Robust Significance Testing in Sparse and High Dimensional Linear Models

Wenjing Yin

Advisor: Professor Jelena Bradic

May 2015
Department of Mathematics
University of California, San Diego

Abstract

Classical statistical theory offers validity under restricted assumptions. However, in practice, it is a common approach to perform statistical analysis based on data-driven model selection [1], which guarantees none of results of classical statistical theory. Those results include hypothesis testings and confidence intervals which are useful tools of measuring fitness of models. Considering that too much information about the true model of the datasets is unknown,we are unable to perform any testing before model selection. However, we are still interested in identifying how well the model we select fits the data, which leads to the problems of testing after model selection.

In this paper, we discuss the robustness in testing after model selection of the lasso. Lasso, as a relatively new estimation procedure, have not been thoroughly explored yet. Especially when working in practice, one may intend to assure that the lasso model he or she chooses is the appropriate one within the assigned significance level.In the last decades, a few papers have been working on the testing problems of the lasso. Among those papers, we choose [2] as a reference paper and prove some of the lemmas of [2] with details. The lemmas help us to understand the properties of the test statistic derived in the paper, named covariance test statistic, and its asymptotic distribution under the null hypothesis. We also determine the exact stoping time for selecting the variables during the second step of the LARS algorithm. The exact stopping time allows us to propose a new LARS algorithm that is robust to the presence of outliers, by using Kendall's τ correlation coefficient. We mimic the successive feature of the famous LARS algorithm and use the exact stopping time to select the second explanatory variables.

Additionally, we propose a new test statistic that tests whether the selected variables are contained in the support of true model. The test statistics compares the covariance between the model selected before the stopping time and the model that includes an additional feature, in terms of Kendall's τ correlation coefficient. Furthermore, we find a connection of our new test statistic with the Wilcoxon ranked-sum test and use that connection to study its distribution properties. The analysis is complicated by the intricate dependencies present in the proposed test statistic. We conjecture that the new test statistic has asymptotically rescaled normal distribution under the null. We also design a simulation to illustrate the finite sample properties of our new test statistic. We observe that the finite sample behavior shows better stability in comparison to the existing covariance test statistic.

Contents

1 Introduction

We consider the usual setup of linear regression problem for an observed vector $\mathbf{y}$ and a matrix of explanatory variables $\mathbf{X}$.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta^*} + \boldsymbol{\varepsilon}, \quad \mathbf{y} \in \mathbb{R}^n,\ \mathbf{X} \in \mathbb{R}^{n\times p}, \text{and } \boldsymbol{\beta}^* \in \mathbb{R}^p, \tag{1}$$

where $\boldsymbol{\varepsilon}$ is a vector of n entries of standard normal distribution. $\boldsymbol{\beta^*}$ is the vector of unknown coefficients which we intend to estimate through model selection. Usually in linear regression problems, we deal with the case that $n > p$. However, the case that $n < p$ is possible in practical problems. We mainly focused on the usual case in this paper and will discuss how the inequality between n and p changes our result in the last section. The estimators of $\boldsymbol{\beta^*}$ have distinct properties according to the distinct models we choose.

Considering that modern statistics deals with large and complex datasets, we are more interested in high-dimensional statistical analysis and problems with sparse models. Sparse models assume that the number of nonzero coefficients in the coefficient vector, $\boldsymbol{\beta}^*$, is much smaller than the sample size n. That is, if we denote the corresponding explanatory variables to the nonzero coefficients as $\text{supp}(\boldsymbol{\beta}^*)$, we have $n \gg \text{supp}(\beta^*)$ as the true model. We wonder how well model selection can perform as a tool to solve such problems within high-dimensional statistical setting and how well different methods can be applied.

Notice that the goal here is to estimate the unknown sparse vector $\boldsymbol{\beta^*}$ using observed dataset $\{\mathbf{y}, \mathbf{X}\}$. An usual approach is Ordinary Least Square(OLS) which minimizes the sum of squares of errors $\boldsymbol{\varepsilon}$. That is,

$$\hat{\boldsymbol{\beta}}_{\text{ols}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \varepsilon_i^2 \quad = \arg\min \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j^* x_{ij})^2. \tag{2}$$

While we intend to obtain sparsity in the estimated vector of coefficients, a lot of statistical methods are chosen such as Ordinary Least Squares(OLS), lasso, and Least Angle Regression (LARS). The first method, OLS, succeeds in low-dimensional setting with $p \leq n$ by providing a consistent and unbiased estimators. However, it fails to offer a sparse vector $\hat{\boldsymbol{\beta}}$ in high dimensional setting [3]. The later two methods, lasso [3] and LARS [4], shrink the value of estimators and use certain thresholds, making sparsity more accessible. The lasso estimators are defined as

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j^* x_{ij})^2 + \lambda \|\boldsymbol{\beta^*}\|. \tag{3}$$

where λ is called the tuning parameter and controls the number of sparse elements to be present in the estimator $\hat{\boldsymbol{\beta}}_{\text{lasso}}$.

However, besides the sparsity in the true coefficient vector $\boldsymbol{\beta^*}$, the unknown covariance of the high-dimensional dataset makes direct OLS an impossible approach. Therefore, we focus on the other two model selection methods, lasso [3] and LARS [4] which include penalties during the minimization process and control the sparsity in estimators of $\boldsymbol{\beta^*}$.

Though compared with OLS solutions, lasso controls better in sparsity of the solution, it only works under appropriate conditions and limitations of the explanatory variables. When proper assumptions are missing, lasso fails to offer consistent solutions [5].The appropriate conditions are Irrepresentable Condition [6] and Restricted Eigenvalue Condition [7]. The former is restricting the empirical column correlation in the design matrix $\mathbf{X}$, and is not verifiable in any given dataset because the empirical column correlation depends on the unknown sparsity of the true regression parameter $\boldsymbol{\beta}^*$. The later is restricting the variation in the sample covariance matrix. However, these two restricted conditions are highly unlikely to be fulfilled in practical problems. We wonder how well lasso estimators performed in practice and how to use lasso estimators to design tests for the purpose of studying the existence of sparsity.

Since only under restricted conditions we can obtain sparsity in the lasso estimators, it is even more necessary for us to test if the set of variables selected fits our expectation of the sparse model, which leads to the problem of testing after model selection. Testing after model selection or with model selection in mind, is a new frontier in high dimensional statistics which allows us to understand more about the methods and how well the methods are performing. [7]

Several obstacles for testing after model selection are difficult to overcome. One is how to precisely measure the difference between our ideal selection and the resulted selection. Second is how to correctly design the test statistic and provide results based on asymptotic distribution of the test statistic. Third is, even we succeed in setting the difference and in comparing the results, how we can improve the selection process of the sparse model to make our method works better. These are all open questions that worth working on.

Traditional statistical tests, like simple linear regression slope test and Durbin-Watson test, do not apply easily considering the shrinkage property of the lasso and LARS unless the selected model is the correct model. However, in high dimensional problems, existence of one true sparse model is questionable and not sustainable often. So it is undeniable that, before testing if the sparse model is true, we are also interested in testing whether there is a sparse model to begin with. When $n \ll p$, this question if a sparse model exists cannot be answered easily.

In this thesis, we first aim to understand the proposed method of [2] and its connections with the famous LARS algorithm. Moreover, we introduce a new LARS algorithm and a new test statistic. The new algorithm is designed to detect non-linear correlations between observed vector $\mathbf{y}$ and matrix $\mathbf{X}$. It is also believed to improve the existing state of the art. The new test statistic is related to Wilcoxon ranked-sum test of nonparametric samples, seeming to perform powerfully in detecting deviations from sparsity.

The thesis is organized as follows. Section 2 defines the lasso and LARS estimators and introduces the famous LARS algorithm. Section 3 proves two lemmas in [2] with details and introduces the new conditional inference idea of [2]. Section 4 and 5 propose the new LARS algorithm using Kendall τ correlation coefficient and study the distribution properties of τ defined in Section 5. In Section 6, we implement the proposed test statistic and perform simulations to show its good empirical properties. Lastly, Section 7 summarizes the paper and discusses about more open questions and possible forms of test statistic.

2 LARS Algorithm

Given the linear model with the error term $\boldsymbol{\varepsilon}$, we assume that each entry of $\boldsymbol{\varepsilon}$ is normally distributed with mean zero and standard deviation one such that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{4}$$

In practical problems, we intend to figure out how a small change in $\mathbf{X}$ would affect the change in $\mathbf{y}$, which leads to the problem of solving the estimated value of $\boldsymbol{\beta}^*$. By minimizing the sum of squares of error, we obtain the OLS estimators $\tilde{\boldsymbol{\beta}}_{\text{OLS}}$.

$$\tilde{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y.$$

Additionally, we can obtain the lasso estimator $\hat{\boldsymbol{\beta}}_{\text{lasso}}$.

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j^* x_{ij})^2 + \lambda\|\boldsymbol{\beta^*}\|. \tag{5}$$

for a tuning parameter λ.

Motivated by OLS and properties of the lasso, Efron and Tibshirani [4] derived a new estimation process and named the method as Least Angle Regression(LARS). In order to improve properties of OLS, LARS is built on an iterative evaluation of the correlation between $\mathbf{y}$ and residuals at each step. However, LARS performs poorly when the explanatory variables are correlated. We will provide simulation results of LARS estimators for $\boldsymbol{\beta}^*$ using correlated explanatory variables in this section. The simulation results motivate us to understand the flaws of the LARS algorithm and to propose modification of the algorithm in the next section.

We denote $\boldsymbol{x}_j$ to be the jth column of the explanatory matrix. That is, for $j = 1, 2, ..., p$, $\boldsymbol{x}_j$ is a column vector of n entries. We first standardize all observed data.

$$\begin{aligned} &\sum_{i=1}^{n} x_{ij} = 0, \sum_{i=1}^{n} x_{ij}^2 = 1 \qquad \text{for } j = 1, 2, ..., p; \\ &\sum_{i=1}^{n} y_i = 0, \sum_{i=1}^{n} y_i^2 = 1. \end{aligned} \tag{6}$$

Before we start with the LARS algorithm, we assume that the OLS solution for the linear model, $\tilde{\boldsymbol{\beta}}$, is known. We use $\hat{\boldsymbol{\beta}}$ to denote the LARS estimator of $\boldsymbol{\beta}^*$.

The idea of LARS is to only enter "as much" of a predictor in successive steps. That is, in each step, LARS will add one explanatory variable to the active set A, which is a subset of $\mathbf{X}$. After s steps, $|A|$ is equal to s and only s entries of the $\hat{\boldsymbol{\beta}}$ are nonzero.

In the first step, LARS identifies the explanatory variable that is most correlated with the response $\mathbf{y}$.

Consider a predictor vector: $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Let $\hat{\boldsymbol{r}} = y - \hat{\boldsymbol{\mu}}$ be the residual of the linear model where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{r}}$ will be updated after every step is finished. We denote the Pearson correlation between $\boldsymbol{x}_j$ and $\mathbf{y}$ for each j as cor($\boldsymbol{x}_j$,$\mathbf{y}$) for each column vector $\boldsymbol{x}_j$.

We know that the LARS algorithm works successively and selects a single variable at each step. In the first step, we start from $\hat{\boldsymbol{\mu}} = \mathbf{0}$. That is,

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \mathbf{0}, \\ \hat{\boldsymbol{r}} &= \mathbf{y} - \hat{\boldsymbol{\mu}} = \mathbf{y}; \\ cor(\boldsymbol{x}_j, \mathbf{y}) &= \frac{\langle \mathbf{x}_j, \mathbf{y}\rangle}{\langle \boldsymbol{x}_j, \boldsymbol{x}_j\rangle\langle \mathbf{y}, \mathbf{y}\rangle} = \langle \boldsymbol{x}_j, \mathbf{y}\rangle = \boldsymbol{x}_j^{\intercal}\mathbf{y} \end{aligned} \tag{7}$$

After comparing values of $cor(\boldsymbol{x}_j, \mathbf{y})$ for each j, we choose the explanatory variable with the biggest value of $cor(\boldsymbol{x}_j, \mathbf{y})$ and label it as $\boldsymbol{x}_{s_1}$. Notice that up to this point, though we are unable to identify the exact value of $\hat{\beta}_{s_1}$, we continue the algorithm to step two and denote the unknown coefficient for $\boldsymbol{x}_{s_1}$ as $\hat{\gamma}_1$.

Observe that the values of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{r}}$ have been updated at this point:

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \mathbf{0} + \hat{\gamma}_1 \boldsymbol{x}_{s_1}, \\ \hat{\boldsymbol{r}} &= \mathbf{y} - \hat{\boldsymbol{\mu}} = \mathbf{y} - \hat{\gamma}_1 \boldsymbol{x}_{s_1}. \end{aligned} \tag{8}$$

We continue with step two. In this step, we intend to identify the second selected variable $\boldsymbol{x}_{s_2}$ and to find the value of $\hat{\beta}_{s_1}$. We first make the value of $\hat{\gamma}_1$ bigger continuously from 0 to the corresponding OLS solution $\tilde{\beta}_{s_1}$. As the value of $\hat{\gamma}_1$ is increasing, we select $\boldsymbol{x}_{s_2}$ which is the first appearing variable to have the same amount of correlation with the residual as $\boldsymbol{x}_{s_1}$.That is, we intend to compare $cor(\boldsymbol{x}_j, \hat{\boldsymbol{r}})$ and $cor(\boldsymbol{x}_{s_1}, \hat{\boldsymbol{r}})$ where

$$\begin{aligned} cor(\boldsymbol{x}_j, \hat{\boldsymbol{r}}) &= \langle \boldsymbol{x}_j, \hat{\boldsymbol{r}} \rangle = \boldsymbol{x}_j^\intercal \hat{\boldsymbol{r}} = \boldsymbol{x}_j^\intercal (\mathbf{y} - \hat{\gamma}_1 \boldsymbol{x}_{s_1}), \\ cor(\boldsymbol{x}_{s_1}, \hat{\boldsymbol{r}}) &= \langle \boldsymbol{x}_{s_1}, \hat{\boldsymbol{r}} \rangle = \boldsymbol{x}_{s_1}^\intercal \hat{\boldsymbol{r}} = \boldsymbol{x}_{s_1}^\intercal (\mathbf{y} - \hat{\gamma}_1 \boldsymbol{x}_{s_1}) \end{aligned} \tag{9}$$

The reason that we are comparing $cor(\boldsymbol{x}_j, \hat{\boldsymbol{r}})$ and $cor(\boldsymbol{x}_{s_1}, \hat{\boldsymbol{r}})$ is to locate the second selected variable. Intuitively, we tend to select a new explanatory variable that equally greatly correlated with the response. However, before we jump to the conclusion of how to practically choose the second variable $\boldsymbol{x}_{s_2}$, we claim that, by comparing Pearson correlation, LARS algorithm fails to choose the best subset of explanatory variable if each explanatory variables are correlated.

For that end, we have designed following simulation under the usual linear regression setting as introduced earlier in this section with each $\boldsymbol{x}_j$ highly correlated to each other.

Lemma 2.1. *After LARS selects $\boldsymbol{x}_{s_1}$ in the first step, the LARS algorithms, in the second step, will return the variable x_{s_2} where cor($\boldsymbol{x}_{s_2}$, $\boldsymbol{y}$) is the second largest element among all cor($\boldsymbol{x}_j$, $\boldsymbol{y}$) for j = 1, 2, ..., p when $\boldsymbol{X}$ is orthogonal. However, the algorithm will return the variable that has most correlation with $\boldsymbol{x}_{s_1}$ in the second step if $\boldsymbol{X}$ is correlated.*

Proof of Lemma 2.1. Assume $\mathbf{X}$ is orthogonal. Suppose in the first step, we have identified $\boldsymbol{x}_{s_1}$ such that cor($\boldsymbol{x}_{s_1}$, $\mathbf{y}$) provides the largest value among all cor($\boldsymbol{x}_j$, $\mathbf{y}$) for j = 1, 2, ..., p. That is,

$$\begin{aligned} \mathrm{cor}(\boldsymbol{x}_j, \mathbf{y}) &\le \mathrm{cor}(\boldsymbol{x}_{s_1}, \mathbf{y}), \\ \langle \boldsymbol{x}_j, \mathbf{y} \rangle &\le \langle \boldsymbol{x}_{s_1}, \mathbf{y} \rangle, \quad \boldsymbol{x}_j^T \mathbf{y} \le \boldsymbol{x}_{s_1}^T \mathbf{y}, \quad \forall j = 1, 2, .., p. \end{aligned} \tag{10}$$

Also, suppose we have updated $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{r}}$. We are looking for $\boldsymbol{x}_{s_2}$ which has the largest value of $cor(\boldsymbol{x}_j, \hat{\boldsymbol{r}})$ among all $\boldsymbol{x}_j$ such that

$$cor(\boldsymbol{x}_j, \hat{\boldsymbol{r}}) \ge cor(\boldsymbol{x}_{s_1}, \hat{\boldsymbol{r}}). \tag{11}$$

We expand both sides of (11) and obtain

$$\begin{aligned} \langle \boldsymbol{x}_{s_2}, \hat{\boldsymbol{r}} \rangle &= \boldsymbol{x}_{s_2}^T \hat{\boldsymbol{r}} \\ &= \boldsymbol{x}_{s_2}^T (\mathbf{y} - \hat{\gamma}_1 \boldsymbol{x}_{s_1}) \\ &= \boldsymbol{x}_{s_2}^T \mathbf{y} - \hat{\gamma}_1 \boldsymbol{x}_{s_2}^T \boldsymbol{x}_{s_1} \\ &= \boldsymbol{x}_{s_2}^T \mathbf{y} - \hat{\gamma}_1 \mathbf{I} \\ \langle \boldsymbol{x}_{s_1}, \hat{\boldsymbol{r}} \rangle &= \boldsymbol{x}_{s_1}^T \hat{\boldsymbol{r}} \\ &= \boldsymbol{x}_{s_1}^T \mathbf{y} - \hat{\gamma}_1 \mathbf{I}. \end{aligned} \tag{12}$$

By (11), we get

$$\begin{aligned} \boldsymbol{x}_{s_2}^T \mathbf{y} - \hat{\gamma}_1 \mathbf{I} \geq \boldsymbol{x}_{s_1} \mathbf{y} - \hat{\gamma}_1 \mathbf{I} \\ \boldsymbol{x}_{s_2}^T \mathbf{y} \geq \boldsymbol{x}_{s_1}^T \mathbf{y}. \end{aligned} \tag{13}$$

Notice that from (10), we have the $\boldsymbol{x}_{s_2}^T \mathbf{y} \leq \boldsymbol{x}_{s_1}^T \mathbf{y}$. From (10) and (11), we are able to conclude that $\boldsymbol{x}_{s_2}^T \mathbf{y} = \boldsymbol{x}_{s_1}^T \mathbf{y}$.

Now we have two situations. First, there exists such $\boldsymbol{x}_{s_2}$ that $\boldsymbol{x}_{s_2}^T \mathbf{y} = \boldsymbol{x}_{s_1}^T \mathbf{y}$ holds. Second, no such $\boldsymbol{x}_{s_2}$ exists, and we are not able to find any $\boldsymbol{x}_j$ satisfying $\boldsymbol{x}_j^T \mathbf{y} \geq \boldsymbol{x}_{s_1}^T \mathbf{y}$ for all j.

If we are at the first situation, we are done with the proof. If we encounter the second situation , we can rewrite the problem to a new question. That is, we have a strictly descending sequence of $cor(\boldsymbol{x}_j, \mathbf{y})$, and we intend to find the closest value $cor(\boldsymbol{x}_{s_2}, \mathbf{y})$ of $cor(\boldsymbol{x}_{s_1}, \mathbf{y})$ such that $\boldsymbol{x}_{s_2}$ provides largest correlation with $\mathbf{y}$ among all columns $\boldsymbol{x}_j$ except $\boldsymbol{x}_{s_1}$. We are able to claim that the $\boldsymbol{x}_{s_2}$ we are looking for provides the second largest correlation with the response among all $\boldsymbol{x}_j$'s.

□

By reasoning of the previous lemma, we believe that LARS algorithm will fail for correlated designs, as it will choose the variable that is most correlated with previously selected $\boldsymbol{x}_{s_1}$, rather than with the response $\mathbf{y}$.

We illustrate the idea in a simulation study. Consider the following setup of a correlated $\mathbf{X}$:

$$\begin{aligned} \mathbf{X} &= (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5) \\ &= (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_1\boldsymbol{x}_2, \boldsymbol{x}_1^2, \boldsymbol{x}_1^3, \boldsymbol{x}_1^4) \end{aligned}$$

where each column of $\mathbf{X}$ is independent. Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$.

We simulate two independent random variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ from normal distribution with mean zero and standard deviation one. Assume the true value of $\boldsymbol{\beta}^*$ is equal to (0,0,1,1,0,-1). That is, we have six explanatory variables and they are highly correlated to one another. We also assume that each entry of the error term $\boldsymbol{\varepsilon}$ is normally distributed with mean zero and standard deviation one. Theoretically, LARS will return the coefficients of $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_5$ with zeros. We use current LARS algorithm to locate the optimal λ which provides the smallest Cross-Validation error. Next, we use the optimal value of λ to obtain the estimated value of the coefficients. We repeat this simulation for 100 times and count the number of times that each variable has a nonzero coefficient. The bar plot of the counts is given by Figure 1.

Notice in Figure 1 , $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ have the tendency to provide less nonzero coefficients in 100 times of repeated simulation as we expected. However, the proportion of nonzero coefficients of $\boldsymbol{x}_5$ is much bigger than the proportion of $\boldsymbol{x}_4$. The true value of β_4^* is equal to 1. However, as we notice in the bar plot, the frequency of nonzero $\hat{\beta}_4$'s is smaller than the frequency of nonzero $\hat{\beta}_5$ which has the true value 0. The simulation results demonstrate that, under our linear model setup, the LARS algorithm does not to choose explanatory variables with positive true $\boldsymbol{\beta}^*$ values but the variables that are highly correlated with $\boldsymbol{x}_1$ even the true β_6^* is zero.

3 A Significance Test for the Lasso

In [2], a covariance test statistic is proposed for testing the significance of the predictor variable that enters the current lasso model, in the sequence of models visited along the lasso

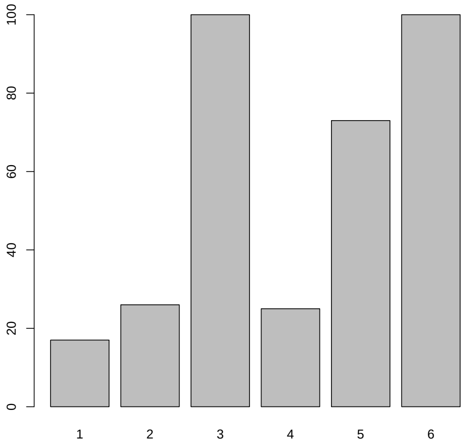

Figure 1: Barplot of Nonzero Coefficients' Counts

solution path. The paper succeeded in proving that, when the true model is linear, this covariance test statistic has exponential distribution with parameter 1 asymptotically under the null hypothesis. The null states that all truly active variables are contained in the current lasso model. The significance of covariance test is that it introduces a new practical way of testing random active set A during the process of lasso.

Explicitly, consider the following linear regression set-up:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}. \tag{14}$$

The lasso estimator is defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|^2 + \lambda|\boldsymbol{\beta}^*|. \tag{15}$$

where λ is the tuning parameter, controlling the level of sparsity in $\hat{\boldsymbol{\beta}}$.

Suppose set A is the active set just before λ taking value as λ_k and that explanatory variable $\boldsymbol{x}_j$ enters into the active set at λ_k. Notice that here $\boldsymbol{x}_j$ with a column vector with n-entires.

We use the following notations in this section. Define two estimators of $\boldsymbol{\beta}^*$ as follows

$$\begin{aligned} \hat{\boldsymbol{\beta}}(\lambda_{k+1}) &= \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*\|^2 + \lambda_{k+1}|\boldsymbol{\beta}^*|, \\ \tilde{\boldsymbol{\beta}}_A(\lambda_{k+1}) &= \arg\min_{\boldsymbol{\beta}_A} \frac{1}{2}\|\mathbf{Y} - \boldsymbol{X_A}\boldsymbol{\beta}_A^*\|^2 + \lambda_{k+1}|\boldsymbol{\beta}_A^*|. \end{aligned} \tag{16}$$

where $\hat{\boldsymbol{\beta}}(\lambda_{k+1})$ is the solution of the lasso at the next knot in the path λ_{k+1}, using explanatory variables in $A \cup \boldsymbol{x}_j$ which is denoted in short as $A \cup \{j\}$ and $\tilde{\boldsymbol{\beta}}_A(\lambda_{k+1})$ is the solution of the Lasso problem using only active set A at $\lambda = \lambda_{k+1}$. That is, $\hat{\boldsymbol{\beta}}(\lambda_{k+1})$ is the lasso estimator of $\boldsymbol{\beta^*}$ at the $(k+1)$st step. After extracting the value of the tuning parameter λ_{k+1} and the selected variables from $\hat{\boldsymbol{\beta}}(\lambda_{k+1})$, we compute the lasso estimator $\tilde{\boldsymbol{\beta}}(\lambda_{k+1})$ which using the tuning parameter λ_{k+1} and the selected variables.

Therefore we can write the covariance test as

$$H_0 : supp(\boldsymbol{\beta}^*) \subseteq A, \qquad H_1 : supp(\boldsymbol{\beta}^*) \nsubseteq A$$

The test statistic of covariance test is defined as

$$T_k = \frac{\langle \mathbf{y}, \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda_{k+1})\rangle - \langle \mathbf{y}, \boldsymbol{X}_{\boldsymbol{A}}\tilde{\boldsymbol{\beta}}_{\boldsymbol{A}}(\lambda_{k+1})\rangle}{\sigma^2}. \tag{17}$$

The above test statistics measures the difference between Pearson correlations of the observed data $\mathbf{y}$ and the predicted data $\mathbf{X}\hat{\boldsymbol{\beta}}(\lambda_{k+1})$ and $\boldsymbol{X}_{\boldsymbol{A}}\tilde{\boldsymbol{\beta}}_{\boldsymbol{A}}(\lambda_{k+1})$.

Lemma 3.1. *The test statistics T_k defined in (17) satisfies the representation*

$$T_k = C(A, \boldsymbol{s}_A, j, s)\lambda_k(\lambda_k - \lambda_{k+1})/\sigma^2,$$

where

$$C(A, \boldsymbol{s}_A, j, s) = \|(\boldsymbol{X}^T_{A\cup\{j\}})^+\boldsymbol{s}_{A\cup\{j\}} - (\boldsymbol{X}^T_A)^+\boldsymbol{s}_A\|^2,$$

and $\mathbf{X}^+$ *denotes the pseudo-inverse of a matrix* $\mathbf{X}$*.*

Proof of Lemma 3.1. From the first order (KKT) condition, we get

$$\hat{\boldsymbol{\beta}}_A(\lambda) = (\mathbf{X}^T_A\mathbf{X}_A)^{-1}\mathbf{X}^T_A\mathbf{y} - \lambda(\mathbf{X}^T_A\mathbf{X}_A)^{-1}s_A, \text{where } \mathbf{s}_A = sgn(\hat{\boldsymbol{\beta}}_A(\lambda)). \tag{18}$$

Let $\mathbf{P}_A$ and $(\mathbf{X}^T_A)^+$ be the projection onto the column space of $\mathbf{X}_A$ and pseudoinverse of $\mathbf{X}_A$ respectively. Therefore, we have

$$\begin{aligned} \mathbf{P}_A &= \mathbf{X}_A(\mathbf{X}^T_A\mathbf{X}_A)^{-1}\mathbf{X}^T_A, \\ (\mathbf{X}^T_A)^+ &= \mathbf{X}_A(\mathbf{X}^T_A\mathbf{X}_A)^{-1}. \end{aligned} \tag{19}$$

Then we obtain

$$\begin{aligned} \mathbf{X}\hat{\boldsymbol{\beta}}_{k+1} &= \mathbf{P}_{A\cup\{j\}}\mathbf{y} - \lambda_{k+1}(\mathbf{X}^T_{A\cup\{j\}})^+\mathbf{s}_{A\cup\{j\}}, \\ \mathbf{X}_A\tilde{\boldsymbol{\beta}}(\lambda_{k+1}) &= \mathbf{P}_A\mathbf{y} - \lambda_{k+1}(\mathbf{X}^T_A)^+\mathbf{s}_A. \end{aligned} \tag{20}$$

Plug in (20) into (17) and expand the inner product, we can transform (17) into the following

$$T_k = \mathbf{y}^T(\mathbf{P}_{A\cup\{j\}} - \mathbf{P}_A)\mathbf{y}/\sigma^2 - \lambda_{k+1}\mathbf{y}^T((\mathbf{X}^T_{A\cup\{j\}})^+\mathbf{s}_{A\cup\{j\}} - (\mathbf{X}^T_A)^+\mathbf{s}_A)/\sigma^2. \tag{21}$$

Recall that by the continuity of the lasso solution path at λ_k, we have

$$\begin{aligned} \mathbf{X}_{X\cup\{j\}}\hat{\boldsymbol{\beta}}_{A\cup\{j\}}(\lambda_k) &= \mathbf{X}_A\hat{\boldsymbol{\beta}}_A + \mathbf{x}_j\hat{\beta}_j \\ &= \mathbf{X}_A\hat{\boldsymbol{\beta}}_A \end{aligned} \tag{22}$$

We rewrite (20),

$$\begin{aligned} \mathbf{X}_A\hat{\boldsymbol{\beta}}_A(\lambda_k) &= \mathbf{P}_A\mathbf{y} - \lambda_k(\mathbf{X}^T_A)^+\mathbf{s}_A \\ &= \mathbf{P}_{A\cup\{j\}}\mathbf{y} - \lambda_k(\mathbf{X}^T_{A\cup\{j\}})^+\mathbf{s}_{A\cup\{j\}} \\ &= \mathbf{X}_{X\cup\{j\}}\hat{\boldsymbol{\beta}}_{A\cup\{j\}}(\lambda_k). \end{aligned} \tag{23}$$

and obatin:

$$(\mathbf{P}_{A\cup\{j\}} - \mathbf{P}_A)\mathbf{y} = \lambda_k((\mathbf{X}^T_{A\cup\{j\}})^+\mathbf{s}_{A\cup\{j\}} - (\mathbf{X}^T_A)^+\mathbf{s}_A). \tag{24}$$

We first square both sides of (24),

$$\mathbf{y}^T(\mathbf{P}_{A\cup\{j\}} - \mathbf{P}_A)^T(\mathbf{P}_{A\cup\{j\}} - \mathbf{P}_A)\mathbf{y} = \lambda_k^2\|(\mathbf{X}_{A\cup\{j\}}^T)^+\mathbf{s}_{A\cup\{j\}} - (\mathbf{X}_A^T)^+\mathbf{s}_A\|^2 \tag{25}$$

Notice that $\mathbf{P}_{A\cup\{j\}} - \mathbf{P}_A$ is a projection onto the column space of $\mathbf{X}_A$, therefore $(\mathbf{P}_{A\cup\{j\}} - \mathbf{P}_A)^T(\mathbf{P}_{A\cup\{j\}} - \mathbf{P}_A) = \mathbf{P}_{A\cup\{j\}} - \mathbf{P}_A$ and we obtain (25) in the following form

$$\mathbf{y}^T(\mathbf{P}_{A\cup\{j\}} - \mathbf{P}_A)\mathbf{y} = \lambda_k^2\|(\mathbf{X}_{A\cup\{j\}}^T)^+\mathbf{s}_{A\cup\{j\}} - (\mathbf{X}_A^T)^+\mathbf{s}_A\|^2 \tag{26}$$

Next we take the inner product of both sides of (24) with $\mathbf{y}$ and plug the new equation into (26).

$$\begin{aligned} \mathbf{y}^T(\mathbf{P}_{A\cup\{j\}} - \mathbf{P}_A)\mathbf{y} &= \mathbf{y}^T\lambda_k((\mathbf{X}_{A\cup\{j\}}^T)^+\mathbf{s}_{A\cup\{j\}} - (\mathbf{X}_A^T)^+\mathbf{s}_A), \\ \lambda_k^2\|(\mathbf{X}_{A\cup\{j\}}^T)^+\mathbf{s}_{A\cup\{j\}} - (\mathbf{X}_A^T)^+\mathbf{s}_A\|^2 &= \mathbf{y}^T\lambda_k((\mathbf{X}_{A\cup\{j\}}^T)^+\mathbf{s}_{A\cup\{j\}} - (\mathbf{X}_A^T)^+\mathbf{s}_A). \end{aligned} \tag{27}$$

Notice that,

$$\|(\mathbf{X}_{A\cup\{j\}}^T)^+\mathbf{s}_{A\cup\{j\}} - (\mathbf{X}_A^T)^+\mathbf{s}_A\|^2 = \|(\mathbf{X}_{\{j\}}^T)^+\mathbf{s}_{\{j\}}\|^2 = 1 \tag{28}$$

Then we plug in (26) and (27) into (21),

$$\begin{aligned} T_k &= \mathbf{y}\mathbf{P}_{A\cup\{j\}} - \mathbf{P}_A)\mathbf{y}/\sigma^2 - \lambda_{k+1}\mathbf{y}((\mathbf{X}_{A\cup\{j\}}^T)^+\mathbf{s}_{A\cup\{j\}} - (\mathbf{X}_A^T)^+\mathbf{s}_A)/\sigma^2 \\ &= \frac{1}{\sigma^2}\lambda_k^2\|(\mathbf{X}_{A\cup\{j\}}^T)^+\mathbf{s}_{A\cup\{j\}} - (\mathbf{X}_A)^T)^+\mathbf{s}_A\|^2 - \frac{1}{\sigma^2}\lambda_{k+1}\lambda_k\|(\mathbf{X}_{A\cup\{j\}}^T)^+\mathbf{s}_{A\cup\{j\}} - (\mathbf{X}_A^T)^+\mathbf{s}_A\|^2 \\ &= \frac{1}{\sigma^2}(\lambda_k^2 - \lambda_{k+1}\lambda_k)\|(\mathbf{X}_{A\cup\{j\}}^T)^+\mathbf{s}_{A\cup\{j\}} - (\mathbf{X}_A^T)^+\mathbf{s}_A\|^2 \\ &= \frac{1}{\sigma^2}(\lambda_k^2 - \lambda_{k+1}k), \end{aligned}$$

where the last line follows from (28). □

Consider a special case of an orthogonal predictor matrix $\mathbf{X}$. We are interested in finding the test statistic T_k for the first predictor to enter the active set A, i.e., T_1.

Lemma 3.2. *Suppose the predictor matrix $\boldsymbol{X}$ is orthogonal. We denote $U_j = \langle \boldsymbol{x}_j, \boldsymbol{y}\rangle = \boldsymbol{x}_j^T\boldsymbol{y}$ for $j = 1, 2, ..., p$, then the knots in the lasso paths(values of λ at which the coefficients become nonzero) are: $\lambda_1 = |U_{(1)}|, \lambda_2 = |U_{(2)}|, ..., \lambda_p = |U_{(p)}|$ where $|U_{(1)}| \geq |U_{(2)}| \geq ... \geq |U_{(p)}|$ are the order statistics of $|U_{(1)}|, |U_{(2)}|, ..., |U_{(p)}|$.*

Lemma 3.3. *Under the null, $U_1, ..., U_p$ are identically independent distributed from normal distribution with mean zero and variance σ^2, so $|U_1|/\sigma, ...|U_p|/\sigma$ follow a χ_1 distribution. Therefore $T_1 = |U_{(1)}|(|U_{(1)}| - |U_{(2)}|)/\sigma^2 \xrightarrow{d} Exp(1)$.*

Proof of Lemma 3.3. Let $F(x)$ be the Cumulative Distribution Function of χ_1 distribution:

$$F(x) = (2\Phi(x) - 1)1(x > 0),$$

where $\Phi(x)$ is the standard normal CDF. In order to make sure that $F(x)$ is in the domain of attraction of G_γ according to Theorem1.1.8 in [8]. We first compute the value of γ by evaluating

$$
\begin{aligned}
\lim_{t\to\infty} \frac{(1-F(t))F''(t)}{(F'(t))^2} &= \lim_{t\to\infty} \frac{F''(t)}{F'(t)} \frac{1-F(t)}{F'(t)} \\
&= \lim_{t\to\infty} \frac{-2t\phi(t)}{2\phi(t)} \frac{1-2\Phi(t)+1}{2\phi(t)} \\
&= \lim_{t\to\infty} -\frac{t(1-\Phi(x))}{\phi(x)} \\
&= \lim_{t\to\infty} -\frac{t}{\lambda(t)} \\
&= \lim_{t\to\infty} -t \times m(t) \\
&= -1 = -\lambda - 1
\end{aligned} \tag{29}
$$

where $\phi(x)$ is the standard normal PDF. According to Mill's ratio, the hazard function $\lambda(t) = \frac{\phi(x)}{1-\Phi(x)}$. When X follows standard normal distribution, Mill's Ratio $m(x) = \frac{1}{\lambda(x)} \sim \frac{1}{x}$. Therefore we have determined that $\lambda = 0$ which implies that $F(x)$ is in the domain of attraction of G_0.

Theorem 2.2.1 in [8] implies, for real constants $a_p = H(p) = F^{-1}(1 - \frac{1}{p})$, where $H(p)$ is the left-continuos inverse of $\frac{1}{1-F(x)}$, and non-negative constants $b_p = \frac{1}{pH'(p)} = pF'(a_p)$, random variables $W_1 = b_p(U_{(1)} - a_p)$ and $W_2 = b_p(U_{(2)} - a_p)$ converge to standard normal in distribution when $p \to \infty$. Our next goal is to find joint converging distribution of W_1 and W_2. First look at $U_{(1)}$ and $U_{(2)}$. $U_{(1)}$ is the maximum of the order statistics, then

$$
\lim_{n\to\infty} F^n(\frac{1}{b_p}x + a_p) = \exp(-e^{-x}), \tag{30}
$$

and $U_{(2)}$ is the second maximum element,

$$
\begin{aligned}
&\lim_{n\to\infty} nF^{n-1}(\frac{1}{b_p}x + a_p)(1 - F(\frac{1}{b_p}x + a_p)) + F^n(\frac{1}{b_p}x + a_p) \\
&= -\log G_0(x) G_0(x) + G_0(x) \\
&= (1 + e^{-x})\exp(-e^{-x}).
\end{aligned} \tag{31}
$$

By extreme value distributions [8], we conclude that

$$
(W_1, W_2) \xrightarrow{d} (-logE_1, -log(E_1 + E_2)),
$$

where E_1 and E_2 are two independent standard exponential distributions.

We rewrite $U_{(1)}(U_{(1)} - U_{(2)})$ in terms of W_1 and W_2.

$$
U_{(1)}(U_{(1)} - U_{(2)}) = (a_p + \frac{W_1}{b_p})(\frac{W_1 - W_2}{b_p}) = \frac{a_p}{b_p}(W_1 - W_2) + \frac{W_1(W_1 - W_2)}{b_p}. \tag{32}
$$

Notice that $a_p = F^{-1}(1 - \frac{1}{p})$ and $b_p = pF'(p)$. In this proof we are dealing with standard normal distributions. Therefore, we can rewrite a_p and b_p in terms of $\phi(x)$ and $\Phi(x)$, the standard normal density and probability functions. That is,

$$
\begin{aligned}
&1 - \frac{1}{p} = 2\Phi(a_p) - 1 \quad \text{i.e.,} 1 - \Phi(a_p) = \frac{1}{2p}, \\
&b_p = 2p\phi(a_p).
\end{aligned} \tag{33}
$$

By (33), we know that when b_p approaches infinity, the term $\frac{W_1(W_1-W_2)}{b_p}$ converges to zero. Next we want to find the limit of $a_p b_p$. Using Mill's inequalities, we obtain

$$\frac{\phi(a_p)}{a_p}\frac{1}{1+\frac{1}{a_p^2}} \leq 1-\Phi(a_p) \leq \frac{\phi(a_p)}{a_p}. \tag{34}$$

We multiply (34) by $2p$ and make a_p approaches infinity. Notice that $b_p = 2p\phi(a_p)$ and $\frac{1}{2p} = 1-\Phi(a_p)$. We get

$$\frac{b_p}{a_p}\frac{1}{1+\frac{1}{a_p^2}} \leq 1 \leq \frac{b_p}{a_p}. \tag{35}$$

Therefore by squeezing theorem, we conclude that $\frac{b_p}{a_p}$ converges to 1, which leads to the result that $\frac{a_p}{b_p}$ converges to 1.

We know that $W_1 - W_2$ converges in distribution to $log(E_2 + E_1) - log(E_1)$, which is also a standard exponential distribution. Therefore by (32), we are able to conclude that the limiting distribution of $U_{(1)}(U_{(1)} - U_{(2)})$ is standard exponential distribution. □

4 τ-LARS algorithm

As motivated by the counterexample we provided in Section 2, we introduce a new LARS algorithm using the Kendall's τ correlation coefficient instead of Pearson correlation. The new algorithm mimics the iterative feature of the LARS algorithm and selects one explanatory variable at each step.

Given two vectors $\mathbf{a}$ and $\mathbf{b}$ of n entries, Kendall's τ correlation coefficient is defined as

$$\tau(\mathbf{a},\mathbf{b}) = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{k=1,k>i}^{n} sgn(a_i-a_k)sgn(b_i-b_k). \tag{36}$$

where the sgn function is defined by

$$sgn(s-t) = \begin{cases} 1 & \text{if } s-t>0 \\ 0 & \text{if } s-t=0 \\ -1 & \text{if } s-t<0 \end{cases} \text{ for real numbers } s \text{ and } t.$$

Croux and Dehon [9] have proved that, if $\mathbf{X}$ and $\mathbf{Y}$ have bivariate distribution, Kendall's τ has infinitesimal robustness as a positive feature. That is, the influence function is bounded.

Like the regular LARS algorithm, we suppose the prediction vector to be $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and residual to be $\hat{\boldsymbol{r}} = \mathbf{y}$ - $\hat{\boldsymbol{\mu}}$ for each step. Notice that in the first step, we consider,

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \mathbf{0}, \\ \hat{\boldsymbol{r}} &= \mathbf{y} - \hat{\boldsymbol{\mu}} = \mathbf{y}. \end{aligned} \tag{37}$$

In step one we need to identify the explanatory variable that has the most correlation with the response. Thus we compare $\tau(\boldsymbol{x}_j, \hat{\boldsymbol{r}})$ for each column vector $\boldsymbol{x}_j$. Notice that each $\boldsymbol{x}_j$ and $\hat{\boldsymbol{r}}$ are column vectors with n entries.

The Kendall's τ correlation coefficient for each $\boldsymbol{x}$ and $\hat{r}$ is:

$$\begin{aligned}\tau(\boldsymbol{x}_j, \hat{\boldsymbol{r}}) &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1,k>i}^{n} sgn(x_{ij} - x_{kj}) sgn(\hat{r}_i - \hat{r}_k) \\ &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1,k>i}^{n} sgn(x_{ij} - x_{kj}) sgn(y_i - y_k).\end{aligned} \tag{38}$$

After comparing values of $\tau(\boldsymbol{x}_j, \hat{\boldsymbol{r}})$ for j = 1, 2,...,p, we choose the explanatory variable which gives the biggest correlation and label it as $\boldsymbol{x}_{s_1}$.

Therefore we continue with the algorithm and move to step two. We first update our prediction vector and residual. After choosing $\boldsymbol{x}_{s_1}$, we have:

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= \hat{\boldsymbol{\mu}} + \hat{\gamma}_1 \boldsymbol{x}_{s_1}, \\ \hat{\boldsymbol{r}} &= \mathbf{y} - \hat{\boldsymbol{\mu}} = \mathbf{y} - \hat{\gamma}_1 \boldsymbol{x}_{s_1}\end{aligned} \tag{39}$$

where $\hat{\gamma}_1$ is the LARS coefficient for explanatory variable x_{s_1} based on the new algorithm. Notice that after selecting the first variable, we are unable to identify the value of $\hat{\gamma}_1$. However, we will solve the problem in step two while we select the second variable by comparing how the rest of the explanatory variables are correlated with residual $\hat{r}$ and the current residual of $\boldsymbol{x}_{s_1}$. That is, we will select both the second variable and the value of $\hat{\gamma}_1$. We first calculate:

$$\begin{aligned}\tau(\boldsymbol{x}_j, \hat{\boldsymbol{r}}) &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1,k>i}^{n} sgn(x_{ij} - x_{kj}) sgn(\hat{r}_i - \hat{r}_k) \\ &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1,k>i}^{n} sgn(x_{ij} - x_{kj}) sgn((y - \hat{\gamma}_1 x_{s_1})_i - (y - \hat{\gamma}_1 x_{s_1})_k) \\ &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1,k>i}^{n} sgn(x_{ij} - x_{kj}) sgn((y_i - \hat{\gamma}_1 x_{is_1}) - (y_k - \hat{\gamma}_1 x_{ks_1})) \\ &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1,k>i}^{n} sgn(x_{ij} - x_{kj}) sgn((y_i - y_k) - \hat{\gamma}_1 (x_{is_1} - x_{ks_1})),\end{aligned} \tag{40}$$

and

$$\tau(\boldsymbol{x}_{s_1}, \hat{\boldsymbol{r}}) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1,k>i}^{n} sgn(x_{is_1} - x_{ks_1}) sgn((y_i - y_k) - \hat{\gamma}_1 (x_{is_1} - x_{ks_1})). \tag{41}$$

While making $\hat{\gamma}_1$ bigger from 0, we select the first appearing $\boldsymbol{x}_j$ such that (40) have the same value as (41). That is, we want $\boldsymbol{x}_{s_2}$ to be the first appearing explanatory variable while the value of $\hat{\gamma}_1$ changes from 0 to $\tilde{\beta}_{s_1}$ such that:

$$\tau(\boldsymbol{x}_{s_2}, \hat{\boldsymbol{r}}) = \tau(\boldsymbol{x}_{s_1}, \hat{\boldsymbol{r}}) \tag{42}$$

We repeat the algorithm successively. That is, at step s,we only select s variables and identify (s-1) LARS coefficient.

5 Distribution Properties of τ

In this section, we define τ to be the correlation between observed data and predicted data. Meanwhile we explore the distribution properties of τ under the linear regressions setting of previous sections. The analysis is non-trivial as the coefficient τ is a sum of dependent random variables. We first establish dependency patterns and then discuss further distribution properties of τ. Notice that in this section we denote $\mathbf{x}_i$ to be the transpose vector of the ith row vector of the matrix $\mathbf{X}$. Observe that conditionally on $\mathbf{X}$, the response vector $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$.

We define τ as

$$\begin{aligned} \tau = \tau(\mathbf{y}, \mathbf{X}\hat{\boldsymbol{\beta}}) &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1,k>i}^{n} sgn(y_i - y_k) sgn({\mathbf{x}_i}^T\hat{\boldsymbol{\beta}} - {\mathbf{x}_k}^T\hat{\boldsymbol{\beta}}) \\ &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1,k>i}^{n} sgn((y_i - y_k)({\mathbf{x}_i}^T\hat{\boldsymbol{\beta}} - {\mathbf{x}_k}^T\hat{\boldsymbol{\beta}})). \end{aligned} \tag{43}$$

By the definition of sign function, we have $sgn(x)sgn(y) = sgn(xy)$ for any real numbers x and y.

We treat $Z_{ik} = (y_i - y_k)({\mathbf{x}_i}^T\hat{\boldsymbol{\beta}} - {\mathbf{x}_k}^T\hat{\boldsymbol{\beta}})$ as a random variable since the estimator $\hat{\boldsymbol{\beta}}$ is a random variable.

We know that each ε_i has standard normal distribution. Then $y_i = {\mathbf{x}_i}^T\boldsymbol{\beta^*} + \epsilon_i$ for $i = 1, ..., n$, where $\boldsymbol{\beta^*}$ is the true value of $\boldsymbol{\beta}$. Each y_i follows a normal distribution with mean ${\mathbf{x}_i}^T\boldsymbol{\beta^*}$ and variance ${\mathbf{x}_i}^T\boldsymbol{\beta^*}{\boldsymbol{\beta^*}}^T\mathbf{x}_i$. Notice that Z_{ik} is defined as the product of two continuous random variables. Therefore, Z_{ik} follows a continuous distribution.

We denote the Cumulative Distribution Function for Z_{ik} as F_{ik}.

Define function $f(Z_{ik}) = sgn(Z_{ik})$ for each pair (i, k). Then we have

$$f(Z_{ik}) = \begin{cases} 1 & \text{if } Z_{ik} > 0 \\ 0 & \text{if } Z_{ik} = 0 \\ -1 & \text{if } Z_{ik} < 0 \end{cases} \tag{44}$$

Since Z_{ik} is a random variable, we can consider $f(Z_{ik})$ as a random variable transformed by function f. Consider its probability mass function:

$$\begin{aligned} &\mathrm{P}(f(Z_{ik}) = 1) = \mathrm{P}(Z_{ik} > 0) = 1 - \mathrm{P}(Z_{ik} < 0) = 1 - F_{ik}(0); \\ &\mathrm{P}(f(Z_{ik}) = 0) = \mathrm{P}(Z_{ik} = 0) = 0 \text{ since } Z_{ik} \text{ is continuously distributed;} \\ &\mathrm{P}(f(Z_{ik}) = -1) = \mathrm{P}(Z_{ik} < 0) = F_{ik}(0). \end{aligned}$$

Lemma 5.1. *For any two Z_{ik} and Z_{pq} where $k > i$ and $q > p$, if $i \neq p$ and $k \neq q$ then Z_{ik} and Z_{pq} are independent. If $i = p$ or $k = q$, then Z_{ik} and Z_{pq} are dependent.*

Proof of Lemma 5.1. Consider Z_{ik} and Z_{pq} where $k > i$ and $q > p$.

If $p \neq i$ and $q \neq k$, then Z_{ik} is independent of Z_{pq} because each y_i is independent and identically distributed..

If $i = p$ or $k = q$, it suffices to show the case when $i = p$.

From the definition of Z, Z_{ik} and Z_{iq} carry information about y_k and y_q which is related to Z_{kq} if $q > k$. Therefore, Z_{ik} and Z_{iq} are dependent of each other.

□

In classical literature the distribution of τ, when represented as a sum of independent components, is known to be asymptotically normal with asymptotic variance of $4/9n$. Based on the result of Lemma 5.1 above, we know that independence assumption is no longer true in our setting.

Let $G_{iq,ik}(z)$ be the joint CDF of Z_{iq} and Z_{ik}. We first compute mean and variance of $f(Z_{ik})$'s.

$$\begin{aligned}
\mathrm{E}(f(Z_{ik})) &= 1 \times (1 - F_{ik}(0) + 0 \times 0 + (-1) \times (F_{ik}(0)) \\
&= 1 - 2F_{ik}(0) \\
\mathrm{Var}(f(Z_{ik})) &= \mathrm{E}(f(Z_{ik})^2) - \mathrm{E}(f(Z_{ik}))^2 \\
&= (1 \times (1 - F_{ik}(0)) + 1 \times F_{ik}(0)) - (1 - 2F_{ik}(0))^2 \\
&= 1 - (1 - F_{ik}(0))^2
\end{aligned} \tag{45}$$

We observe $f(Z_{iq})f(Z_{ik})$ first:

$$f(Z_{iq})f(Z_{ik}) = \begin{cases} 1 & \text{if } f(Z_{iq}) = 1 \text{ and } f(Z_{ik}) = 1 \\ -1 & \text{if } f(Z_{iq}) = -1,\ f(Z_{ik}) = 1 \text{ or } f(Z_{iq}) = 1,\ f(Z_{ik}) = -1 \end{cases}$$

Consider the probabilities of $f(Z_{iq})f(Z_{ik})$:

$$\begin{aligned}
\mathrm{P}(f(Z_{iq})f(Z_{ik}) = 1) &= \mathrm{P}(f(Z_{iq}) = 1, f(Z_{ik}) = 1) \\
&= \mathrm{P}(Z_{iq} > 0, f(Z_{ik}) > 0) \\
&= 1 - \mathrm{P}(Z_{iq} < 0, Z_{ik} < 0) = 1 - G_{iq,ik}(0,0); \\
\mathrm{P}(f(Z_{iq})f(Z_{ik}) = -1) &= \mathrm{P}(f(Z_{iq}) = 1, f(Z_{ik}) = -1) + \mathrm{P}(f(Z_{iq}) = -1, f(Z_{ik}) = 1) \\
&= \mathrm{P}(Z_{iq} > 0, Z_{ik} < 0) + \mathrm{P}(Z_{iq} < 0, Z_{ik} > 0) \\
&= (\mathrm{P}(Z_{ik} < 0) - \mathrm{P}(Z_{iq} < 0, Z_{ik}) < 0)) + (\mathrm{P}(Z_{iq} < 0) - \mathrm{P}(Z_{iq} < 0, Z_{ik} < 0)) \\
&= (F_{ik}(0) - G_{iq,ik}(0,0)) + (F_{iq}(0) - G_{iq,ik}(0,0)) \\
&= F_{ik}(0) + F_{iq}(0) - 2G_{iq,ik}(0,0)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Cov}_{ik,iq} &= \mathrm{Cov}(f(Z_{iq}), f(Z_{ik})) \\
&= \mathrm{E}(f(Z_{iq})f(Z_{ik}))) - \mathrm{E}(f(Z_{iq}))\mathrm{E}(f(Z_{ik})) \\
&= 1(1 - G_{iq,ik}(0,0)) + (-1)(F_{ik}(0) + F_{iq}(0) - 2G_{iq,ik}(0,0)) \\
&- (1 - 2F_{ik}(0)) \times (1 - 2F_{iq}(0)) \\
&= F_{ik}(0) + F_{iq}(0) + G_{iq,ik}(0,0) - 4F_{ik}(0)F_{iq}(0)
\end{aligned} \tag{46}$$

Using the definition of expectation and variance, we obtained the mean and variance of $\tau = \tau(\mathbf{y}, \mathbf{X}\hat{\boldsymbol{\beta}})$ using Eq (45).

$$\begin{aligned}
\mathrm{E}(\tau) &= \mathrm{E}(\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1,k>i}^{n} f(Z_{ik})) \\
&= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1,k>i}^{n} \mathrm{E}(f(Z_{ik})) \\
&= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{k=1,k>i}^{n} (1 - 2F_{ik}(0))
\end{aligned} \tag{47}$$

By Eq (46)

$$\begin{aligned}
\mathrm{Var}(\tau) &= \mathrm{Var}(\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{k=1,k>i}^{n} f(Z_{ik})) \\
&= (\frac{2}{n(n-1)})^2 \mathrm{Var}(\sum_{i=1}^{n}\sum_{k=1,k>i}^{n} f(Z_{ik})) \\
&= (\frac{2}{n(n-1)})^2 \left(\sum_{i=1}^{n}\sum_{k=1,k>i}^{n} \mathrm{Var}(f(Z_{ik}) + \sum_{i=1}^{n}\sum_{k=1,q\neq k,k>i,q>i}^{n} \mathrm{Cov}(f(Z_{ik}), f(Z_{iq}))\right) \\
&= (\frac{2}{n(n-1)})^2 \left(\sum_{i=1}^{n}\sum_{k=1,k>i}^{n} (1-(1-F_{ik}(0))^2 + \sum_{i=1}^{n}\sum_{k=1,q\neq k,k>i,q>i}^{n} \mathrm{Cov}_{ik,iq}\right)
\end{aligned} \tag{48}$$

By Lemma 5.1, we notice that τ is composed of two groups of sums of $f(Z_{ik})$. Define two index set B and D as the following

$$B = \{(i,k) : 1 \leq i, p \leq n, 1 \leq k, q \leq n, k > i, q > p, \text{ and } i \neq p, k \neq q, ie.T_{ik} \text{ is indenpendent of } T_{pq}.\}$$
$$D = \{(i,k) : 1 \leq i, p \leq n, 1 \leq k, q \leq n, k > i, q > p, \text{ and } i = p \text{ or } k = q, ie.T_{ik} \text{ is dependent of } T_{pq}.\}$$

Then we can rewrite τ and manipulate with constant such that

$$\begin{aligned}
\frac{n(n-1)}{2}\tau &= \sum_{i=1}^{n}\sum_{k=1,k>i}^{n} f(Z_{ik}) \\
&= \sum_{(i,k)\in B} H_{ik} + \sum_{(i,k)\in D} H_{ik} \\
&= R + S
\end{aligned} \tag{49}$$

where R is the sum of all independent pairs of $f(Z_{ik})$ which take values in $\{1,-1\}$ and S is the sum of all dependent pairs. Remember that

$$H_{ik} = f(Z_{ik}) = sgn\left((y_i - y_k)(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}} - \mathbf{x}_k{}^T\hat{\boldsymbol{\beta}})\right).$$

For example, if we have $n = 4$, the possible pairs of (i,k) are $\{(1,2), (1,3), (1,4), (2,3), (2,4),$ and $(3,4)\}$. According to the definition of set B and D, we have $R = H_{12} + H_{34}$ since H_{12} and H_{34} are independent of each other. Therefore, S will equal to the sum of rest of the H_{ik}'s.

Observe that, according to the discussion above, R is the sum of independent Bernoulli random variables with different probabilities of success. Therefore, it follows a Poisson Binomial distribution with probability of each Bernoulli success $(1 - F_{ik}(0))$. The distribution of R can be approximated with Poisson distribution [10] with parameter

$$\lambda = \sum_{(i,k)\in B} (1 - F_{ik}(0)). \tag{50}$$

The random variable S, on the other hand, is the sum of all dependent Bernoulli random variables with different probabilities of success. The probability mass function of S follows the following model for each binary random variable H_{ik} [11].

Let $|D| = d$ and $H(D) = \{H_{ik} : (i,k) \in D\}$. Assume that all H_{ik}'s are sorted by the increasing order of i and k. Define S_t to be the sum of the first t elements in $H(D)$. That is, if $n = 4$, then $H(D) = \{H_{13}, H_{14}, H_{23}, H_{24}\}$ and $S_2 = H_{13} + H_{14}$ for $t = 2$. Then

$$P(S = s) = \binom{d}{s} \sum_{j=0}^{d-s} (-1)^j \binom{d-s}{j} \theta_{s+j}, \tag{51}$$

where

$$\theta_t = \mathrm{P}(S_t = t) = \mathrm{P}(\text{ first } t \text{ elements of } H_{ik}\text{'s are equal to } 1 \text{ }) \quad \text{for } t = 1, 2, ...$$

The analysis above, leads to the new result presented below.

Lemma 5.2. *Random variable τ as defined in (43) is a sum of a Poisson random variable with parameter λ (50) and another discrete random variable with probability mass function as in (51).*

6 Kendal's Significance Test (KEST)

After exploring the new distribution properties of KEndal's τ, we propose a new high dimensional, robust covariance Significance Test statistic (KEST for short) which mimics the test ideas stated in Section 3.

Remaining working on the null hypothesis that the active set A contains supp($\boldsymbol{\beta}^*$), we treat the data as a triple $(y_i - y_j, \hat{y_i} - \hat{y_j}, \tilde{y_i} - \tilde{y_j})$ where y_i's are observed data and $\hat{y}_i$'s and $\tilde{y}_i$'s are predicted values of different estimators of $\boldsymbol{\beta}^*$ such that

$$\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}(\lambda_{k+1}), \quad \tilde{y}_i = \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}_A(\lambda_{k+1}). \tag{52}$$

where two estimators of $\boldsymbol{\beta^*}$ are defined as the following

$$\begin{aligned} \hat{\boldsymbol{\beta}}(\lambda_{k+1}) &= \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_{k+1}|\boldsymbol{\beta}|, \\ \tilde{\boldsymbol{\beta}}_A(\lambda_{k+1}) &= \arg\min_{\boldsymbol{\beta}_A} \frac{1}{2}\|\mathbf{Y} - \boldsymbol{X_A}\boldsymbol{\beta}_A\|^2 + \lambda_{k+1}|\boldsymbol{\beta}_A|. \end{aligned} \tag{53}$$

Additionally, the null hypothesis and the alternative of our covariance test statistic are

$$H_0 : supp(\boldsymbol{\beta}^*) \subseteq A, \qquad H_1 : supp(\boldsymbol{\beta}^*) \nsubseteq A.$$

Our new test statistic preserves the test idea in [2] to compare the correlation between observed vector $\mathbf{y}$ and predicted vectors $\hat{\mathbf{y}}$ and $\tilde{\mathbf{y}}$. We define our τ_k as

$$\tau_k = \frac{\tau(\mathbf{y}, \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda_{k+1})) - \tau(\mathbf{y}, \mathbf{X}_A\tilde{\boldsymbol{\beta}}_A(\lambda_{k+1}))}{\sigma^2}.$$

where some constant σ^2.

Using the Wilcoxon ranked sum test as a reference, we rewrite our test statistic τ_k.

$$\tau_k = \frac{1}{\sigma^2} \sum_{N_r} sgn(y_i - y_j) R_{ij},$$

where N_r is the number of pairs of i, j that satisfy the requirement of indices of Kendall τ correlation coefficient and

$$R_{ij} = sgn(\hat{y_i} - \hat{y_j}) - sgn(\tilde{y_i} - \tilde{y_j}). \tag{54}$$

We define R_{ij} in our test statistic to be the "sign difference of the predicted $\mathbf{y}$", which is similar to how R_i is defined in Wilcoxon ranked sum test.

Conjecture 6.1. *The τ-test statistic τ_k converges to a rescaled normal distribution under the null, when: 1) the errors have standard normal distribution; 2) $n \leq p$;*

We design a simulation according to the formula of τ_k. We draw $\mathbf{X}$ of size 1000×200 from multi-normal distribution with mean zero and variance σ^2. We choose the true values of beta to be 190 zeros and ones.

Figures 2 and 3 contain results of the simulation for different values of k according to the claim. We perform the same simulation for the original covariance test statistic with Pearson Correlation and compare two groups of graphs.

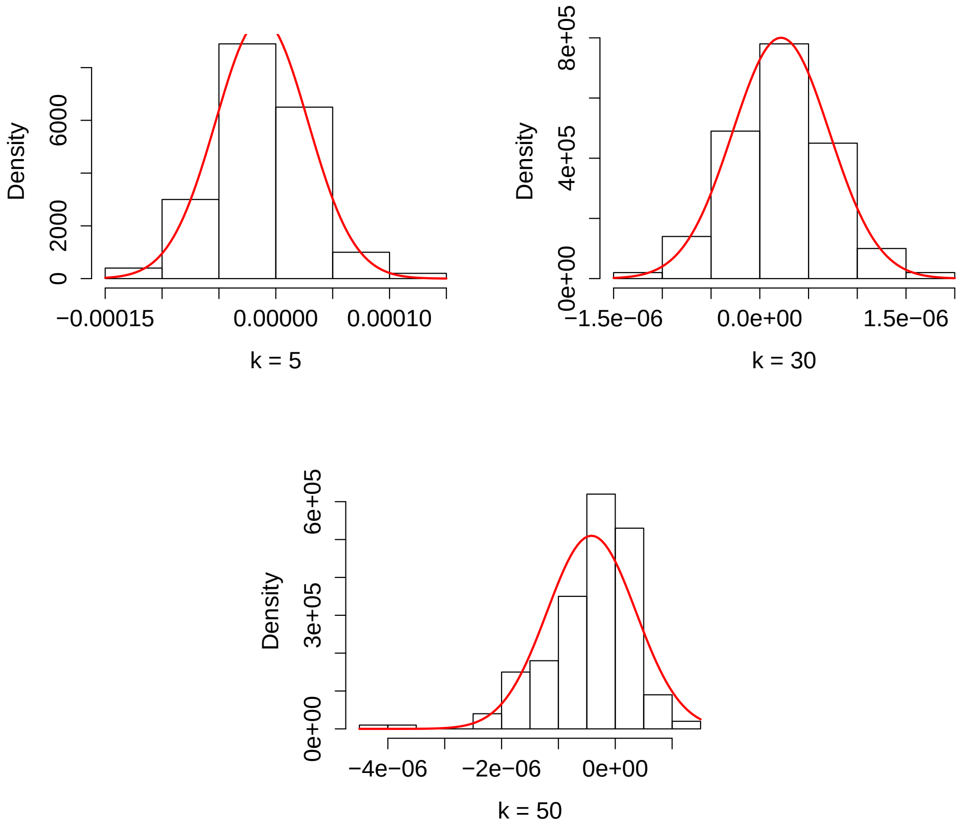

Figure 2: covariance test statistic using Pearson correlation

In the figures, as we intentionally select the values of k's, we noticed that, compared with Pearson correlation, the normal pattern of τ_k is more obvious. For $k = 5$, T_5 has less bins

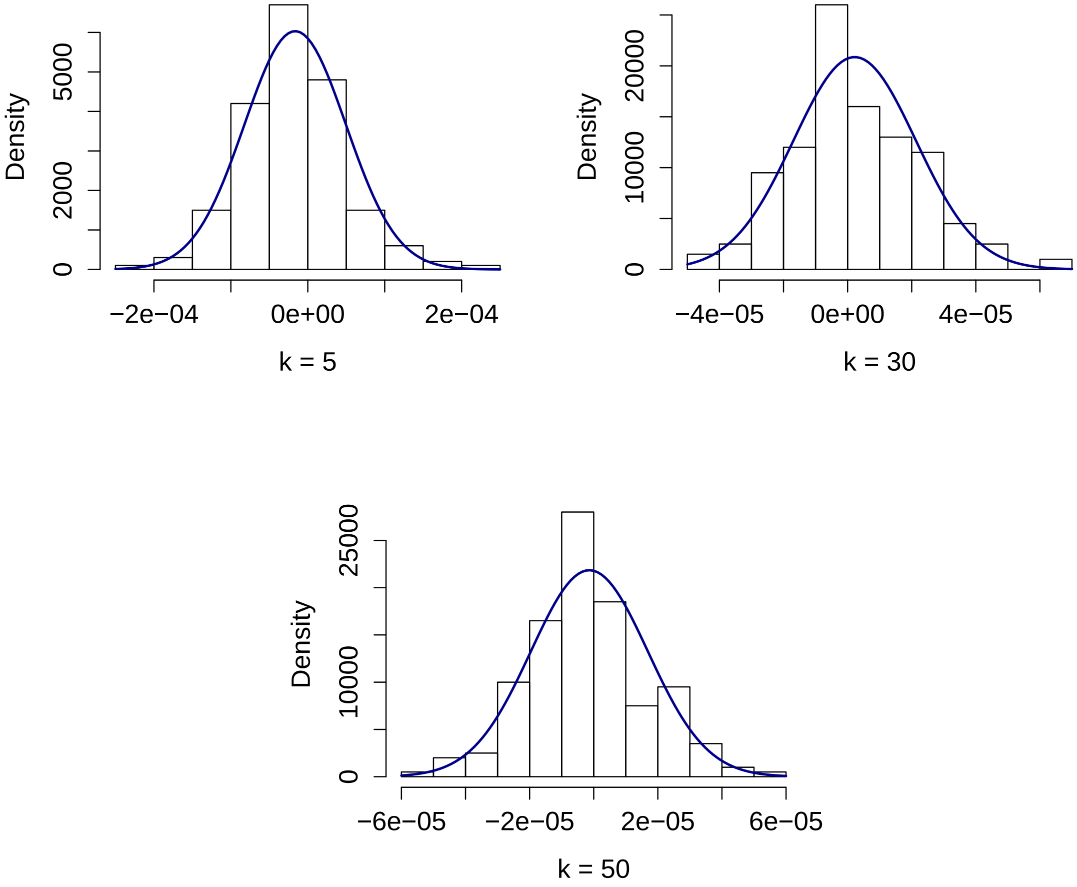

Figure 3: covariance test statistic using Kendall τ

compared to τ_5 and has less bell-shaped pattern. Similarly, for $k = 30$ and $k = 50$, we notice that the graphs of test statistic with Pearson correlation are less symmetric and right-skewed, while graphs of test statistic with Kendall τ are more symmetric respect to zero.

7 Discussion

In previous sections, we have shown how the LARS algorithm selects the set of active variables and how it fails to function properly under correlated setting. We also proved two useful lemmas stated in [2] with full details, which may help our readers to understand how the exponential distribution is derived. By offering a new LARS type algorithm, we showed that our new test statistic,τ_k, works better under limited simulation setting compared with the covariance test statistic of [2]. Up to this point, we succeeded in understanding the reason that the LARS algorithm fails and propose a solution to the problem. However, it is still significant to explore the robust approaches to the significance testing with model selection in mind. Additionally, the analysis of the newly proposed τ -LARS algorithm is still an open question. In Section 6, we conjecture about the connection between our new test statistic and the Wilcoxon ranked sum test. The connection could be used for proving the limiting distribution of our new test statistic under the null. Although in this paper, we have discussed the distribution properties of our proposed new test statistic, it is undeniably that there might exist many other forms of test statistic which worth exploring. For instance, we may consider the following comparison of the covariance test statistic

$$T_k = \frac{1}{2}\log\frac{1-\tau(\mathbf{y},\mathbf{X}\hat{\boldsymbol{\beta}}(\lambda_{k+1}))-\tau(\mathbf{y},\mathbf{X}_A\tilde{\boldsymbol{\beta}}_A(\lambda_{k+1}))}{1-\tau(\mathbf{y},\mathbf{X}\hat{\boldsymbol{\beta}}(\lambda_{k+1}))+\tau(\mathbf{y},\mathbf{X}_A\tilde{\boldsymbol{\beta}}_A(\lambda_{k+1}))},$$

which is inspired by the early work of Fisher exact test of independence in contingency tables.

References

[1] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao *Valid Post-Selection Inference* The Annals of Statistics, 2013

[2] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani *A significance test for the lasso.* The Annals of Statistics, 2014.

[3] Robert Tibshirani *Regression Shrinkage and Selectio via the Lasso.* Journal of the Royal Statistical Society. Series B, Volume 58, Issue 1, 1996.

[4] Bradley Efron, Trevor Hatie, Iain Johnstone, and Robert Tibshirani. *Least Angle Regression.* The Annals of Statistics, 2004.

[5] Keith Knight and Wenjiang Fu. *Asymptotics for Lasso-type Estimators.* The Annals of Statistics, 2000.

[6] Peng Zhao and Bin Yu. *On Model Selection Consistency of Lasso.* Journal of Machine Learning Research, 2007.

[7] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. *Restricted Eigenvalue Properties for Correlated Gaussian Designs.* Journal of Machine Learning Research, 2010.

[8] Laurens de Haan and Ana Ferreira *Extreme Value Theory: An Introduction.* Springer, 2006

[9] Christophe Croux and Catherine Dehon *Influence functions of the Spearman and Kendall correlation measures* Statistical Methods and Applications, 2010.

[10] J. Michael Steele *Le Cam's Inequality and Poisson Approximations* The American Mathematical Monthly, 1994

[11] Chang Yu and Daniel Zelterman *Sums of Dependent Bernoulli Random Variables and Disease Clustering* Statistics and Probability Letters, 2002