이 학 석 사 학 위 논 문

Variable Selection Methods in High-Dimensional Regression Analysis

고차원 회귀분석에서의 변수선택 방법론

FEBRUARY 2023

서울대학교 대학원
통계학과
신 도 협

Variable Selection Methods in High-Dimensional Regression Analysis

지도 교수 PARK JUN YONG

이 논문을 이학석사 학위논문으로 제출함

2022년 10월

서울대학교 대학원

통계학과

신 도 협

신도협의 이학석사 학위 논문을 인준함

2022년 12월

위 원 장: 이 상 열 **(인)**

부위원장: PARK JUN YONG **(인)**

위 원: 임 요 한 **(인)**

Abstract

Variable Selection Methods in High-Dimensional Regression Analysis

Dohyup Shin
The Department of Statistics
The Graduate School
Seoul National University

High-dimensional data analysis is attracting attention in many fields these days. In particular, it is a difficult and important problem to select a variable that has a significant effect among numerous variables. Several statistical methods exist to solve this problem, such as multiple testing and LASSO in linear regression models. In this paper, we introduce the case of Lasso, adaptive Lasso, Elastic net, and generalized linear models in Bühlmann and Van De Geer (2011) [3]. Also, we review and cover the multiple testing procedures and introduce a recent method of false discovery rate(FDR) control via data splitting proposed by Dai et al.(2022)[4]. Finally, if relevant variables are sparse, we check whether the adaptive Lasso estimator gives better results than the Lasso estimator through simulation. In addition, we confirm that the MDS method is more stable and has higher empirical power than the DS method by simulation.

keywords: Variable Selection, Lasso, Adaptive Lasso, Elastic net, Generalized Lasso, False Discovery Rate, BHq procedure, Mirror Statistic, Data Splitting, Multiple Data Splitting
student number: 2021-28605

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

Let $\mathbf{X}_{n\times p} = (X_1, X_2, \ldots, X_p)$ be the explanatory features with $p \gg n$. For each feature has been normalized with zero mean and unit variance. Let $Y = (y_1, \ldots, y_n)$ be the vector of n independent response variable. Consider the linear regression model

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \boldsymbol{\epsilon} \tag{1.1}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ is a noise vector with $\epsilon_i \overset{iid}{\sim} (0, \sigma^2)$. Let $S_0 = \{i : \beta_i = 0\}$ be the index set of the null features and $S_1 = \{i : \beta_i \neq 0\}$ be the index set of the relevant features.

Various methods for high-dimensional linear regression have been developed. It is important to select the relevant variables. The classical methods of variable selection include forward selection, backward selection, stepwise selection, Akaike Information Criterion(AIC) (Akaike (1998))[1], Bayesian Information criterion(BIC) (Schwarz (1978)) [6], etc. Furthermore, there is another method by using a penalty term. The most widely known method is the Lasso regression proposed by Tibshirani (1996) [7]. Alternatively, there is a method of selecting significant variables by controlling the false discovery rate. Benjamini and Hochberg (1995) [2] procedure first proposed a method of using FDR control. After that, many methods of controlling

FDR in various assumptions were announced. But, the BHq procedure is difficult to apply in a high-dimensional model. This is because the BHq procedure requires the calculation of p-values, which are difficult to compute in high-dimensional models. To solve the above problem, we introduce the false discovery rate control via data splitting, which is proposed by Dai et al. (2022) [4].

In chapter 2, we introduce the theory and good properties of LASSO regression. In addition, we will cover the adaptive Lasso estimator and Elastic net estimator, which are more advanced models, and introduce LASSO estimators in generalized linear models.

In chapter 3, we briefly discuss multiple testing and the concept of False Discovery Rate(FDR). And we introduce the variable selection method through FDR control in the linear regression model and deal with FDR control via data splitting recently announced by Dai et al. (2022) [4].

In chapter 4, we experimentally show that the Adaptive Lasso estimator performs better in variable selection than LASSO in sparse linear models. Next, we show through experiments that the MDS method is more stable and has higher power than DS in FDR control by using data splitting.

Chapter 2

Lasso Regression

2.1 The Lasso estimator

If $p > n$, the least square estimator of β is not unique and greatly overfits the data. Therefore, we use the l_1 penalty method among the regularization methods. The Lasso estimator is defined as

$$\hat{\beta}(\lambda) = \underset{\beta}{\text{argmin}} \left(||Y - X_{n\times p}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \right) \tag{2.1}$$

where $\lambda \geq 0$ is the penalty parameter.

In general, we can select the proper λ by using a cross-validation procedure. The estimator has the property that it does variable selection in the sense that $\beta_j(\lambda) = 0$ for some j's (depending on the choice of λ) and $\hat{\beta}_j(\lambda)$ can be thought of as a shrunken least squares estimator. Since the optimization in (2.1) is convex, (2.1) is equivalent to

$$\hat{\beta}_{\text{Primal}}(R) = \underset{\beta:||\beta||_1 \leq R}{\text{argmin}} \left(||Y - X_{n\times p}\boldsymbol{\beta}||_2^2 / n \right) \tag{2.2}$$

with 1-1 correspondence between λ and R, depending on the data.

In Bühlmann and Van De Geer (2011) [3], there are many properties for the Lasso estimator. Under several conditions, the lasso estimator has variable screening and

variable selection properties. The variable screening property is

$$P[\hat{S}(\lambda) \supseteq S_0] \to 1 \ (p \geq n \to \infty)$$

which means that all relevant variables are included. Also, the variable selection property is

$$P[\hat{S}(\lambda) = S_0] \to 1 \ (p \geq n \to \infty)$$

which means that all relevant are chosen exactly and where $\hat{S}(\lambda) = \{j : \hat{\beta}_j(\lambda) \neq 0, j = 1, 2, \ldots, p\}$. The proof of the above properties and other properties are detailed in Bühlmann and Van De Geer (2011) [3].

2.2 Adaptive Lasso

In Zou, H. (2006) [8], the Adaptive Lasso estimator is defined by reweighed penalty term

$$\hat{\beta}_{\text{adapt}}(\lambda) = \underset{\beta}{\text{argmin}} \left(||Y - X\beta||_2^2/n + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{\text{init},j}|} \right) \tag{2.3}$$

where $\lambda \geq 0$ is the penalty parameter and $\hat{\beta}_{\text{init}}$ is an initial estimator. There is the two-stage procedure for obtaining adaptive lasso in Bühlmann and Van De Geer (2011) [3].

In the first stage, we use a Lasso estimator as the initial estimator which means

$$\hat{\beta}_{\text{init}} = \hat{\beta}(\lambda) = \underset{\beta}{\text{argmin}} \left(||Y - X_{n \times p}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \right).$$

In the second stage, we use cross-validation again to choose the parameter λ in the adaptive Lasso (2.3). By proceeding in this way, the regularization parameters can be selected sequentially. This is computationally much cheaper because it optimizes twice for a single parameter instead of simultaneously optimizing for two tuning parameters.

There is trivial property of the adaptive Lasso such that

$$\hat{\beta}_{\text{init},j} = 0 \Rightarrow \hat{\beta}_{\text{adapt},j} = 0.$$

In addition, if $|\hat{\beta}_{\text{init},j}|$ is large, the adaptive Lasso uses a small penalty for the jth coefficient β_j. Therefore, we can use the adaptive lasso to generate sparse solutions and reduce the number of false positives in the first step.

2.3 Elastic net

In Zou (2006) [9], the Elastic net estimator is defined by using a combination of the l_1, l_2 penalties

$$\hat{\beta}_{EN}(\lambda_1, \lambda_2) = \underset{\beta}{\text{argmin}} \left(||Y - X\beta||_2^2/n + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2 \right) \tag{2.4}$$

where $\lambda_1, \lambda_2 \geq 0$ are regularization parameters.

The motivation for adding the l_2-norm penalty is as follows. For strongly correlated covariates, Lasso can choose one, but generally not both. From a sparsity point of view, this method works well. However, in terms of interpretation, we may want to have two strongly correlated variables between the selected variables. This is motivated by the idea that we do not want to miss a true variable due to the selection of non-true variables which are highly correlated with the true variable. The computation of the elastic net estimator can be done by using an algorithm for the Lasso.

2.4 Lasso for Generalized Linear Models

Generalized Linear Models (McCullagh and Nelder (2019)) [5] are useful for processing many extensions of a linear model. Let Y be the response variable and $X \in \mathcal{X} \subset$

$\mathbb{R}^p$ be the p-dimensional covariates:

$$Y_1, Y_2, \ldots, Y_n \text{ are independent.}$$
$$g(\mathbb{E}([Y_i|X_i = x]) = f(x) = f_{\mu,\beta}(x) = \mu + \sum_{j=1}^{p} \beta_j x^{(j)}$$

$g(\cdot)$ is known as the link function and μ is the intercept term. The conditional probability density function(pdf) of $Y|X = x$ can be defined as $p(y|x) = p_{\mu,\beta}(y|x)$. This means the conditional pdf of $Y|X = x$ depends on μ, β.

The Lasso can be applied to Generalized linear models. In this case, the Lasso estimator is defined by penalizing the negative log-likelihood with the l_1-norm. The negative log-likelihood is $-\sum_{i=1}^{n} \log\left(\mathrm{p}_{\mu,\beta}\left(Y_i \mid X_i\right)\right)$. The negative log-likelihood can be rewritten with a loss function $\rho(.,.)$:

$$n^{-1} \sum_{i=1}^{n} \rho_{\mu,\beta}\left(X_i, Y_i\right),$$
$$\rho_{\mu,\beta}(x, y) = -\log(p_{\mu,\beta}(y|x)).$$

For many examples, for all x, y, the loss function $\rho_{\mu,\beta}(x, y)$ is often convex in μ, β.

The l_1-norm penalized Lasso estimator is then defined as:

$$\hat{\mu}(\lambda), \hat{\beta}(\lambda) = \underset{\mu,\beta}{\arg\min} \left(n^{-1} \sum_{i=1}^{n} \rho_{\mu,\beta}\left(X_i, Y_i\right) + \lambda \|\beta\|_1 \right).$$

The properties of the Lasso estimator in the generalized linear models are very similar to those of the Lasso estimator in the linear model. There are consistency and variable screening(selection) properties.

2.4.1 Logistic regression

We consider a model with binary response variable Y and p-dimensional covariates $X \in R^p$. Let $Y_i|X_i = x \sim$ Bernoulli$(\pi(x))$ with

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \mu + \sum_{j=1}^{p} \beta_j x^{(j)} = f_{\mu,\beta}(x)$$

where the link function $g(\pi) = \log(\frac{\pi}{1-\pi})$ with $\pi \in (0,1)$.

So, the negative log-likelihood function is

$$-\sum_{i=1}^{n} \log\left(\mathrm{p}_{\mu,\beta}\left(Y_i \mid X_i\right)\right) = \sum_{i=1}^{n} \left\{-Y_i f_{\mu,\beta}\left(X_i\right) + \log\left(1 + \exp\left(f_{\mu,\beta}\left(X_i\right)\right)\right)\right\}$$

The corresponding loss function is

$$\rho_{\mu,\beta}(X_i, Y_i) = -Y_i\left(\mu + \sum_{i=1}^{p} \beta_j X_i^{(j)}\right) + \log\left(1 + \exp\left(\mu + \sum_{i=1}^{p} \beta_j X_i^{(j)}\right)\right).$$

Thus, we can define the Lasso estimator as:

$$\hat{\mu}(\lambda), \hat{\beta}(\lambda) = \underset{\mu,\beta}{\arg\min}\left(n^{-1}\sum_{i=1}^{n} \rho_{\mu,\beta}\left(X_i, Y_i\right) + \lambda\|\beta\|_1\right).$$

Chapter 3

FDR control via data splitting

3.1 Multiple testing

Suppose we test N hypotheses simultaneously, defined H_{0i} vs H_{1i} with $i = 1, \ldots, N$. The multiple testing or multiple comparisons problem is how to determine which null hypothesis is rejected when observing a large number of test statistics. There are several methods to solve this problem such as family-wise error rate(FWER) control and false discovery rate(FDR) control.

First, the FWER is the probability of incorrectly rejecting a true null hypothesis at least once,

$$FWER = P\{\text{reject any true null } H_{0i}\}.$$

We'll show that Bonferroni's procedure controls FWER at significance level α. Let I_0 be the set of index the true H_{0i} with $N_0 = \#I_0$. Define p_i be the p-values from each hypothesis. Then, by using Boole's inequality,

$$\begin{aligned} \text{FWER} &= \mathrm{P}\left\{\bigcup_{I_0}\left(p_i \leq \frac{\alpha}{N}\right)\right\} \leq \sum_{I_0} \mathrm{P}\left\{p_i \leq \frac{\alpha}{N}\right\} \\ &= N_0 \frac{\alpha}{N} \leq \alpha \end{aligned}$$

Therefore, if the significance level of each hypothesis is set to α/N and then tested, the FWER is controlled at the significance level α.

The Bonferroni procedure is too conservative. That is, we reject too few hypotheses. There is a more advanced method than this, Holm's procedure. This method can also control FWER at the same level α. Here's holm's procedure.

- Sort the observed p-values from smallest to largest,

$$p_{(1)} \leq p_{(2)} \leq p_{(3)} \leq \ldots \leq p_{(i)} \leq \ldots \leq p_{(N)}$$

  with $H_{0(i)}$ denoting the corresponding null hypotheses.

- Let i_0 be the smallest index $i \in \{1, \ldots, N\}$ such that $p_{(i)} > \alpha/(N - i + 1)$. This means,

$$i_0 = \min\{i : p_{(i)} > \alpha/(N - i + 1)\}$$

- Reject all $H_{0(i)}$ for $i < i_0$ and accept all $H_{0(i)}$ with $i \geq i_0$.

Next, we define the false discovery rate(FDR). Similar to FWER, assume that when N hypotheses are simultaneously tested to account for the false discovery rate, the actual number of null hypotheses is N_0. When the R null hypotheses were rejected, we define the number of hypotheses that were incorrectly rejected as a. Let D be a de-

	Accept Null	Reject Null	Total
Null hypothesis is true	$N_0 - a$	a	N_0
Alternative hypothesis is true	$N_1 - b$	b	N_1
Total	$N - R$	R	N

Table 3.1: The possible outcomes when testing multiple null hypotheses.

cision rule that rejects R out of N null hypotheses. Then, the false discovery proportion (Fdp) is defined as

$$Fdp(D) = \frac{a}{R}.$$

We define $Fdp(D) = 0$ if $R = 0$. Since a is a random variable that cannot be observed, the false discovery rate(FDR) is the expected value of $Fdp(D)$, denoted

$$FDR(D) = E\{Fdp(D)\}.$$

We can find the decision rule D that controls FDR at level q, with $q \in (0, 1)$ a preselected value. This means $FDR(D) = E\{Fdp(D)\} \leq q$.

3.2 BHq procedure

There are many ways to control FDR, the Benjamin-Hochberg procedure, which was first introduced in 1995. There is an assumption that all variables are independent.

Theorem 1 (Benjamini and Hochberg (1995) [2]) *Let $H_1, \ldots, H_N$ be the N hypotheses and $p_1, \ldots, p_N$ be the p-value corresponding to each hypothesis. Define $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$ as ordered p-value with $H_{(i)}$ denoting the corresponding hypotheses. For given control level $q \in (0, 1)$, define $i_{max} = \max\{i : p_{(i)} \leq \frac{i}{N}q\}$. We set the decision rule D_q to reject all hypotheses $H_{(i)}$ for $i \leq i_{max}$. Then,*

$$FDR(\mathcal{D}_q) = \frac{N_0}{N}q \leq q$$

Furthermore, a method of controlling FDR in the case of dependency between features was also introduced. Benjamini and Yekutieli (2001) [10] generalized the BHq procedure to handle when p-values are dependent. Sarkar (2002) [11] generalized the BHq procedure for general step-wise multiple testing procedures with positive dependence. Storey et al. (2004) [12] handled the case of weak dependence.

3.3 FDR control in Regression models

In high-dimesional linear regression model, the hypotheses are

$$H_{0i} : \beta_i = 0 \text{ vs } H_{1i} : \beta_i \neq 0$$

Then, we can select the significant features by applying the FDR control procedures as follows.

$$FDR = \mathbb{E}[FDP], \quad FDP = \frac{\#\{j : j \in S_0, j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} \vee 1} \tag{3.1}$$

where $\hat{S} = \{j : H_{0j} \text{ is rejected } \}$ is the selected relevant features.

3.4 Single Data Splitting(DS)

In Dai et al. (2022) [4], it is difficult to find the p-values and the joint distribution of explanatory features in high dimensional. So, there are limitations and difficulties in applying the BHq procedure.

Unlike the traditional method of selecting the features with a regression coefficient value estimated using the entire data, Dai et al. (2022) [4] introduced the single data splitting(DS) method. We split the data into two halves, denoted as $(y^{(1)}, X^{(1)})$, $(y^{(2)}, X^{(2)})$. Then, we estimate two independent coefficients by applying two potentially different statistical procedures to each of the two groups, denoted as $\hat{\beta}_j^{(1)}, \hat{\beta}_j^{(2)}$. Since we do not use the p-values, Dai et al. (2022) [4] defined new statistics, called the mirror statistics. The mirror statistics M_j for each feature X_j is

$$M_j = sign(\hat{\beta}_j^{(1)}\hat{\beta}_j^{(2)}) f(|\hat{\beta}_j^{(1)}|, |\hat{\beta}_j^{(2)}|)$$

Here, $f(u, v)$ has several conditions. It is non-negative, symmetric for u and v. Also, it is a monotone increasing function of u and v. Then, Dai et al. (2022) [4] mentioned

that the mirror statistics satisfy the following properties.

(a) If a feature has a large mirror statistic, the feature is likely to be a significant feature.

(b) The mirror statistic of a non significant feature has symmetric sampling distribution about 0.

Since we don't know exactly the number of false positives $\{j \in S_0 : M_j > t\}$ for given threshold value t and the mirror statistics, we'll find the upper bound of $\{j \in S_0 : M_j > t\}$ which we can estimate. The symmetric assumption of the mirror statistic for the null feature is used to show the following upper bound of $\{j \in S_0 : M_j > t\}$.

$$\#\{j \in S_0 : M_j > t\} \approxeq \#\{j \in S_0 : M_j < -t\} \leq \#\{j : M_j < -t\}, \quad \forall t > 0. \tag{3.2}$$

We can select the index of relevant variables denoted $\hat{S}_t = \{j : M_j > t\}$. Then, the FDP(t) of the $\hat{S}_t$ is given by

$$FDP(t) = \frac{\#\{j : M_j > t, j \in S_0\}}{\#\{j : M_j > t\} \vee 1} \tag{3.3}$$

Thus, we use the estimator of FDP(t) by

$$\widehat{FDP}(t) = \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\} \vee 1} \tag{3.4}$$

Next, we set the level $q \in (0, 1)$ of FDR control. So we can define the cutoff value

$$\tau_q = \min\{t > 0 : \widehat{FDP}(t) \leq q\}$$

Thus, we can finally choose the index set of the relevant variables by $\hat{S}_{\tau_q} = \{j : M_j > \tau_q\}$.

There are more assumptions to obtain a good estimate of the number of false pos-

itives. First, the mirror statistics for null features should not be too correlated. Second, the variance of mirror statistics converges to finite. Then, Dai et al. (2022) [4] showed that $FDR(\tau_q)$ is controlled to the level q. The proof of this is detailed in Dai et al. (2022) [4].

3.5 Multiple Data Splitting(MDS)

Dai et al. (2022) [4] mentioned that there are two problems with single data splitting(DS). First, splitting the data into two halves inflates the variance of the estimated regression coefficient. So, DS can potentially suffer a power loss. Second, the selection result of DS may not be stable and can vary substantially across different sample splits.

To solve this problem, Dai et al. (2022) [4] proposed that we collect the selection results obtained from independent replication of DS using multiple data segmentation procedures. Suppose we use random sample splits to independently repeat DS m times. Define the $\hat{S}^{(k)}$ to be the set of selected features in the kth trials for $k = 1, 2, \ldots, m$ and the associated inclusion rate I_j and its estimate $\hat{I}_j$ as

$$I_j = \mathbb{E}\left[\frac{\mathbf{1}(j \in \hat{S})}{|\hat{S}| \vee 1} | X, y\right], \quad \hat{I}_j = \frac{1}{m}\sum_{k=1}^{m} \frac{\mathbf{1}(j \in \hat{S}^{(k)})}{|\hat{S}^{(k)}| \vee 1}.$$

This ratio is a measure of the importance of each feature related to the data splitting procedure. In other words, if a feature is selected with a small frequency, it tends to be not a significant feature. Then, similar to DS, the cutoff value of the inclusion rate that controls the FDR can be found.

Dai et al. (2022) [4] mentioned that there are three steps to finding the cutoff value. First, we sort the estimated values of the inclusion rate, denoted $0 \leq \hat{I}_{(1)} \leq \hat{I}_{(2)} \leq \cdots \leq \hat{I}_{(p)}$. Second, we find the largest index $l \in \{1, \ldots, p\}$ such that $\hat{I}_{(1)} + \hat{I}_{(2)} + \cdots + \hat{I}_{(l)} \leq q$. Finally, we select the relevant features $\hat{S} = \{j : \hat{I}_j > \hat{I}_{(l)}\}$. Similar to DS, the FDR of MDS is controlled to the level q. The proof of this is detailed in Dai et

al. (2022) [4].

3.6 Application for linear models

Let the design matrix X be random which means each row of X follows independently p-dimensional distribution and β^* be the true coefficient with $p \gg n$. We estimate the $S_1 = \{j : \beta_j \neq 0\}$ which is contained true relevant features.

Dai et al. (2022) [4] proposed LASSO + OLS procedure. Firstly, to reduce the dimension, we apply the LASSO to the first data $(y^{(1)}, X^{(1)})$. Then, $\hat{\beta}^{(1)}$ is defined as the lasso estimator. Let $\hat{S}^{(1)} = \{j, \hat{\beta}_j^{(1)} \neq 0\}$. Next, we use the restricted features set $\hat{S}^{(1)}$. Let $X^{(2)}_{\hat{S}^{(1)}}$ be the restricted design matrix of $X^{(2)}$. Then, we proceed the ordinary least square method for $(y^{(2)}, X^{(2)}_{\hat{S}^{(1)}})$. Let $\hat{\beta}^{(2)}$ be the least square estimator. Finally, using $\hat{\beta}^{(1)}, \hat{\beta}^{(2)}$, we obtain the mirror statistics. So, we can apply the DS or MDS procedures.

Chapter 4

Simulation Study

In this chapter, we assume the linear model and do simulations. Firstly, we will use the Lasso, adaptive Lasso. Secondly. We will use DS and MDS methods. We compare which methods find more significant features.

4.1 Lasso vs Adaptive Lasso

Using $p = 500$ and $n = 50$, we describe Lasso and adaptive Lasso in some simulation data of the linear model. We select $\beta_1 = 1.5, \beta_2 = 2.3, \beta_3 = -1.3, \beta_4 = 0.7$ and $\beta_5 = \cdots = \beta_{500} = 0, \epsilon \sim N(0,1)$ and $X^{(1)}, \ldots, X^{(500)} \overset{i.i.d}{\sim} N(0,1)$ where $Y = f(X) + \epsilon = X\beta + \epsilon$.

Figure 4.1 shows the coefficient estimates for Lasso and adaptive Lasso, respectively, using the initial estimator from Lasso. The tuning parameters are chosen as follows: For Lasso, we use the optimal λ in 10-fold cross-validation. This Lasso fit is used as an initial estimator and then re-optimizes the 10-fold cross-validation to select λ for the second step of the adaptive Lasso. We can empirically see that Lasso is a powerful screening method here. All four relevant variables $\beta_1, \beta_2, \beta_3, \beta_4$ are chosen. This means that $\hat{S} \supseteq S_0$, but it also selects 26 noise covariates. The adaptive Lasso provides a fairly sparse fit. All of the 4 relevant variables and 2 noise covariates are

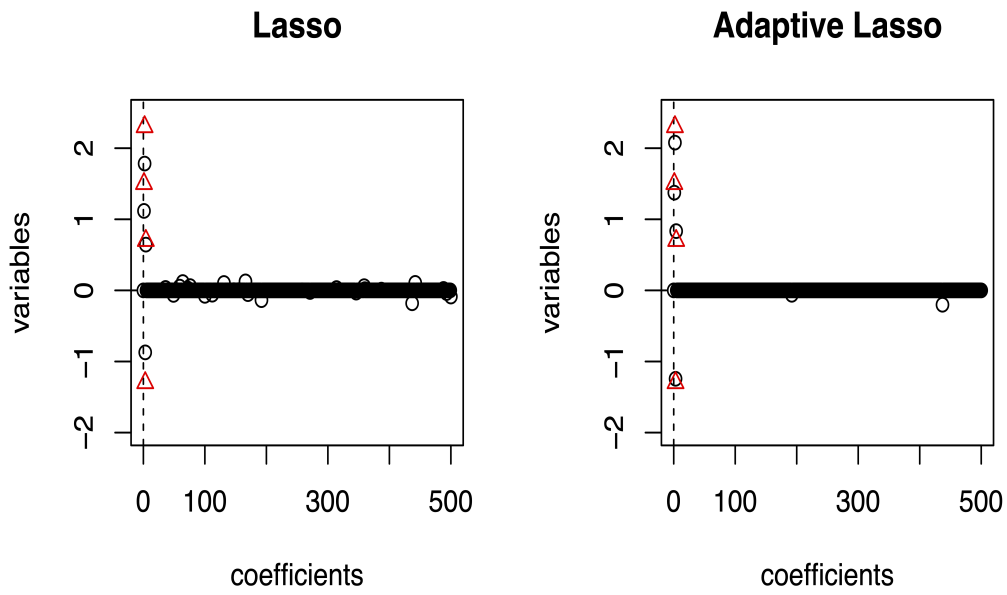

Figure 4.1: Estimated regression coefficients using lasso and adaptive lasso

selected. Therefore, if the number of significant variables is sparse, the Adaptive Lasso provides a better estimate.

4.2 DS vs MDS

Also, we will check that the MDS method is more stable than the DS method. We proceed with the simulation setting in Dai et al. (2022) [4] in more various ways. Let $y_{n\times 1} = X_{n\times p}\beta^*_{p\times 1} + \epsilon_{n\times 1}$ with $\epsilon \sim N(\mathbf{0}, I_n)$ and randomly locate the significant index set S_1. The distribution of β^*_j for $j \in S_1$ is $\beta^*_j \sim N(0, \delta\sqrt{\log p/n})$ where δ is signal strength. The distribution of each row of $X_{n\times p}$ follows $N(\mathbf{0}, \Sigma)$ where Σ is a Toeplitz covariance matrix with correlation ρ. We set the FDR control level $q = 0.1$. In this simulation, we set various situations and measure the empirical power of MDS.

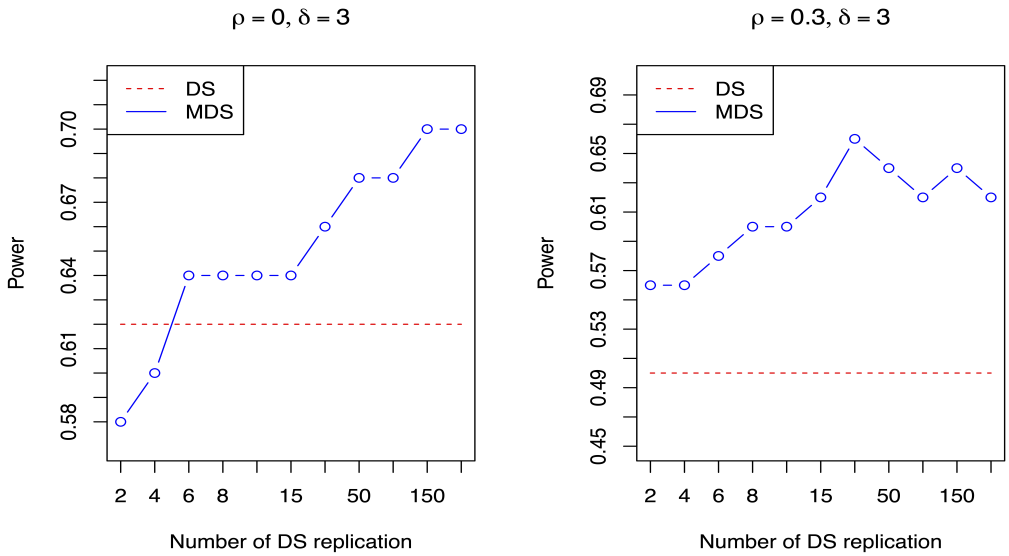

Figure 4.2: $\rho = 0, 0.3$ with $\delta = 3, n = 500, p = 500, p_1 = 50$

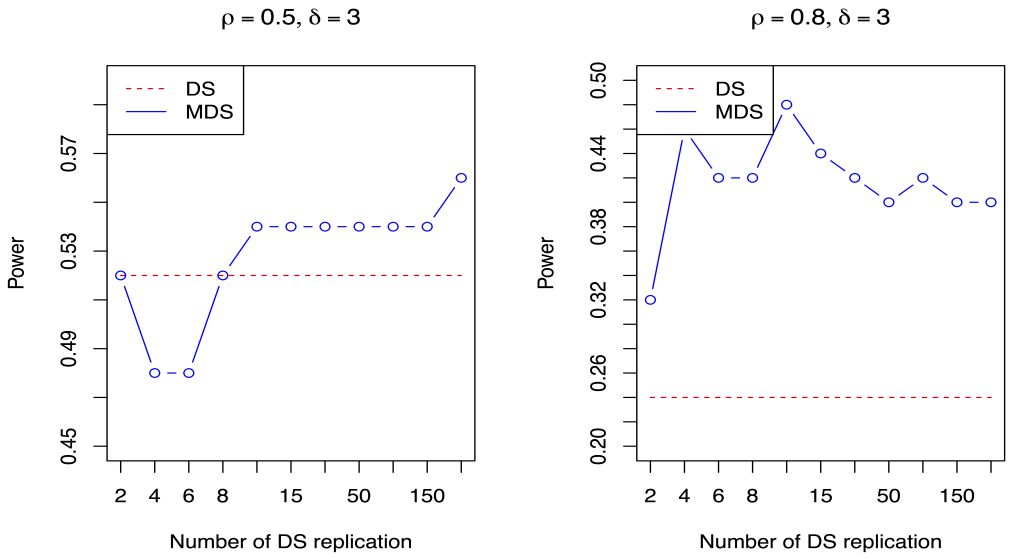

Figure 4.3: $\rho = 0.5, 0.8$ with $\delta = 3, n = 500, p = 500, p_1 = 50$

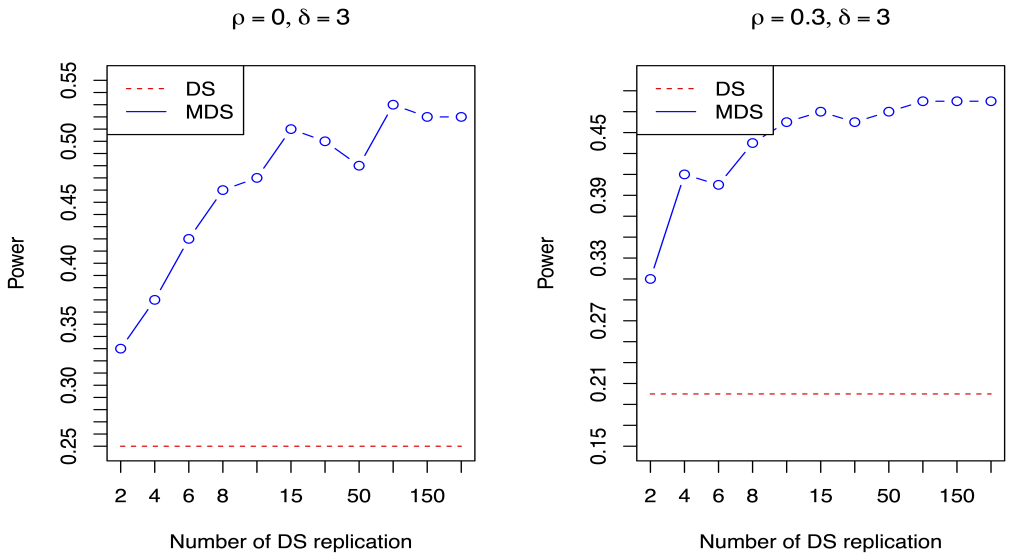

Figure 4.4: $\rho = 0, 0.3$ with $\delta = 3, n = 500, p = 1000, p_1 = 100$

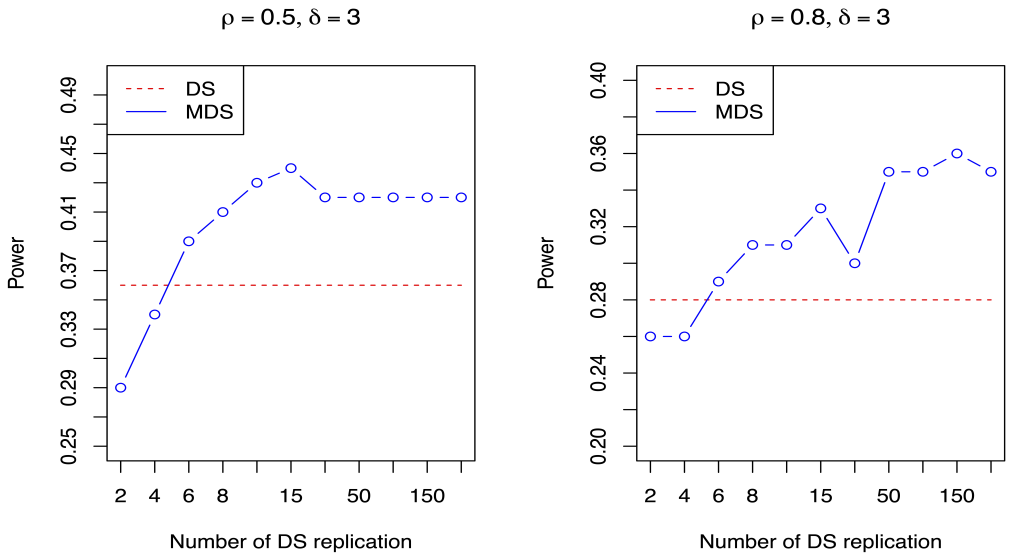

Figure 4.5: $\rho = 0.5, 0.8$ with $\delta = 3, n = 500, p = 1000, p_1 = 100$

Let m be the number of DS replication. The empirical power of MDS is obtained as

$$\widehat{Power} = \frac{\#(S_1 \cap \hat{S}_1)}{p_1}.$$

We calculate the empirical power of MDS with different m. For all case, the empirical power of MDS increases with m and is stable after $m \geq 50$. Therefore, the MDS method is empirically more stable and has more power than the DS method.

Chapter 5

R Code

We attach the R code for the simulation results. The following code compares Lasso and Adaptive Lasso.

```
library(glmnet)
library(MASS)

covariance <- diag(500)

### Generate Data
generate_data <- function(p, n){
  beta_star <- rep(0, p)
  signal = c(1,2,3,4)
  beta_star[signal] <- c(1.5, 2.3, -1.3, 0.7)
  set.seed(123)
  X = mvrnorm(n, mu = rep(0, p),  Sigma = covariance)
  y <- X%*%beta_star + rnorm(n, mean = 0, sd = 1)

  return(list(X = X, y = y))
```

```
}

a = generate_data(500, 50)
X = a$X # design matrix : 50 x 1000
y = a$y # dependent variable : 50 x 1

# Lasso
set.seed(123)
cv1 <- cv.glmnet(X, y, alpha = 1, intercept = F)
best_lambda <-  cv1$lambda.min  #best lambda

## lasso estimator
beta = coef(cv1, s = "lambda.min")
tmp <- as.data.frame(as.matrix(beta))

rel_coef_index = which(tmp$s1 != 0)
est_coef = tmp$s1[rel_coef_index]

par(mfrow = c(1, 2))
plot(seq(0,500), tmp$s1, type = "p", main = "Lasso",
    xlab="coefficients", ylab = "variables", ylim = c(-2, 2.5))
abline(h=0,col="black",lty=2)
abline(v=0, col ="black",lty=2)
points(c(1,2,3,4), c(1.5, 2.3, -1.3, 0.7), col = "red", pch = 2)

# adaptive lasso
set.seed(123)
alasso1 <- cv.glmnet(X, y, alpha = 1,
    penalty.factor = 1 / abs(tmp$s1[2:501]), intercept = F)
```

```
best_lambda_adaptive <- alasso1$lambda.min

## adaptive lasso estimator
beta_adaptive <- coef(alasso1, s = "lambda.min")
tmp_adap <- as.data.frame(as.matrix(beta_adaptive))

rel_coef_index_adap = which(tmp_adap$s1 != 0)
est_coef = tmp_adap$s1[rel_coef_index_adap]

plot(seq(0,500), tmp_adap$s1, type = "p", main = "Adaptive Lasso",
        xlab="coefficients", ylab = "variables", ylim = c(-2, 2.5))
abline(h=0,col="black",lty=2)
abline(v=0, col ="black",lty=2)
points(c(1,2,3,4), c(1.5, 2.3, -1.3, 0.7), col = "red", pch = 2)
```

Next, the code to compare DS and MDS is on the Dai et al. (2022) [4] Github. The link to GitHub is "https://github.com/Jeremy690/False-Discovery-Rate-via-Data-Splitting".

Chapter 6

Conclusion

In high-dimensional regression analysis, we introduce the classic methods of variable selection methods to the latest methods. Among the classical methods, we explain several properties of the Lasso estimator and the Adaptive Lasso, Elastic net and generalized linear models.

Next, we introduce a method of variable selection by controlling the false discovery rate. Among the multiple test methods, the BHq procedure method and the method for controlling FDR in linear models are introduced. In addition, the FDR control via data splitting proposed by Dai et al. (2022) [4] is introduced.

Dai et al. (2022) [4] define a new statistic, the mirror statistic, and use the property that this statistic has a distribution symmetrical to 0 in the case of the null hypothesis, and large positive values in the case of the alternative hypothesis. In this way, FDP, which is an approximate value of the FDR value, is estimated, and significant variables are selected by adjusting the threshold so as not to exceed a predetermined FDR level.

Dai's method allows us to select relevant features without looking for p-values or joint distributions of high dimensional features.

Bibliography

[1] Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. *Selected papers of hirotugu akaike*, 199–213.

[2] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289–300.

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

[4] Dai, C., Lin, B., Xing, X., & Liu, J. S. (2022). False discovery rate control via data splitting. *Journal of the American Statistical Association*, 1-18.

[5] McCullagh, P. (2019). *Generalized linear models*, Routledge.

[6] Schwarz, G. (1978), Estimating the dimension of a model. *The annals of statistics*, 461–464.

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

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

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

[10] Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165-1188.

[11] Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *The Annals of Statistics*, 30(1), 239-257.

[12] Storey, J. D., Taylor, J. E., & Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66(1)*, 187-205.

초 록

최근 많은 분야에서 고차원 데이터 분석이 주목받고 있다. 특히 수많은 변수 중에서 유의미한 영향을 미치는 변수를 선택하는 것은 어렵고도 중요한 문제이다. 이 문제를 해결하기 위해 선형 회귀 모델에서 다중 검정 및 LASSO와 같은 몇 가지 통계적 방법이 있다.

본 논문에서 우리는 Bühlmann and Van De Geer (2011) [3]에서 다루는 Lasso, Adaptive Lasso, Elastic net 및 일반화된 선형 모델 사례를 소개한다. 또한, 우리는 여러 다중 검정 절차를 다루고 Dai ei al. (2022) [4]가 최근에 제안한 데이터 분할을 통한 허위 발견률(FDR) 제어 방법을 소개한다. 마지막으로 유의한 변수가 희소한 경우 Lasso 추정량보다 적응형 Lasso 추정량이 더 좋은 결과를 주는지 시뮬레이션으로 확인해 본다. 그리고 DS와 MDS 방법으로 FDR을 통제하는 경우 MDS 방법이 DS 방법보다 더 안정적이고 경험적 검정력이 높은 것을 시뮬레이션을 통해 확인해 본다.

주요어: 변수 선택 방법론, Lasso, Adaptive Lasso, Elastic net, Generalized Lasso, 거짓 발견율(FDR), BHq 절차, 거울 통계량, 데이터 분할 (DS), 다중 데이터 분할 (MDS)

학번: 2021-28605