Simulation-based Significance Test for Lasso-Type Problems

A thesis submitted in partial satisfaction of the requirements for the degree
Master of Science in Statistics

by

Wenjia Wang

2014
Lasso has been shown to be effective in variable selection and sparse modeling. It can be applied to select a parsimonious set for the efficient prediction of a response variable. The goal of the thesis is to do significance test to know whether all truly active variables are contained in the current lasso model.

We design the following test statistics to do the test: $T_1 = \|\hat{\beta}\|_1$, $T_2 = \|\hat{\beta}\|_\infty$ and $T_3 =$ covariance test statistic. Simulation-based method, like direct sampling and importance sampling, are applied to draw samples and calculate the first two test statistics. The third statistic, covariance test statistic, is constructed based on lasso fitted values. Its null distribution is tractable and asymptotically Exp(1). Power curves of $T_1$ and $T_2$ are slightly different. To minimize type II error, $T_2$ performs better. Another test aims to test the significance of the predictor variable in the sequence of models visited along the lasso solution path. All of the three statistics are effective to select truly active variable; however, in terms of efficiency, $T_3$ is a better choice.
The thesis of Wenjia Wang is approved.

Ying Nian Wu

Frederic R. Paik Schoenberg

Qing Zhou, Committee Chair

University of California, Los Angeles

2014
To my beloved family
for their unconditional support and encouragement.
# Table of Contents

1 Introduction ................................................................................. 1

2 Methods .................................................................................... 3
   2.1 Basic concepts and properties of lasso ........................................... 3
   2.2 Covariance test statistic ........................................................... 4
   2.3 Simulation methods ................................................................. 5
      2.3.1 Direct sampling ............................................................... 5
      2.3.2 Importance sampling ....................................................... 5
      2.3.3 Simulation-based significance test ..................................... 6

3 Numerical examples ..................................................................... 8
   3.1 Simple case ($H_0 : \beta = 0$) ......................................................... 8
   3.2 Complicated case
      ($H_0$: All truly active predictors are contained in the current lasso model.)... 14

4 Conclusion and Discussion .......................................................... 19

References ..................................................................................... 21
List of Figures

3.1 Power curves for significance tests with dataset A ......................... 10
3.2 Power curves for significance tests with dataset B ......................... 11
3.3 Power curves for significance tests with dataset C ......................... 12
3.4 power-λ for dataset A ...................................................... 13
3.5 p-values for dataset D ...................................................... 15
3.6 p-values for dataset E ...................................................... 16
3.7 p-value for dataset F ...................................................... 16
3.8 p-values for dataset G ...................................................... 17
3.9 p-value for dataset H ...................................................... 17
List of Tables

3.1 Simulated datasets for $k = 1$ ................................................. 9
3.2 Estimation of p-values and powers for datasets A .................................. 9
3.3 Estimation of p-values and powers for datasets B .................................. 9
3.4 Simulated dataset C ............................................................... 12
3.5 Simulated datasets for one true active coefficient .................................. 15
3.6 Simulated datasets for $n = 100$, $p = 50$ ........................................ 17
Acknowledgments

First and foremost, I would like to thank my advisor, Professor Qing Zhou, for his supervision and guidance in my master thesis. I appreciate his patience in answering my seemingly endless questions and help providing me with recommendation letters. I learnt a lot from his invaluable advice and ideas. I feel lucky to have him as my master advisor.

I am grateful to all the professors, staff in our department. Thanks for leading me to the fascinating world of statistics. I also would like to thank my friends. I do have two wonderful years at UCLA.

Last but not least, I am grateful to my parents. Thanks for their unconditional support and love.
CHAPTER 1

Introduction

Consider the usual linear regression setup, for an outcome vector \( y \in \mathbb{R}^n \) and matrix of predictor variables \( X \in \mathbb{R}^{n \times p} \):

\[
y = X\beta^* + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)
\] (1.1)

where \( \beta^* \in \mathbb{R}^p \) are unknown coefficients to be estimated. Recently, L1-penalized estimation methods have been widely used to find sparse estimates of the coefficient vector. The lasso estimator\[7\] is defined as

\[
\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1
\] (1.2)

where \( \lambda \geq 0 \) is a tuning parameter, controlling the level of sparsity in \( \hat{\beta} \). Here we assume that the columns of \( X \) are in general position in order to ensure uniqueness\[11\] of the lasso solution.

There has been a considerable amount of recent work dedicated to the lasso problem, both in terms of computation and theory. To seek some sort of inferential guarantee for the computed lasso model, we need some simulation methods to investigate the null distribution of the test statistic (the null being that all truly active variables are contained in the current lasso model). We choose to apply the direct sampling approach to get statistical inference like p-values and power to see the performance of lasso estimators along the lasso solution path.

Yet, there are still major gaps in our understanding of the lasso as an estimation procedure, as the usual constructs like p-values are too small to estimate through simulation. As the joint distribution of a Lasso-type estimator \( \hat{\beta} \) and the subgradient \( S \) of \( \|\beta\|_1 \) evaluated
at $\hat{\beta}$ has a closed-form expression assuming a normal error distribution[1], importance sampling can be used to draw samples and accurately calculate a tail probability with respect to the sampling distribution under the null hypothesis.

In addition, we review some recent work, and find that Lockhart et al (2014) proposed a test statistic that has a simple and exact asymptotic null distribution, that is covariance test statistic[2]. We would like to know whether this test statistic has advantage over other statistics in variable selection.

The remaining part of the article is organized as follows. Chapter 2 reviews the basic properties of lasso, gives a brief view of the simulation method that we use, and how to get the statistical inference based on the method; and also gives the details of the covariance test statistic. In Chapter 3, we will give some numerical examples in both low-dimensional ($p < n$) and high-dimensional settings ($p > n$), to compare the performance of different statistics along the lasso solution path, mostly in terms of their effectiveness and efficiency. We conclude with a discussion in Chapter 4.
CHAPTER 2

Methods

2.1 Basic concepts and properties of lasso

Let \( W = \text{diag}(w_1, \ldots, w_p) \), where \( w_j \) is a positive weight, \( j = 1, \ldots, p \). The minimizer \( \hat{\beta} \) of (1.2) is given by the Karush-Kuhn-Tucker (KKT) condition

\[
\frac{1}{n} X^\top Y = \frac{1}{n} X^\top X \hat{\beta} + \lambda WS
\]

where \( S = (S_j)_{1:p} \) is the subgradient of the function \( g(\beta) = \|\beta\|_1 \) evaluated at the solution \( \hat{\beta} \). Therefore,

\[
\begin{cases}
S_j = \text{sgn}(\hat{\beta}_j), & \text{if } \hat{\beta}_j \neq 0, \\
S_j \in [-1, 1], & \text{if } \hat{\beta}_j = 0
\end{cases}
\]

for \( j = 1, \ldots, p \). Hereafter, we may simply call \( S \) the subgradient.

The vector \((\hat{\beta}, S)\) is called the augmented estimator in an L1-penalized regression problem, as the solution to Equation (2.1). In Zhou(2014), the joint distribution of \((\hat{\beta}, S)\) is derived with a density that can be calculated explicitly assuming a normal error distribution, regardless of the relative size between \( n \) and \( p \).

Let \( \mathcal{A} = \text{supp}(\hat{\beta}) \triangleq \{j : \hat{\beta}_j \neq 0\} \) be the active set of \( \hat{\beta} \) and \( \mathcal{I} = \{1, \ldots, p\} \setminus \mathcal{A} \) be the inactive set, i.e., the set of the zero components of \( \hat{\beta} \). The vector \((\hat{\beta}, S)\) can be equivalently represented by the triple \((\hat{\beta}_\mathcal{A}, S_{\mathcal{I}}, \mathcal{A})\). They are equivalent because from \((\hat{\beta}_\mathcal{A}, S_{\mathcal{I}}, \mathcal{A})\) one can unambiguously recover \((\hat{\beta}, S)\), by setting \( \hat{\beta}_{\mathcal{I}} = 0 \) and \( S_{\mathcal{A}} = \text{sgn}(\hat{\beta}_\mathcal{A}) \), and vice versa.

Under a normal error distribution, we can derive a closed-form density of the joint distribution[1] of \((\hat{\beta}_\mathcal{A}, S_{\mathcal{I}}, \mathcal{A})\). (See in Zhou(2014) for proof.) The density of interest is written as \( \pi_r(b_{\mathcal{A}}, s_{\mathcal{I}}, \mathcal{A}) \).
2.2 Covariance test statistic

At a given step in the lasso path, we consider testing the significance of the variable that enters the active set. We constructed the test statistic, called covariance test statistic, from the lasso solution path, i.e., the solution $\hat{\beta}(\lambda)$ in (1.2) a function of the tuning parameter $\lambda \in [0, \infty)$. The lasso path can be computed by the well-known LARS algorithm of Efron et al.(2004)[3], which traces out the solution as $\lambda$ decreases from $\infty$ to 0. We assume that the columns of $X$ are in general position, implying that there is a unique lasso solution at each $\lambda > 0$ and hence a unique path[11].

Before giving the definition of the statistic, some properties of the lasso path need to be addressed[2]:

(1) The path $\hat{\beta}(\lambda)$ is a continuous and piecewise linear function of $\lambda$, with knots (changes in slope) at values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$ (those knots depend on $y, X$).

(2) At $\lambda = \infty$, the solution $\hat{\beta}(\infty)$ has no active variables (i.e., all variables have zero coefficients); for decreasing $\lambda$, each knot $\lambda_k$ marks the entry or removal of some variable from the current active set (i.e., its coefficient becomes nonzero or zero, respectively). Therefore the active set, and also the signs of active coefficients, remain constant in between knots.

(3) For a matrix $X$ satisfying the positive cone condition (a restrictive condition that covers, e.g., orthogonal matrices), there are no variables removed from the active set as $\lambda$ decreases, and therefore the number of knots is $\min\{n, p\}$. Let $A$ be the active set just before $\lambda_k$, and suppose that predictor $j$ enters at $\lambda_k$. Denote by $\hat{\beta}(\lambda_{k+1})$ the solution at the next knot in the path $\lambda_{k+1}$, using predictors $A \cup \{j\}$. Finally, let $\hat{\beta}_A(\lambda_{k+1})$ be the solution of the lasso problem using only the active predictors $X_A$, at $\lambda = \lambda_{k+1}$, i.e.

$$\hat{\beta}_A(\lambda_{k+1}) = \arg\min_{\beta_A \in \mathbb{R}^{|A|}} \frac{1}{2} \|y - X_A \beta_A\|^2_2 + \lambda_{k+1} \|\beta_A\|_1$$

The covariance test statistic is defined as

$$T_k = \left( \langle y, X \hat{\beta}(\lambda_{k+1}) \rangle - \langle y, X_A \hat{\beta}_A(\lambda_{k+1}) \rangle \right)/\sigma^2$$

(2.3)

We can show that under the null hypothesis that the current lasso model contains all truly
active variables, $A \supseteq \text{supp}(\beta^*)$,

$$T_k \xrightarrow{d} \text{Exp}(1),$$

i.e., $T_k$ is asymptotically distributed as a standard exponential random variable, given reasonable assumption on $X$ and the magnitudes of the nonzero true coefficients[2]. See more details in Lockhart et al. (2014).

### 2.3 Simulation methods

As the density $\pi(b_A, s_I, A)$ has a closed-form expression given $\beta$ and $\sigma^2$, we can apply importance sampling to calculate expectations (so that to accurately estimate a tail probability) with respect to the distribution $\pi(b_A, s_I, A)$ after drawing proposals via a direct sampling approach.

#### 2.3.1 Direct sampling

**Routine 1** (Direct sampler). Assume the error distribution is $\mathcal{N}_n(0, \sigma^2 I_n)$. For $t = 1, \ldots, N$,

1. Draw $e(t) \sim \mathcal{N}_n(0, \sigma^2 I_n)$ and set $Y(t) = X\beta + e(t)$;
2. Find the minimizer $\hat{\beta}(t)$ of (1.2) with $Y(t)$ in place of $Y$;
3. If needed, calculate the subgradient vector $S(t) = \left((n\lambda W)^{-1}X^t(Y(t) - X\hat{\beta}(t))\right)$.

This approach directly draws $Y(t)$ from its sampling distribution and requires a numerical optimization algorithm in step (2) for each sample.

Yet estimating a tail probability (small p-value) will be extremely difficult using direct sampling, therefore we propose importance sampling (IS) to solve the problem.

#### 2.3.2 Importance sampling

Importance sampling[4] is a more efficient approach to simulation. In essence, we take draws from an alternative distribution whose support is concentrated in the truncation region.
Principle of importance sampling:
\[
\int_{F} sf(s) \, ds = \int_{G} s \frac{f(s)}{g(s)} g(s) \, ds
\]  
That is, sampling \( s \) from \( f(s) \) distribution equivalent to sampling \( s \ast w(s) \) from \( g(s) \) distribution, with importance sampling weight \( w(s) \equiv \frac{f(s)}{g(s)} \). (\( f \) and \( g \) should have the same support.)

Importance sampling is therefore of considerable interest since it puts very little restriction on the choice of the instrumental distribution \( g \), which can be chosen from distributions that are either easy to simulate or efficient in the approximation of integral.

2.3.3 Simulation-based significance test

As introduced before, for a fixed \( X \), the density \( \pi_r(b_A, s_I, A) \) has a closed-form expression. Suppose we are given a Lasso-type estimate \( \hat{\beta}^* \) for an observed dataset with a penalty parameter \( \lambda^* \). The hypothesis test is as followed:

Null hypothesis \( (H_0) \): \( \beta = \beta_0 \)

Alternative hypothesis \( (H_1) \): \( \beta = \beta_1 \)

Under such model, we want to calculate the p-value and power of some test statistic \( T(\hat{\beta}) \in \mathbb{R} \) constructed from the Lasso-type estimator \( \hat{\beta} \) for \( \lambda = \lambda^* \). Precisely, the desired p-value is

\[
p - value = P(\left| T(\hat{\beta}) \right| \geq T^*; H_0, \lambda^*) = \int_{\Omega^*_{r}} \pi_r(b_A, s_I, A; \beta_0, \sigma^2_0, \lambda^*) \xi_n (db_A d\hat{s}); \quad (2.5)
\]

similarly, power is calculated by

\[
\text{power} = P(\left| T(\hat{\beta}) \right| \geq T^*; H_1, \lambda^*) = \int_{\Omega^*_{r}} \pi_r(b_A, s_I, A; \beta_1, \sigma^2_1, \lambda^*) \xi_n (db_A d\hat{s}); \quad (2.6)
\]

where \( T^* = |T(\hat{\beta})| \), \( \Omega^*_{r} = \{(b_A, s_I, A) \in \Omega_r : |T(b)| \geq T^* \} \), and \( \xi_n \) denotes \( n \)-dimensional Lebesgue measure. For power calculation, we can directly sample from \( \pi_r(b_A, s_I, A; \beta_1, \sigma^2_0, \lambda^*) \).

Yet estimating p-value like this is extremely difficult as it is very small. Therefore, we use importance sampling to calculate p-value.

Our target distribution is \( \pi_r(\bullet; \beta_0, \sigma^2_0, \lambda^*) \) and we propose to use \( \pi_r(\bullet; \beta_0, (\sigma^2)^{1}, \lambda^1) \) as a trial distribution to estimate expectations with respect to the target distribution via IS.
As the trial and target distribution have the same support and the importance weight for a sample \( \pi_r(b_A, s_I, A) \) from the trial distribution can be calculated efficiently, we can sample from the trial distribution and allocate the sample with different weights to calculate the test statistic \( T(b^{(t)}) \). The importance weight is denoted by \( w(b_A, s_I, A; \sigma_0^2, \lambda^*) \). See the proof of IS weights calculation in Zhou(2014).

**Routine 2** (IS estimation)[1]. By direct sampling, draw \((b_A, s_I, A)^{(t)}\), for \( t = 1, \ldots, N \), from the trial distribution \( \pi_r(\cdot; \beta_0, (\sigma^2)^\dagger, \lambda^\dagger) \). Then the IS estimate for the p-value is given by

\[
\hat{p}_{val}^{(IS)} = \frac{\sum_{t=1}^{N} w((b_A, s_I, A)^{(t)}; \sigma_0^2, \lambda^*) \mathbb{1}(|T(b^{(t)})| \geq T^*)}{\sum_{t=1}^{N} w((b_A, s_I, A)^{(t)}; \sigma_0^2, \lambda^*)}
\]

(2.7)
CHAPTER 3

Numerical examples

The goal of the significance test is to test whether all truly active variables are contained in the current lasso model. We apply different test statistic to the test, like L1-norm of \( \hat{\beta} \), covariance test statistic, etc. Comparing the results of these statistic, we would like to know which one works better under certain settings.

For the sake of understanding, we first focus on the special case, which assumes the active set is \( \emptyset \). Following this, we study the more complicated case, which assumes active set in the lasso step along the solution path is the true active set in true model.

3.1 Simple case \( (H_0 : \beta = 0) \)

The null hypothesis is \( H_0 : \beta = 0 \) and \( \sigma^2 = \sigma_0^2 \). We alter the value of \( \lambda^* \), i.e. the target distribution \( \pi_r(\bullet; \beta, \sigma_0^2, \lambda^*) \) and observed statistic \( T^* \), to see along the path, which knot’s test statistic works better for the hypothesis test. We designed the following test statistics, \( T_1^{(j)} = \|\hat{\beta}^{(j)}\|_1 \) and \( T_2^{(j)} = \|\hat{\beta}^{(j)}\|_\infty \), to calculate the p-values of the \( j \)th knot’s test statistic under the null hypothesis and powers of the statistic under the alternative hypothesis (alternative hypothesis being that \( \beta = \beta_0 \)).

We simulated two high-dimensional cases \( (p > n) \) to calculate p-value and power of the test. The predictors \( X \) were generated from \( N_p(0, \Sigma_X) \), where the diagonal and the off-diagonal elements of \( \Sigma_X \) are 1 and 0.05, respectively. Given the predictors \( X \), the response vector \( Y \) was drawn from \( N_n(X\beta_0, \sigma_0^2 I_n) \). Table 1 gives the values of \( n, p, \sigma_0^2 \), and \( \beta_0 \) for the two datasets.

By IS to calculate p-values for the test statistic, we need to set \( (\sigma^2)^\dagger \) and \( \lambda^\dagger \) for the trial
distribution so that there is a substantial fraction of samples for which \(|T(b^{(t)})| \geq T^*\). We chose \((\sigma^2)^\dagger = 5\sigma_0^2\) and used the following Routine 3 to choose \(\lambda^\dagger\). See more details in Zhou (2013). The values of \(\lambda^\dagger\) for the two datasets are also given in Table 3.1.

| Table 3.1: Simulated datasets for \(k = 1\) |
|-----------------|-----|--------|-----------------|-----|
| Dataset | n  | p    | \(\sigma_0^2\) | \(\beta_0\) | \(\lambda^\dagger\) |
| A     | 5  | 10   | 1/4             | (2, −2, 0, . . ., 0) | 0.622 |
| B     | 5  | 10   | 1/4             | (1/4, . . ., 1/4)  | 0.625 |

**Routine 3** (Set \(\lambda^\dagger\) given \((\sigma^2)^\dagger\)). Draw \(Y^{(t)}\) from \(N_n(0, (\sigma^2)^\dagger I_n)\) and calculate \(\lambda^{(t)} = \frac{1}{n} \|X^{\dagger} Y^{(t)}\|_{\infty}\), for \(t = 1, \ldots, 100\). Then set \(\lambda^\dagger\) to the first quartile of \(\{\lambda^{(t)} : t = 1, \ldots, 100\}\).

The IS method (Routine 2) was applied with \(N = 5,000\) to estimate p-values, and the DS method (Routine 1) was also applied with \(N = 5,000\) to estimate powers for all the above tests. Table 3.2 and Table 3.3 summarizes the results for the dataset A and B.

<table>
<thead>
<tr>
<th>Step</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\lambda^*)</td>
<td>2.41</td>
<td>1.88</td>
<td>0.891</td>
<td>0.860</td>
<td>0.143</td>
<td>0</td>
</tr>
<tr>
<td>Pval(T1)</td>
<td>3.45 \times 10^{-10}</td>
<td>1.07 \times 10^{-7}</td>
<td>5.81 \times 10^{-30}</td>
<td>8.28 \times 10^{-32}</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pval(T2)</td>
<td>9.81 \times 10^{-5}</td>
<td>8.20 \times 10^{-8}</td>
<td>2.06 \times 10^{-17}</td>
<td>6.09 \times 10^{-19}</td>
<td>2.71 \times 10^{-34}</td>
<td>5.82 \times 10^{-27}</td>
</tr>
<tr>
<td>Power(T1)</td>
<td>0.794</td>
<td>0.852</td>
<td>0.809</td>
<td>0.714</td>
<td>0.723</td>
<td>0.725</td>
</tr>
<tr>
<td>Power(T2)</td>
<td>0.794</td>
<td>0.846</td>
<td>0.893</td>
<td>0.849</td>
<td>0.826</td>
<td>0.956</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\lambda^*)</td>
<td>1.31</td>
<td>0.944</td>
<td>0.323</td>
<td>0.291</td>
<td>0.0796</td>
<td>0</td>
</tr>
<tr>
<td>Pval(T1)</td>
<td>0.743</td>
<td>0.329</td>
<td>1.19 \times 10^{-3}</td>
<td>7.27 \times 10^{-4}</td>
<td>1.16 \times 10^{-6}</td>
<td>1.20 \times 10^{-9}</td>
</tr>
<tr>
<td>Pval(T2)</td>
<td>0.743</td>
<td>0.263</td>
<td>1.12 \times 10^{-2}</td>
<td>8.56 \times 10^{-3}</td>
<td>4.91 \times 10^{-3}</td>
<td>2.34 \times 10^{-4}</td>
</tr>
<tr>
<td>Power(T1)</td>
<td>0.989</td>
<td>0.989</td>
<td>0.994</td>
<td>0.969</td>
<td>0.994</td>
<td>0.996</td>
</tr>
<tr>
<td>Power(T2)</td>
<td>0.989</td>
<td>0.989</td>
<td>0.995</td>
<td>0.987</td>
<td>0.991</td>
<td>0.995</td>
</tr>
</tbody>
</table>
In dataset A, there are two large coefficients, which represents the case that the true model is sparse; while in dataset B, it mimics the scenario in which the true model has many relatively small coefficients. We analyze these two cases separately.

For dataset A, all the test statistics have tail probability (small p-value), ranging from $10^{-5}$ to $10^{-34}$ or even smaller. Thus we have sufficient evidence to reject the null hypothesis ($\beta = 0$), which is the same as we expected. To evaluate whether it is a good test or not, we also want to minimize its type II error (i.e, maximize the power of the test). Figure 3.1 shows the trend of powers with different $\lambda^*$.

![Figure 3.1: Power curves for significance tests with dataset A](image)

It seems that mostly $T_2$ has greater power than $T_1$. Then we investigate the power of the statistics of different knots along the lasso path. As $\lambda^*$ decreases, the power of $T_2$ first increases, then decreases, and the maximum power of $T_1$ and $T_2$ both occur at Step 3. Step 3 is the step that the correct number of active coefficients enter. Thus, we may infer that...
for the sparse model like dataset A, we can choose \( T = \|\hat{\beta}\|_\infty \) as the test statistic, and \( \lambda^* \) as the first \( \lambda \) along the solution path such that the Lasso estimate \( \hat{\beta}^* \) gives the correct number of active coefficients.

For dataset B, p-values for the first two steps are not significant, while those for other steps range from \( 10^{-2} \) to \( 10^{-9} \). Therefore, there is no need to investigate the test for the first two steps any more. In terms of powers, Figure 3.2 shows its trend with different \( \lambda^* \).

![Figure 3.2: Power curves for significance tests with dataset B](image)

Eliminating the first two steps, it seems that \( T_2 \) has greater power than \( T_1 \). So next we investigate the power of \( T_2 \) of different knots along the lasso solution path. There is no big difference or trend among these knots, and the powers stay over 0.9 steadily. Therefore, we may infer that for the true model which has many relatively small coefficients like dataset B, we can choose \( T = \|\hat{\beta}\|_\infty \) as the test statistic, and under the assumption of controlling p-value less than significance level, the choice of \( \lambda^* \) has little effect on the test.
Summarizing the results from the above two datasets, it appears that $T_2$ performs better than $T_1$, and for convenience, we can choose $\lambda^*$ as the first $\lambda$ that gives true active coefficients. To verify the results, we propose another dataset C, which is given in Table 3.4.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>n</th>
<th>p</th>
<th>$\sigma_0^2$</th>
<th>$\beta_0$</th>
<th>$\lambda^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>10</td>
<td>20</td>
<td>1/4</td>
<td>(1, 1, -1, -1, 0, . . . , 0)</td>
<td>0.572</td>
</tr>
</tbody>
</table>

We applied the Lars algorithm to calculate the lasso solution path, and found that the 5th step gave the correct active coefficients. According to the results concluded above, we can choose $T = T_2 = \|\hat{\beta}\|_\infty$ and 5th $\lambda$ along the path as $\lambda^*$. Figure 3.3 shows the power curve for the test. On the curve, we can see that $T_2$ in the 5th step has the largest power among all the statistic, which corresponds to the test statistic we proposed.

![Figure 3.3: Power curves for significance tests with dataset C](image-url)
Remark 1. For convenience, we choose $\lambda^*$ as the $\lambda$ at knots along the solution path. We have no idea at some arbitrary value of the tuning parameter $\lambda$, what the p-value and power is, whether test statistic determined by such $\lambda$ perform even better. For example, read Figure 3.4 directly, it is hard to tell the relationship between $\lambda$ and the power.

Remark 2. When doing these simulation-based tests, mostly we focus on the power curves of the significance tests, do summarizations and make some inference. Yet the reason why we choose statistic like this is still not clear, and it requires more work to do in the future.
3.2 Complicated case

\((H_0):\text{ All truly active predictors are contained in the current lasso model.}\)

Suppose exactly \(k_0\) components of the true coefficient vector are nonzero, and consider testing the entry of the predictor at step \(k = k_0' + 1\) (\(k_0'\) is the step at which all true active coefficients are included). Let \(A = \text{supp}(\beta)\) denote the true active set (so \(k_0 = |A|\)), and let \(B\) denote the event that all truly active variables are contained at step \((k-1)\).

We design the following test statistics, \(T_1^{(k)} = \|\hat{\beta}^{(k)}\|_1\), \(T_2^{(k)} = \|\hat{\beta}^{(k)}\|_\infty\) and \(T_3^{(k)} = \) covariance test statistic at step \(k\). Under the null hypothesis (i.e., conditional on \(B\)), we would like to see which test statistic works better in the test. To make comparison, we calculated p-values of the three statistics along the path in different models.

For \(T_1\) and \(T_2\), the p-value estimation procedure is as follows. With a fixed \(X\), given \(A^{(k)}\), \(\lambda^{(k)}\) and \(\hat{\beta}^*\) for an observed dataset of the lasso model,

1. estimate least square estimators \(\hat{\beta}^{LS^{(k)}}\) and \(\hat{\sigma}^2^{(k)}\) for the linear model \(Y = X_{A^{(k)}}\beta + \epsilon\);
2. given \(\beta = \hat{\beta}^{LS^{(k)}}\) and \(\sigma^2 = \hat{\sigma}^2^{(k)}\), apply direct sampling (Routine 1) to draw samples, and get \(\hat{\beta}^{lasso}(\lambda^{(k)})\);
3. the p-value estimator is calculated by \(p^{(k)} = P\left(T(\hat{\beta}^{lasso}(\lambda^{(k)})) > T^* \mid H_0\right)\).

For \(T_3\), the covariance test statistic, according to Lockhart et al.(2014), we know that the distribution of this statistic is asymptotically \(\text{Exp}(1)\), under the null hypothesis that all truly active predictors are contained in the current active set. We can directly apply the R package covTest to compute p-values for the covariance test.

We simulated several datasets as numerical examples. The predictors \(X\) were generated from \(N_p(0, \sum_X)\), where the diagonal and the off-diagonal elements of \(\sum_X\) are 1 and 0.05, respectively. Given the predictors \(X\), the response vector \(Y\) was drawn from \(N_n(X\beta_0, \sigma_0^2 I_n)\).
First we set the first coefficient of the true coefficient vector equal to 2, and the rest zero. Table 3.5 gives the values of $n$, $p$, $\sigma_0^2$ and $\beta_0$ for the datasets.

Table 3.5: Simulated datasets for one true active coefficient

<table>
<thead>
<tr>
<th>Dataset</th>
<th>n</th>
<th>p</th>
<th>$\sigma_0^2$</th>
<th>$\beta_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>100</td>
<td>10</td>
<td>1/4</td>
<td>(2, 0, \ldots, 0)</td>
</tr>
<tr>
<td>E</td>
<td>100</td>
<td>50</td>
<td>1/4</td>
<td>(2, 0, \ldots, 0)</td>
</tr>
<tr>
<td>F</td>
<td>10</td>
<td>20</td>
<td>1/4</td>
<td>(2, 0, \ldots, 0)</td>
</tr>
</tbody>
</table>

Figure 3.5: p-values for dataset D

Figure 3.5 shows the p-value curve for the three statistics for dataset D. The red line is the significance level of the test, 0.05. Checking the lasso path, 1st predictor enters at the 2nd step. As the true coefficient vector has only one non-zero element, it is reasonable to reject the null model in the 1st step, and then accept the null. This is why p-value is small at
the 1st step, and increases greatly in the next step. This example implicitly assumes that one might stop entering variables into the model when the computed p-value rises above some threshold. In addition, we find a growth trend for p-value of all three statistics. Usually as tuning parameter \( \lambda \) decreases and more predictors enter the model, the probability that null model contains all truly active models increases. That’s why p-value has a growth trend.

Figure 3.6 shows the p-value curve for dataset E \((n = 100, p = 50)\). The curves seem similar with that in dataset D. Figure 3.7 shows the p-value curve for dataset F \((n = 10, p = 20)\). This is a high-dimensional case. As it is difficult to estimate error variance, we assume it to be 0.25. P-value increases sharply at step 2, indicating there is only one active coefficient, which corresponds to the true model. Yet, compared to the power curves in Figure 3.5 and Figure 3.6, though the increasing trend exists, p-value of all three statistics oscillate up and down. Besides, though insignificant after step 2, p-value is relatively small for \( T_2 \). The performance of the test statistic in high-dimensional setting \((p > n)\) seems not as good as that in low-dimensional setting \((p < n)\). We infer it may caused by the estimation of error variance.

Figure 3.6: p-values for dataset E

Figure 3.7: p-value for dataset F

**Remark 1.** Estimation of error variance \( \sigma^2 \). In practice, the error variance is typically unknown, thus we have to make an estimation. For the low-dimensional cases \((p < n)\), we can easily estimate it and proceed by analogy to standard linear model theory, i.e.,
\[ \hat{\sigma}^2 = \| Y - X \hat{\beta}_{LS} \|_2^2 / (n - p). \] Yet, for high-dimensional cases \((p \geq n)\), estimation of \(\sigma^2\) is not nearly as straightforward. In this thesis, in high-dimensional cases, we assume \(\sigma^2 = 0.25\) for convenience.

Next, we fix \(n = 100\) and \(p = 50\). We set the first \(k\) coefficients of the true coefficient vector equal to 2, and the rest zero, for \(k = 5, 10\). The setup is shown in Table 3.6.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>(n)</th>
<th>(p)</th>
<th>(\sigma^2_0)</th>
<th>(\beta_0)</th>
<th>number of true active coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>100</td>
<td>50</td>
<td>1/4</td>
<td>((2, 2, \ldots, 0))</td>
<td>5</td>
</tr>
<tr>
<td>H</td>
<td>100</td>
<td>50</td>
<td>1/4</td>
<td>((2, 2, \ldots, 0))</td>
<td>10</td>
</tr>
</tbody>
</table>

Figure 3.8: p-values for dataset G

Figure 3.9: p-value for dataset H

Remark 2. For the covariance test statistic, we only get p-values in which a predictor enters the model or leaves the model or occurs simultaneously (i.e., for the lasso step in which there is predictor neither entering nor leaving, covariance test statistic cannot give such step p-value). Thus, for convenience of comparison, we choose the lasso path in which there is predictor either entering or leaving for every step along the solution path.

Figure 3.8 and Figure 3.9 show the p-value curves with number of truly active coefficients \(k\) be 5 and 10. As expected, p-value increases sharply at step \(k+1\), indicating that the lasso
model at step $k+1$ includes all truly active coefficients. For the following steps after step $k+1$, all the p-value curves show growth trend. Thus, in terms of effective solution of truly active coefficients, all the three test statistics work well. However, it takes much more time to calculate $T_1$ and $T_2$ than $T_3$. To get p-values of $T_1$ and $T_2$, samples have to be drawn via direct sampling, and it requires a numerical optimization (i.e., minimizing (1.2)) for every iteration. In contrast, we know that covariance test statistic asymptotically distributed as Exp(1), so there is no need to iterate many times to calculate its p-value. Thus, in terms of efficiency, we should choose covariance test statistic.
Lasso is effective in variable selection and sparse modeling. When using lasso as an estimation procedure, how to get statistical inference for the lasso seems more and more important. Therefore, we propose two significance tests to investigate the performance of lasso estimators which are at the knots along solution path. One is to test the null being that all the coefficients are zero, while alternative being coefficients are the same as the true model. In this setting, to minimize type II error, we choose tuning parameter as the first $\lambda$ that gives true active coefficients, and test statistic $T = ||\hat{\beta}||_\infty$. The other is to test the null being that all truly active variables are contained in the current lasso model. Covariance test statistic has a simple asymptotic distribution, Exp(1), and by using it, there is no need to resample. Thus, by using covariance test statistic to do significance test, we can infer the active predictors both effectively and efficiently.

The problem of assessing significance in an adaptive linear model fit by the lasso is a difficult and complicated one. Though much effort has been made to compare the results from different test statistics, there are still many problems to be solved:

1. In high-dimensional cases ($p > n$), how to estimate the error variance $\sigma^2$.
2. For a real dataset, if we would like to include an intercept in the model, we should run the lasso on centered data (i.e., centering $Y$ and column centering $X$)[8]. Then it would create a weak dependence between the components of the error vector. In this case, what effect will have on the test statistics and significance tests.
3. By design, our test statistic $T(\hat{\beta}^{(j)})$, where $\hat{\beta}^{(j)}$ is the lasso estimator of $j$th step along the solution path (i.e., tuning parameter is chosen as the $\lambda$ at knots along the solution path). If the tuning parameter in the lasso model is some arbitrary value, what about the performance
of the test statistics?
References


