||_{max} &\leq ||X_{\mathcal{V}}||_2 ||P_{-\mathcal{V},-\mathcal{A}} X_{-\mathcal{V}}^T||_{max} \\
&\leq \mathcal{B}_2 \mathcal{N}_2 \max_{i,j} |X_{i,j}|, \qquad \text{where } ||X_{\mathcal{V}}||_2 \leq \mathcal{B}_2.
\end{aligned}$$

Therefore,

$$\max_{i,j} |A_{i,j}| = ||A||_{max} \leq \mathcal{C} \max_{i,j} |X_{i,j}|$$

where $\mathcal{C} = \max\{\mathcal{N}_1, \mathcal{N}_1, (\frac{1}{k} + \mathcal{B}_1\mathcal{N}_2), \mathcal{B}_2\mathcal{N}_2\}$

3.6.4 Proof of Lemma 3.3.4

See proof of Lemma 3 in Tian and Taylor (2017)

3.6.5 Proof of Theorem 3.3.5

From Lemma 3.3.2 we have

$$|\mathcal{S}_n| \leq \mathcal{K}_n$$

Let $\mathcal{E}_n^*$ be a sequence of affine selection procedures. From Lemma 3.3.3 we have that

$$M(\mathcal{E}_n^*, \eta_n) \leq \mathcal{C} \cdot \max_{i,j} |X_{i,j}|, \ \max_{i,j} |X_{i,j}| = O(n^{-1/2}) \quad \text{and} \ \ \eta_n = X_{En}(X_{E_n}{}^T X_{En})^{-1} e$$

choose $\gamma_n = ((\log p_n)^{\min\{\frac{1}{1+\delta}, \frac{1}{2}\}})^{-1-\omega}$ for Assumption 1. Also, we have that $r(\mathcal{E}_n^*) \leq 2(n+p) < 2(2p)$. Then $\frac{1}{\gamma_n^6} \cdot M(\mathcal{E}_n^*, \eta_n)^3 \cdot n[\ \log(r(\mathcal{E}_n^*)) + \log(|\mathcal{S}_n|)]^{\ 4} \to 0$, as $n \to \infty$ (2.3.1) simplifies to

$$n^{-1/2}[\log(2p_n) + \log \mathcal{K}_n]^4 [(\log p_n)^{\min\{\frac{1}{1+\delta}, \frac{1}{2}\}}]^{6+6\omega} \to 0$$

hence from Theorem 2.3.1, the proof is complete.

Chapter 4

Post Selection Inference With Randomization

The randomization technique involves adding a noise term to the response variable in the model, specifically, we draw $\omega \sim Q$ and use the randomized response $y^*(y, \omega) = y + \omega$ for selection, where Q can be gaussian, logistic, etc, while inference based on the selected model is performed with the original data (data without randomization). The use of a randomized response variable for selective inference has several benefits. These procedures tend to yield more powerful statistical tests, while only incurring a small cost in terms of the quality of the selected models. In other words, the inclusion of a small amount of randomization has a minimal impact on the model selection process but results in a significant increase in the power of inferences made using the model. One reason for the improved power of these procedures is the concept of *leftover Fisher information* which was first introduced by Fithian et al. (2017). This concept refers to the additional information about the parameters of a statistical model that is gained through the use of a randomized response variable. By incorporating this additional information, inferences made using the model are more accurate and reliable. Overall, the use of a randomized response variable in linear regression analysis can enhance the validity and precision of statistical inferences. In this chapter, we will formulated the randomized version of (3.1), and then establish the technique for conditional post selection inference.

4.1 Selective Inference for Randomized Huber Regression

Recall that, from the Moreau-Yosida regularization of the absolute value function, problem (3.1) is equivalent to

$$\min_{\beta\in\mathbb{R}^p}\min_{v\in\mathbb{R}^n}\sum_{i=1}^{n}\{\frac{1}{2}(y_i-\beta^T x_i-v_i)^2+M|v_i|\}+\lambda||D\beta||$$

where $v=(v_1\,,\,\ldots\,,\,v_n)^T$. Writing in a compact form, we have

$$\min_{\beta,v}\frac{1}{2}||y-X\beta-v||_2^2+M||v||_1+\lambda||D\beta||_1 \tag{4.1}$$

For simplicity, we take $D=I$, the randomized version of (4.1) becomes

$$\min_{\beta,v}\frac{1}{2}||y-X\beta-v||_2^2+M||v||_1+\lambda||\beta||_1-\omega^T\beta+\frac{\epsilon}{2}||\beta||_2 \tag{4.2}$$

In this problem, ω represents the added randomization, which is modeled as a random variable drawn from a known distribution $\mathcal{N}(0,\tau^2 I_p)$. After solving the randomized objective (4.2), variable selection output can be described as $E:\hat{E}(y,\omega)=E$, $s_E\in\{\pm 1\}^{|E|}$ are their signs, which is the set of indexes corresponding to non-zero components of $\hat{\beta}$. And the outlier selection output can be described as $\mathcal{V}:\hat{\mathcal{V}}(y,\omega)=\mathcal{V}$, $z_{\mathcal{V}}\in\{\pm 1\}^{|\mathcal{V}|}$ are their signs, which is the set of indexes of observations with large residuals. To conduct valid post selection inference for (4.2) conditional on both the variable selection event and the outlier selection

event, $\mathcal{S} = \{E, s_E, \mathcal{V}, z_{\mathcal{V}}\}$, as discussed in (2.5.3), we employ the pull-back measure technique developed by Harris et al. (2016) to simplify the representation of a complex set of constraints $\mathcal{S}$. This involves introducing optimization variables, which are natural random variables in the problem, to describe the selection region in conjunction with the data. By doing so, we can re-parameterize the selection region in terms of these optimization variables and sample from a simpler region that only imposes constraints on the optimization variables. This approach allows us to streamline the sampling process by avoiding the need to sample from the more complex original region $\mathcal{S}$. First, we define the randomization reconstruction affine map (2.11) for (4.2).

Lemma 4.1.1 *Given the randomized objective (4.2), a linear map between randomization and the augmented vector (D, O), called randomization reconstruction is given by*

$$\omega = \omega(D, O) = A_0 D + BO + \gamma \tag{4.3}$$

where, the optimization variables $O = \begin{pmatrix} \hat{\beta}_E \\ \mu_{-E} \end{pmatrix}$, *the observed data* $D = \begin{pmatrix} \bar{\beta}_E \\ X^T_{-\mathcal{V},E}(y_{-\mathcal{V}} - X_{-\mathcal{V},E}\bar{\beta}_E) \end{pmatrix}$, *and we take* $\bar{\beta}_E = \left(X^T_{-\mathcal{V},E} X_{-\mathcal{V},E}\right)^{-1} X^T_{-\mathcal{V},E} y_{-\mathcal{V}}$ *to be the MLE for the unpenalized regression with only variables in E and observations in $\mathcal{V}^c$. A_0, B are fixed matrices and γ is a fixed vector.*

PROOF. To do this, apply the KKT conditions on (4.2) and using equation (3.3) from section (3.2.1), we have

$$-X^T_{-\mathcal{V}} y_{-\mathcal{V}} + X^T_{-\mathcal{V}} X_{-\mathcal{V}} \hat{\beta} - M X^T_{\mathcal{V}} z_{\mathcal{V}} + \lambda s + \epsilon \hat{\beta} - \omega = 0$$

By partitioning according to E and $-E$ we have

$$-X^T_{-\mathcal{V},\,E}(y_{-\mathcal{V}} - X_{-\mathcal{V},\,E}\hat{\beta}_E) - MX^T_{\mathcal{V},\,E}z_{\mathcal{V}} + \lambda s_E + \epsilon\hat{\beta}_E - \omega_E = 0$$

$$-X^T_{-\mathcal{V},\,-E}(y_{-\mathcal{V}} - X_{-\mathcal{V},\,-E}\hat{\beta}_{-E}) - MX^T_{\mathcal{V},\,-E}z_{\mathcal{V}} + \lambda\mu_{-E} - \omega_{-E} = 0$$

μ_{-E} is the subgradient vector for the penalty corresponding to inactive variables.

Simplifying and concatenating the two equations above gives:

$$\omega = -X^T_{-\mathcal{V}}y_{-\mathcal{V}} + \begin{pmatrix} X^T_{-\mathcal{V},\,E}X_{-\mathcal{V},\,E} + \epsilon I_{|E|} \\ X^T_{-\mathcal{V},\,-E}X_{-\mathcal{V},\,E} \end{pmatrix}\hat{\beta}_E + \lambda\begin{pmatrix} s_E \\ \mu_{-E} \end{pmatrix} - MX^T_{\mathcal{V}}z_{\mathcal{V}}$$

Rearranging to have

$$\omega = -\begin{pmatrix} X^T_{-\mathcal{V},\,E}X_{-\mathcal{V},\,E} & 0 \\ X^T_{-\mathcal{V},\,-E}X_{-\mathcal{V},\,E} & I_{p-|E|} \end{pmatrix}\begin{pmatrix} \bar{\beta}_E \\ X^T_{-\mathcal{V},\,E}(y_{-\mathcal{V}} - X_{-\mathcal{V},\,E}\bar{\beta}_E) \end{pmatrix} + \begin{pmatrix} X^T_{-\mathcal{V},\,E}X_{-\mathcal{V},\,E} + \epsilon I_{|E|} & 0 \\ X^T_{-\mathcal{V},\,-E}X_{-\mathcal{V},\,E} & \lambda I_{p-|E|} \end{pmatrix}\begin{pmatrix} \hat{\beta}_E \\ \mu_{-E} \end{pmatrix} + \lambda\begin{pmatrix} s_E - \frac{M}{\lambda}X_{\mathcal{V},\,E}z_{\mathcal{V}} \\ -\frac{M}{\lambda}X_{\mathcal{V},\,-E}z_{\mathcal{V}} \end{pmatrix}$$

with $\text{sign}(\hat{\beta}_E) = s_E$, $||\mu_{-E}||_\infty \leq \lambda$, $\text{sign}(y_{\mathcal{V}} - X_{\mathcal{V}}\hat{\beta}) = z_{\mathcal{V}}$, and $||y_{-\mathcal{V}} - X_{-\mathcal{V}}\hat{\beta}||_\infty \leq M$. The optimization variables are chosen such that we can recover ω by the sub-gradient equation

of our objective function (4.2).

Therefore, we have

$$\omega = \omega(D, O) = A_0 D + BO + \gamma$$

where $A_0 = \begin{pmatrix} X^T_{-\mathcal{V},\,E} X_{-\mathcal{V},\,E} & 0 \\ X^T_{-\mathcal{V},\,-E} X_{-\mathcal{V},\,E} & I_{p-|E|} \end{pmatrix}$, $B = \begin{pmatrix} X^T_{-\mathcal{V},\,E} X_{-\mathcal{V},\,E} + \epsilon I_{|E|} & 0 \\ X^T_{-\mathcal{V},\,-E} X_{-\mathcal{V},\,E} & \lambda I_{p-|E|} \end{pmatrix}$, and $\gamma = \begin{pmatrix} s_E - \frac{M}{\lambda} X_{\mathcal{V},\,E} z_{\mathcal{V}} \\ -\frac{M}{\lambda} X_{\mathcal{V},\,-E} z_{\mathcal{V}} \end{pmatrix}$.

From lemma (4.1.1) and a change of measure from (D, ω) to (D, O) as introduced by Harris et al. (2016), the selection event $\mathcal{S} = (E, s_E, \mathcal{V}, z_{\mathcal{V}})$ from the solver in (4.2) is reparametrized and now $\mathcal{S}$ is described by the map $\omega(D, O)$ where optimization variables O are constrained to the region

$$\mathcal{K} = \{o \in \mathbb{R}^p : \text{sign}(o_E) = s_E, ||o_{-E}||_\infty \leq \lambda, \text{ and } \text{sign}(y_{\mathcal{V}} - X_{\mathcal{V}}\hat{\beta}) = z_{\mathcal{V}}, ||y_{-\mathcal{V}} - X_{-\mathcal{V}}\hat{\beta}||_\infty \leq M\} \tag{4.4}$$

Therefore, selective inference will now be based on the joint law of data and the optmization variables (D, O), conditional on the event that constrains the optimization variables O to lie in $\mathcal{K}$. The truncated joint law of (D, O) at (d, o) becomes

$$f_D(d) \times g(w(d, o)) \times \mathbb{I}_{o \in \mathcal{K}} \tag{4.5}$$

where $f_D(d)$ is the pre-selection density of D and D is asymptotically normal, $g(.)$ is the density of the randomization ω. The method of changing variables explained above addresses

the sampling difficulties that arise when dealing with the selective density of data conditional on a randomized selection region. Additionally, there is another computational hurdle of sampling the relevant section of the data vector that corresponds to the selected parameter of interest, while conditioning on the section of the data vector that pertains to the nuisance parameters. We circumvent this hurdle through linear decomposition as follows, in testing the hypothesis $H_0 : \beta_E = \theta$, we utilize $||T - \theta||_2^2$ as the test statistic, where $T = \bar{\beta}_E$ is referred to as the target statistic. In order for this test to be valid while treating E as non-random (pre-selection), we employ the asymptotic normality of T to establish a reference distribution. However, for post-selection, we must base our inference on the post-selection distribution of T. Notably, we can perform inference for any parameter θ if the pre-selection Central Limit Theorem (CLT) holds for the target statistic T and the data vector D:

$$\begin{pmatrix} T \\ D \end{pmatrix} \to \mathcal{N}\left(\begin{pmatrix} \theta \\ \mu_D \end{pmatrix}, \begin{pmatrix} \Sigma_T & \Sigma_{T,D} \\ \Sigma_{D,T} & \Sigma_D \end{pmatrix} \right) \quad \text{as } n \to \infty$$

To obtain the post-selection distribution of the target statistic T under the null hypothesis, we perform a decomposition of the affine map (4.3). Let $F = D - \hat{\Sigma}_{D,T}\hat{\Sigma}_T^{-1}T$, where $\hat{\Sigma}_{D,T}$ and $\hat{\Sigma}_T$ are the corresponding covariance estimates. By decomposing D into $F + T$ and conditioning on F, it is sufficient to sample $(T, \hat{\beta}_E, \mu_{-E})$. The plugin sampling density of $(T, \hat{\beta}_E, \mu_{-E})$ is proportional to this distribution

$$\phi_{(\theta, \hat{\Sigma}_T)}(T) \times g(\mathcal{A}T + BO + \mathcal{C}) \times \mathbb{I}_{O \in \mathcal{K}} \tag{4.6}$$

where $\mathcal{A} = A_0\hat{\Sigma}_{D,T}\hat{\Sigma}_T^{-1}$, $\mathcal{C} = \gamma + A_0F$ and $\phi_{(\cdot,\cdot)}$ represents the density of the multivariate normal distribution with the mean and covariance matrix specified in the subscript. The estimation of $\Sigma_{D,T}$ and Σ_T can be achieved using pairs bootstrap.

Even though we can effectively compute the selective pivot to construct p-values and confidence intervals by sampling from the density in (4.6) to obtain samples of $(T, \hat{\beta}_E, \mu_{-E})$, performing multiple tests simultaneously requires running a separate sampler for each test. For instance, to provide selective confidence intervals for all the chosen coefficients $\beta_{E\cdot j}, j \in E$, we must run $|E|$ samplers and set the target T for each one to be $\bar{\beta}_{E\cdot j}, j \in E$. To enhance efficiency, we utilize the weighted optimization sampler that is described below. By sampling the optimization variables from the selective density that fixes the data at its observed value, we can reuse the same optimization samples across different tests. Consequently, we can run the sampler only once while still delivering inference for multiple tests at the same time. Overall, the use of the weighted optimization sampler and selective density in this manner enables us to reduce computational inefficiency and enhance the efficiency of multiple testing procedures.

Steps for constructing the selective pivot.

1. To sample the optimization variables $(\hat{\beta}_E, \mu_{-E})$ given the observed data vector $D = D^{obs}$, we use a density proportional to

$$g(\omega(D^{obs}, \hat{\beta}_E, \mu_{-E}))$$

with $\text{sign}(\hat{\beta}_E) = s_E$, $||\mu_{-E}||\infty \leq \lambda$, $\text{sign}(y_{\mathcal{V}} - X_{\mathcal{V}}\hat{\beta}) = z_{\mathcal{V}}$, and $||y_{-\mathcal{V}} - X_{-\mathcal{V}}\hat{\beta}||_\infty \leq M$. We denote the resulting samples as $(\hat{\beta}^s_E, \mu^s_{-E})$, where $s = 1, 2, \ldots, S$, and S represents the sample size.

2. To obtain samples for the target, we sample from its pre-selection normal distribution, resulting in samples $T^s \sim \mathcal{N}(0, \hat{\Sigma}_T)$, where $s = 1, 2, \ldots, S$.

3. To compute the selective pivot, we first combine the samples $(T^s + \theta, \hat{\beta}^s_E, \mu^s_{-E})$, where $s = 1, 2, \ldots, S$, obtained from the first and second steps. We then weight and tilt each of these triples $(T^s + \theta, \hat{\beta}^s_E, \mu^s_{-E})$ using importance sampling with the ratio

$$w(T^s, \hat{\beta}^s_E, \mu^s_{-E}) = \frac{g(\mathcal{A}(T^s + \theta) + BO^s + \mathcal{C})}{g(A_0 D^{obs} + BO^s + \mathcal{C})}$$

where $O^s = \begin{pmatrix} \hat{\beta}^s_E \\ \mu^s_{-E} \end{pmatrix}$. We then compute the selective pivot as the weighted sum of the indicator function of the condition $||T^s||_2 \leq ||T^{obs} - \theta||_2$, normalized by the sum of the importance weights:

$$\sum_{s=1}^{S} \mathcal{I}_{||T^s||_2 \leq ||T^{obs} - \theta||2} \cdot \frac{w(T^s, \hat{\beta}^s_E, \mu^s_{-E})}{\sum_{s'=1}^{S'} w(T^s, \hat{\beta}^s_E, \mu^s_{-E})}$$

If the target T is one-dimensional and we want to compute a confidence interval, we need to repeat the third step for different values of θ to invert the pivot. In the case of multiple tests, we need to repeat the second and third steps above.

Having reviewed the discussions above, we can now present our main findings in the form of a theorem:

Theorem 4.1.2 (Main Result3) *If E and s_E are the model and the corresponding signs selected from the randomized procedure (4.2) with $E \supseteq supp(\beta)$, and $\mathcal{V}$ is the outlier selection output with corresponding signs $z_{\mathcal{V}}$, where $\omega \sim \mathcal{N}(0, \tau^2 I_p)$ independent of (X, y). While conditioning on $(E, s_E, \mathcal{V}, z_{\mathcal{V}})$, samples $(T, \hat{\beta}_E, \mu_{-E})$ from the joint distribution of (D, O) as in (4.5) are used for inference for β_E by approximating the selective pivot*

$$\mathcal{P} = \sum_{s=1}^{S} \mathcal{I}_{||T^s||_2 \leq ||T^{obs} - \theta||_2} \cdot \frac{w(T^s, \hat{\beta}_E^s, \mu_{-E}^s)}{\sum_{s'=1}^{S'} w(T^s, \hat{\beta}_E^s, \mu_{-E}^s)}$$

A two sided p-value can be computed as $P_{val} = 2 \cdot min(\mathcal{P}, 1 - \mathcal{P})$. The $100(1-\alpha)\%$ two sided confidence intervals can be computed by inverting the approximate selective pivot $\mathcal{P}$.

4.2 Simulation

In this experiment, we have adopted the same structure as in (3.4) for our design. In addition to that, we have introduced two new parameters, namely ω and ϵ. The former is sampled from a normal distribution, i.e., $\omega \sim \mathcal{N}(0, 2.2 * I_p)$. The latter is set to a fixed value of 3.1. To evaluate the effectiveness of our methodology, we have compared two versions of it. The first one is the randomized version, which was discussed earlier in this Chapter. The second version, which we referred to in Chapter 3, does not involve randomization. The purpose of this comparison is to demonstrate the improved inferential power that randomization provides in terms of average interval length. The average interval length is accompanied by its standard deviation, highlighted in orange.

Table 4.1: (n, p) = (400, 10), $\omega \sim \mathcal{N}(0, 1.1 * I_p)$, $\epsilon = 3.1$

Huber-Cond			**Rand-Huber-Cond**		**Rand-lasso-Cond**	
Errors	Coverage Prob	Ave Length	Coverage Prob	Ave Length	Coverage Prob	Ave Length
$t_{1.5}$	95.8%	5.89 1.15	95%	1.0 0.55	94.2%	2.70 0.94
Mix-Normal	95.1%	4.78 1.11	94.92%	1.05 0.58	94.7%	1.22 0.66
Normal	95.01%	2.34 0.76	94.89%	0.92 0.51	94.97%	0.72 0.40

Table 4.2: (n, p) = (1000, 10), $\omega \sim \mathcal{N}(0, 1.1 * I_p)$, $\epsilon = 3.1$

Huber-Cond			**Rand-Huber-Cond**		**Rand-lasso-Cond**	
Errors	Coverage Prob	Ave Length	Coverage Prob	Ave Length	Coverage Prob	Ave Length
$t_{1.5}$	95.5%	3.72 0.92	95.02%	0.91 0.49	94.4%	2.01 0.89
Mix-Normal	95.02%	2.97 1.08	94.93%	0.94 0.51	94.82%	0.98 0.53
Normal	94.99%	1.26 0.73	94.92%	0.90 0.53	95.01%	0.71 0.40

Upon examining the tables above, it is evident that all methods exhibit good coverage probability. It is worth mentioning that when considering the conditional confidence intervals for Rand-Huber-Cond, which were discussed and developed earlier in Chapter 4, we have conditioned on both the variable selection event and the outlier identification event. However, a notable reduction in the average interval lengths is observed when a small amount of randomization is introduced, particularly when comparing Huber-Cond from Chapter 3 and Rand-Huber-Cond. This outcome underscores the substantial inferential power associated with randomized procedures. In comparison to the randomized lasso (Rand-lasso-Cond) method introduced by Tian (2018), our method (Rand-Huber-Cond) demonstrates superior performance in terms of shorter interval lengths when the errors are heavy-tailed as seen in

Table 4.3: (n, p) = (2000, 10), $\omega \sim \mathcal{N}(0, 1.1 * I_p)$, $\epsilon = 3.1$

	Huber-Cond			Rand-Huber-Cond			Rand-lasso-Cond		
Errors	Coverage Prob	Ave Length		Coverage Prob	Ave Length		Coverage Prob	Ave Length	
$t_{1.5}$	95.3%	2.13	0.78	95%	0.88	0.48	94.8%	1.79	0.90
Mix-Normal	95.01%	1.94	0.97	94.98%	0.85	0.45	94.9%	0.93	0.47
Normal	95%	1.02	0.57	94.97%	0.79	0.41	95.1%	0.68	0.34

Table 4.4: (n, p) = (100, 200), $\omega \sim \mathcal{N}(0, 1.1 * I_p)$, $\epsilon = 3.1$

	Huber-Cond			Rand-Huber-Cond			Rand-lasso-Cond		
Errors	Coverage Prob	Ave Length		Coverage Prob	Ave Length		Coverage Prob	Ave Length	
$t_{1.5}$	92.4%	5.97	1.93	94.94%	2.29	0.89	94.51%	3.87	1.18
Mix-Normal	93.9%	5.04	1.54	94.96%	1.97	0.61	94.87%	2.72	0.92
Normal	94.37%	2.48	0.87	94.97%	1.17	0.54	94.98%	0.98	0.51

Tables 4.1 - 4.3. In Table 4.5, a very small randomization scale was taken into consideration which resulted in a decrease in power when measured by the average length of intervals, in comparison to the results obtained from Tables 4.1-4.3. This finding is unsurprising as the addition of such a minute amount of noise has limited potential to significantly enhance inferential power. Furthermore, we conducted experiments by varying the randomization scale but the results remained largely consistent. Furthermore, as demonstrated in Table 4.5, our method exhibited superiority when applied to large values of p.

Table 4.5: (n, p) = (400, 10), $\omega \sim \mathcal{N}(0, 0.1 * I_p)$, $\epsilon = 3.1$

Huber-Cond			**Rand-Huber-Cond**		**Rand-lasso-Cond**	
Errors	Coverage Prob	Ave Length	Coverage Prob	Ave Length	Coverage Prob	Ave Length
$t_{1.5}$	95.82%	5.15 1.10	95.1%	2.07 0.76	94.62%	2.94 0.91
Mix-Normal	95.21%	3.82 1.06	94.97%	1.46 0.63	94.83%	2.19 0.87
Normal	95.03%	2.23 0.73	94.94%	1.07 0.57	94.8%	0.87 0.49

4.3 Real data analysis (Acute Lymphocytic Leukemia)

Acute lymphocytic leukemia (ALL) or acute lymphoblastic leukemia is a type of cancer affecting white blood cells known as lymphocytes. ALL is characterized by the rapid production of immature lymphocytes in the bone marrow, which can cause anemia, infection, and bleeding. Although it is the most common type of cancer in children, ALL can also affect adults. The exact causes of ALL are not fully understood, but it is believed to result from genetic mutations in developing lymphocytes. Symptoms of ALL include weakness, fever, infections, bleeding, and bone pain. Diagnosis usually involves a combination of blood tests, bone marrow biopsy, and imaging studies. Treatment typically involves chemotherapy, radiation therapy, and bone marrow transplant, depending on the patient's age, overall health, subtype, and stage of the disease. Modern treatment approaches have significantly improved the prognosis for ALL, with survival rates of up to 90% in children and 40-50% in adults. However, prognosis depends on various factors, such as age, extent of the disease, and specific genetic mutations involved. Analyzing gene expression data, also known as omics data, can provide insights into the biological mechanisms underlying ALL. For example, gene expression profiling can identify molecular subtypes of ALL, which can inform personalized

treatment strategies. In addition, gene expression data can be used to identify specific genes or pathways that are dysregulated in ALL, which can be targeted with new therapies. For example, BCL2L1 has been identified as a potential therapeutic target in ALL based on its role in promoting disease progression and drug resistance. BCL2L1 is a gene that plays a critical role in promoting cell survival by inhibiting apoptosis. It is commonly overexpressed in many types of cancer, including ALL, and has been implicated in disease progression and resistance to chemotherapy. High expression of BCL2L1 is associated with poor prognosis in ALL patients, indicating its role in disease severity and progression. BCL2L1 promotes the survival of leukemia cells by preventing apoptosis in response to chemotherapy, leading to drug resistance and reduced treatment efficacy. Additionally, BCL2L1 contributes to the maintenance of leukemia stem cells responsible for disease relapse and progression. Targeting BCL2L1 could sensitize leukemia cells to chemotherapy and prevent disease relapse.

In this study, we aim to identify genes that play a significant role in dysregulation (overexpression or underexpression) of the BCL2L1 gene in acute lymphocytic leukemia patients. We utilize gene expression levels from the leukemia dataset, originally introduced by Golub et al. (1999), which can be accessed at `https://hastie.su.domains/CASI_files/DATA/leukemia.html`. The dataset includes 47 patients with acute lymphocytic leukemia, and genetic activity was measured for a panel of 3,571 genes. Based on a thorough review of the relevant literature, we selected a subset of 306 genes that are known to be involved in the pathogenesis of acute lymphocytic leukemia. We apply our developed methodology, the randomized Huber-lasso, to identify potential genes that may explain variation in the expression of BCL2L1. To measure the strength of the association between each potential

gene and BCL2L1, we computed conditional confidence intervals. Our results show that several genes were identified as being significantly associated with dysregulation of BCL2L1 expression in acute lymphocytic leukemia patients, including genes involved in the regulation of apoptosis, cell proliferation, and cell differentiation. Our approach is compared to the vanilla randomized lasso, and the results of our analysis are presented below.

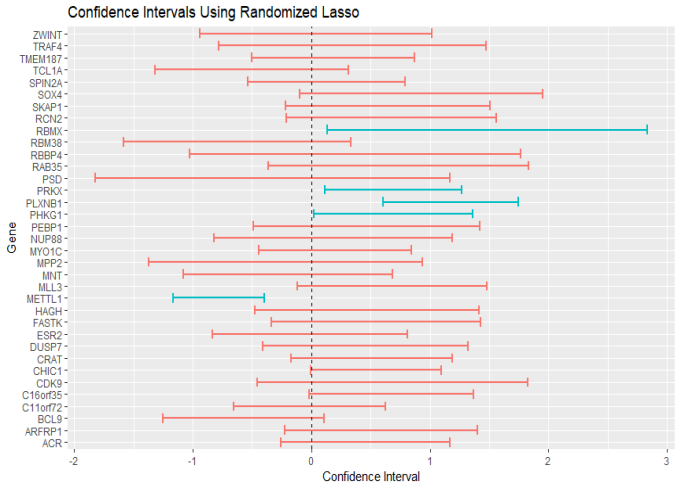


Figure 4.1: Selective intervals with randomized lasso

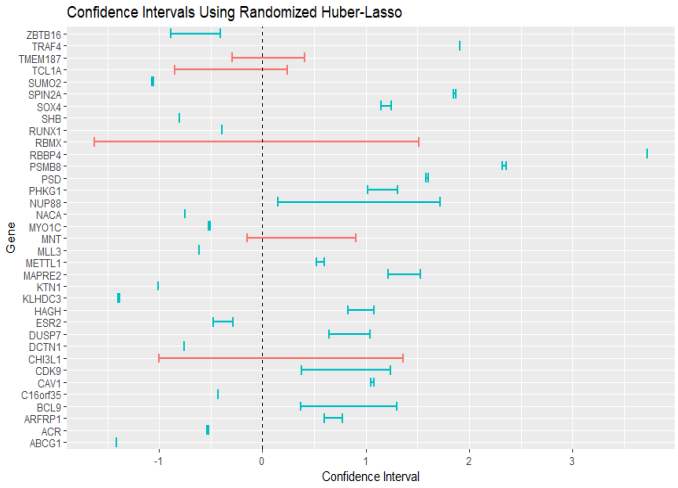


Figure 4.2: Selective intervals with randomized huber-lasso

To include outliers in the response variable, we identified the smallest three observations of BCL2L1 and decreased their values by 10, and similarly, we identified the largest three observations of BCL2L1 and increased their values by 10. Using the randomized lasso technique by Fithian et al. (2017), we selected 35 genes and found 5 to be statistically significant, as shown in Table 4.1. However, this method failed to select the RUNX1 and SHB genes, which are known to play a role in regulating BCL2L1 expression in leukemic cells, as discussed in Mercher et al. (2001) and Thiriet and Thiriet (2013), respectively. Additionally, the randomized lasso method failed to establish the significance of SOX4, CDK9, and SPIN21, despite

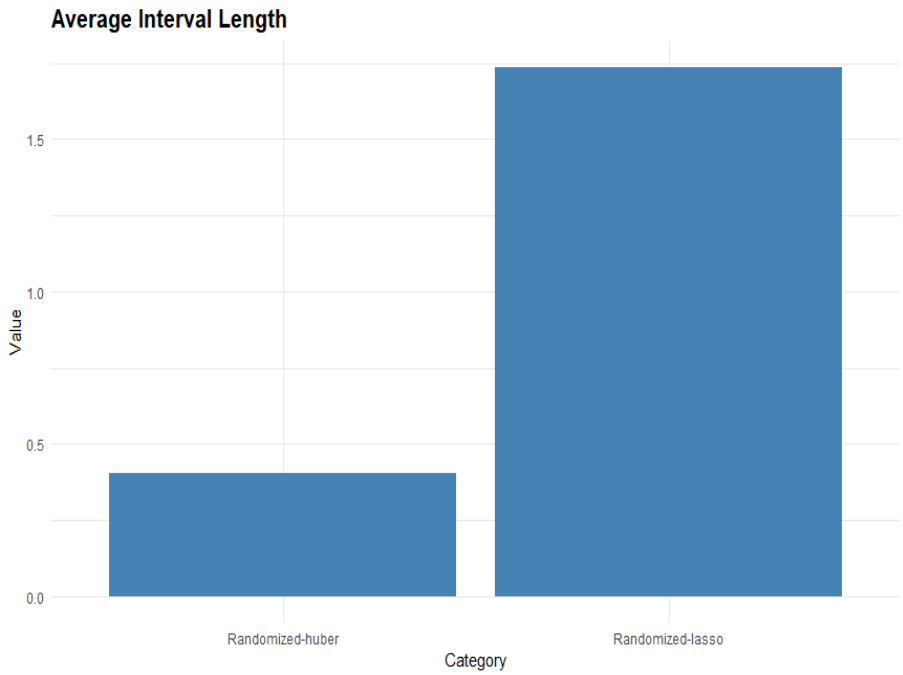


Figure 4.3: Average Length of the Selective Intervals

previous evidence linking them to BCL2L1 dysregulation in acute lymphocytic leukemia. For instance, CDK9 has been shown to promote the survival of leukemic cells by upregulating BCL2L1 expression (see Huang et al. (2014)), SPIN2A has been implicated in regulating BCL2L1 splicing (see Shaw et al. (2021)), and SOX4 can directly bind to the BCL2L1 promoter and activate its expression in leukemic cells, contributing to BCL2L1 dysregulation (see Puissant et al. (2014)). In contrast, our developed method, the randomized Huber loss, selected RUNX1 and SHB and declared SOX4, CDK9, and SPIN21 statistically significant, as shown in Table 4.2. Our method found 30 of the 35 selected genes to be significant, and the conditional selective confidence interval is shorter and more precise, as also shown in Table 4.2. Furthermore, our method has a significantly shorter average interval length, as demonstrated in figure 4.3. These results highlight the trustworthiness and superiority of our method in the presence of outliers or heavy-tailed errors.

Chapter 5

Future Works

In our research, we plan to explore several other scenarios and problems in the future. First, we aim to extend our approach of conditional confidence intervals, as proposed in Chapter 3, to other robust penalized regression models in high dimensions. We will conduct simulations to determine the effectiveness of the approach on other models and make any necessary revisions. One of the models we plan to investigate is the penalized quantile regression, which is widely used in survival analysis, especially when the survival data is skewed or censored, and traditional methods like Cox regression may not provide accurate results. Additionally, it is used in pharmaceutical research to model the relationship between drug dose and response and estimate the median effective dose (ED50) or other quantiles of the dose-response curve. This is important for optimizing drug dosing and minimizing adverse effects, especially when the dose-response relationship is non-linear or heterogeneous. Next, we aim expand the methodology presented in Chapter 4 to develop a Monte Carlo-free approach for post-selection inference following randomization, since as explained in section 2.5.3, our focus is on making inferences about the parameter b, and we can obtain the marginal density of data conditional on the selection event, $T|(T, \omega)$, by marginalizing over the randomization part. This can be achieved by integrating over the ω's, resulting in the expression $exp(-(t-b)^2/2\sigma^2) \times P((T, \omega) \in S_{(E,s_E)}, |, T = t)$. Here, $P((T, \omega) \in S_{(E,s_E)}, |, T = t)$ is the probability of randomization landing in the selection region $S_{(E,s_E)}$ conditional on the

data T. Instead of using a sampler to estimate this intractable probability, we aim to obtain a good approximation of $P((T,\omega) \in S_{(E,s_E)}, |, T = t)$. Also in Chapter 4, we observed that making a trade-off between model-selection power and subsequent inference power can lead to a considerable improvement in the latter. In our ongoing research, we aim to further enhance the reliability of our statistical inferences by theoretically analyzing the expected length of the randomized procedure for the penalized Huber regression, as discussed in Chapter 4. By determining the expected length, we can gain a more comprehensive understanding of the precision of our estimates and the potential impact of sample size on statistical power.

Bibliography

Andrews, D. F. and F. R. Hampel (2015). Robust estimates of location. In *Robust Estimates of Location*. Princeton University Press.

Anscombe, F. J. (1967). Topics in the investigation of linear relations fitted by the method of least squares. *Journal of the Royal Statistical Society: Series B (Methodological) 29*(1), 1–29.

Bai, X. (2014). *Robust mixtures of regression models*. Kansas State University.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *The Annals of Statistics*, 802–837.

Bickel, P. J. (1975). One-step huber estimates in the linear model. *Journal of the American Statistical Association 70*(350), 428–434.

Bickel, P. J. (1984). Robust regression based on infinitesimal neighbourhoods. *The Annals of Statistics*, 1349–1368.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics 37*(4), 1705–1732.

Boos, D. D. (1980). A new method for constructing approximate confidence intervals from m estimates. *Journal of the American Statistical Association 75*(369), 142–145.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli 19*(4), 1212–1242.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Bunea, F., A. Tsybakov, and M. Wegkamp (2007a). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics 1*, 169–194.

Bunea, F., A. B. Tsybakov, and M. H. Wegkamp (2007b). Aggregation for gaussian regression. *The Annals of Statistics 35*(4), 1674–1697.

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

Chen, Y., S. Jewell, and D. Witten (2022). More powerful selective inference for the graph fused lasso. *Journal of Computational and Graphical Statistics*, 1–11.

Dezeure, R., P. Bühlmann, L. Meier, and N. Meinshausen (2015). High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical science*, 533–558.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of statistics 32*(2), 407–499.

El Karoui, N., D. Bean, P. J. Bickel, C. Lim, and B. Yu (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences 110*(36), 14557–14562.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Fithian, W., D. Sun, and J. Taylor (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Fithian, W., D. Sun, and J. Taylor (2017). Optimal inference after model selection. arxiv.

Fu, W. and K. Knight (2000). Asymptotics for lasso-type estimators. *The Annals of statistics 28*(5), 1356–1378.

Gao, L. L., J. Bien, and D. Witten (2022). Selective inference for hierarchical clustering. *Journal of the American Statistical Association* (just-accepted), 1–27.

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

Greenshtein, E. and Y. Ritov (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli 10*(6), 971–988.

Harris, X. T., S. Panigrahi, J. Markovic, N. Bi, and J. Taylor (2016). Selective sampling after solving a convex problem. *arXiv preprint arXiv:1609.05609*.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). Statistical learning with sparsity. *Monographs on statistics and applied probability 143*, 143.

Huang, C.-H., A. Lujambio, J. Zuber, D. F. Tschaharganeh, M. G. Doran, M. J. Evans, T. Kitzing, N. Zhu, E. de Stanchina, C. L. Sawyers, et al. (2014). Cdk9-mediated transcription elongation is required for myc addiction in hepatocellular carcinoma. *Genes & development 28*(16), 1800–1814.

Huber, P. (1981). Robust statistics. new york: John wiley and sons. *HuberRobust statistics1981*.

Huber, P. J. (1964). Robust estimation of a location parameter: Annals mathematics statistics, 35.

Huber, P. J. (1972). The 1972 wald lecture robust statistics: A review. *The Annals of Mathematical Statistics 43*(4), 1041–1067.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The annals of statistics*, 799–821.

Hyun, S., K. Z. Lin, M. G'Sell, and R. J. Tibshirani (2021). Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics 77*(3), 1037–1049.

Javanmard, A. and A. Montanari (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research 15*(1), 2869–2909.

Javanmard, A. and A. Montanari (2014b). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory 60*(10), 6522–6554.

Jewell, S., P. Fearnhead, and D. Witten (2019). Testing for a change in mean after change-point detection. *arXiv preprint arXiv:1910.04291*.

Kivaranovic, D. and H. Leeb (2020). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *Journal of the American Statistical Association*, 1–13.

Koenker, R. and G. Bassett Jr (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Lambert-Lacroix, S. and L. Zwald (2011). Robust regression through the huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics 5*, 1015–1053.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics 44*(3), 907–927.

Lee, J. D. and J. E. Taylor (2014). Exact post model selection inference for marginal screening. *arXiv preprint arXiv:1402.5596*.

Liu, Y., P. Zeng, and L. Lin (2021). Degrees of freedom for regularized regression with huber loss and linear constraints. *Statistical Papers 62*(5), 2383–2405.

Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *The Annals of Statistics*, 382–400.

Markovic, J. and J. Taylor (2016). Bootstrap inference after using multiple queries for model selection. *arXiv preprint arXiv:1612.07811*.

Mehrizi, R. V. and S. Chenouri (2021). Valid post-detection inference for change points identified using trend filtering. *arXiv preprint arXiv:2104.12022*.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics 34*(3), 1436–1462.

Meinshausen, N., L. Meier, and P. Bühlmann (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association 104*(488), 1671–1681.

Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The annals of statistics 37*(1), 246–270.

Mercher, T., M. B.-L. Coniat, R. Monni, M. Mauchauffé, F. N. Khac, L. Gressin, F. Mugneret, T. Leblanc, N. Dastugue, R. Berger, et al. (2001). Involvement of a human gene related to the drosophila spen gene in the recurrent t (1; 22) translocation of acute megakaryocytic leukemia. *Proceedings of the National Academy of Sciences 98*(10), 5776–5779.

Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science 27*(4), 538–557.

Neufeld, A. C., L. L. Gao, and D. M. Witten (2021). Tree-values: selective inference for regression trees. *arXiv preprint arXiv:2106.07816*.

Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. *Contemporary Mathematics 443*(7), 59–72.

Pan, X., Q. Sun, and W.-X. Zhou (2021). Iteratively reweighted 1-penalized robust regression. *Electronic Journal of Statistics 15*(1), 3287–3348.

Portnoy, S. (1984). Asymptotic behavior of m-estimators of p regression parameters when p 2/n is large. i. consistency. *The Annals of Statistics*, 1298–1309.

Portnoy, S. (1985). Asymptotic behavior of m estimators of p regression parameters when p^2/n is large; ii. normal approximation. *The Annals of Statistics 13*(4), 1403–1417.

Portnoy, S. (1986). On the central limit theorem in r p when p→. *Probability theory and related fields 73*(4), 571–583.

Portnoy, S. (1987). A central limit theorem applicable to robust regression estimators. *Journal of multivariate analysis 22*(1), 24–50.

Puissant, A., N. Fenouille, G. Alexe, Y. Pikman, C. F. Bassil, S. Mehta, J. Du, J. U. Kazi, F. Luciano, L. Rönnstrand, et al. (2014). Syk is a critical regulator of flt3 in acute myeloid leukemia. *Cancer cell 25*(2), 226–242.

Relles, D. A. (1968). *Robust regression by modified least-squares*. Yale University.

Ronchetti, E. M. and P. J. Huber (2009). *Robust statistics*. John Wiley & Sons.

Rousseeuw, P. and V. Yohai (1984). Robust regression by means of s-estimators. In *Robust and nonlinear time series analysis*, pp. 256–272. Springer.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association 79*(388), 871–880.

Rousseeuw, P. J. and A. M. Leroy (2005). *Robust regression and outlier detection*. John wiley & sons.

Shaw, T. I., L. Dong, L. Tian, C. Qian, Y. Liu, B. Ju, A. High, K. Kavdia, V. R. Pagala, B. Shaner, et al. (2021). Integrative network analysis reveals usp7 haploinsufficiency

inhibits e-protein activity in pediatric t-lineage acute lymphoblastic leukemia (t-all). *Scientific Reports 11*(1), 5154.

She, Y. and A. B. Owen (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association 106*(494), 626–639.

Sun, Q., W.-X. Zhou, and J. Fan (2020). Adaptive huber regression. *Journal of the American Statistical Association 115*(529), 254–265.

Taylor, J., R. Lockhart, R. J. Tibshirani, and R. Tibshirani (2014). Post-selection adaptive inference for least angle regression and the lasso. *arXiv preprint arXiv:1401.3889 354*.

Taylor, J. and R. Tibshirani (2018). Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics 46*(1), 41–61.

Thiriet, M. and M. Thiriet (2013). Preamble to cytoplasmic protein kinases. *Intracellular Signaling Mediators in the Circulatory and Ventilatory Systems*, 109–135.

Tian, X. and J. Taylor (2017). Asymptotics of selective inference. *Scandinavian Journal of Statistics 44*(2), 480–499.

Tian, X. and J. Taylor (2018). Selective inference with a randomized response. *The Annals of Statistics 46*(2), 679–710.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of statistics 7*, 1456–1490.

Tibshirani, R. J., A. Rinaldo, R. Tibshirani, and L. Wasserman (2018). Uniform asymptotic inference and the bootstrap after model selection. *The Annals of Statistics 46*(3), 1255–1287.

Tibshirani, R. J., J. Taylor, R. Lockhart, and R. Tibshirani (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association 111*(514), 600–620.

Van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

Van De Geer, S. A. and P. Bühlmann (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics 3*, 1360–1392.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using -constrained quadratic programming (lasso). *IEEE transactions on information theory 55*(5), 2183–2202.

Yohai, V. J. (1974). Robust estimation in the linear model. *The Annals of Statistics*, 562–567.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of statistics*, 642–656.

Yohai, V. J. and R. A. Maronna (1979). Asymptotic behavior of m-estimators for the linear model. *The Annals of Statistics*, 258–268.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics 38*(2), 894–942.

Zhang, C.-H. and J. Huang (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics 36*(4), 1567–1594.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242.

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

Zhao, Q., D. S. Small, A. Ertefaie, et al. (2022). Selective inference for effect modification via the lasso. *Journal of the Royal Statistical Society Series B 84*(2), 382–413.

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

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology) 67*(2), 301–320.

Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics 36*(4), 1509.

Zou, H. and H. H. Zhang (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics 37*(4), 1733.