IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 August 2023, accepted 30 August 2023, date of publication 4 September 2023, date of current version 11 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3312015

RESEARCH ARTICLE

Variable Selection of Lasso and Large Model

HUIYI XIA
School of Big Data and Artificial Intelligence, Chizhou University, Chizhou 240007, China
e-mail: xiayz_88@163.com

ABSTRACT In order to clarify the variable selection of Lasso, Lasso is compared with two other variable selection methods AIC and forward stagewise. First, the variable selection of Lasso was compared with that of AIC, and it was discovered that Lasso has a wider application range than AIC. The data simulation shows the variable selection of Lasso under orthonormal design is consistent with AIC, Lasso under orthonormal design can be solved by using the stepwise selection algorithm. The removed variables of Lasso appear again under nonorthonormal design, the variable selection of Lasso under nonorthonormal design isn't consistent with AIC. We continue to compare the variable selection of Lasso and forward stagewise. Based on the analysis of these studies, it is pointed out that the variable selection of Lasso is complex. An infinite number of parameters enable the design matrix to achieve orthonormalization, so that the solution of Lasso can be found with the stepwise selection algorithm, which may be the reason for the success of the large model represented by ChatGPT.

INDEX TERMS Variable selection, Lasso, AIC, forward stagewise, complexity, ChatGPT.

I. INTRODUCTION

With the development of science, more complex and large data sets appear, statisticians and researchers are also developing different statistical models to extract valuable information from these data-sets, performing parameter estimation, hypothesis testing or statistical inference. To simplify the calculation of data, methods of variable selection are widely applied in data analysis.

The traditional variable selection method is subset selection. If the model has p variables, the subset selection obtains the optimal model by comparing $2^p - 1$ sub models, and the amount of calculation is too large. To reduce the number of calculations, many scientists have conducted research. Breaux proposed a stepwise regression that includes two regression methods. The first method is forward selection, and the second is backward elimination [1]. Stepwise regression improves computational efficiency, but it cannot guarantee that the obtained model is optimal. The model obtained by subset selection is the optimal model, but the number of calculations is too large. To determine the optimal model, Breiman proposed a nonnegative garotte [2]. Tibshirani inspired by the garotte, put forward a new technique, called the Lasso (least absolute shrinkage and selection operator). it shrank some coefficients and set others to 0, and hence tried to retain the features of both subset selection and ridge regression [3]. To reduce the computational cost of Lasso, Efron et al. proposed least angle regression (Lars) and proved that modified Lars could solve Lasso [4].

Owing to the fast-computing speed of Lars, it has become the main algorithm for studying high-dimensional data. Khan and Shaw considered variable selection methods for the AFT modeling of censored data, and introduced classes of elastic net type regularized variable selection techniques based on SWLS [5]. Kane and Mandal proposed applying the adaptive Lasso regression as an analytical tool for designs with complex aliasing [6]. Febrero-Bande et al. considered the problem of variable selection in regression models for functional variables [7]. Xia studied the disturbance phenomenon in the process of Lasso's variable selection, and pointed out the complexity of Lasso's variable selection [8]. Borboudakis and Tsamardinos proposed a heuristic that significantly improved its running time, while preserving predictive performance. The idea is to temporarily discard variables that are conditionally independent with the outcome given the selected variable set. Depending on how these variables are reconsidered and reintroduced, this heuristic gives rise to a family of algorithms with increaseingly

The associate editor coordinating the review of this manuscript and approving it for publication was Chong Leong Gan.

stronger theoretical guarantees [9]. Febrero-Bande et al. considered the problem of variable selection in regression models for functional variables [7]. Liang et al. proposed VSOLasso-Bag, which is a variable selection-oriented Lasso bagging algorithm for biomarker discovery in omic-based translational research [10]. Wasserman and Roeder performed variable selection in high-dimensional models, and considered three screening methods: the Lasso, marginal regression, and forward stepwise regression [11]. Austin et al. studied penalized regression and risk prediction in a genome-wide association study by Lasso [12]. Ahrens and Bhattacharjee exploited the Lasso estimator and mimicked two-step least squares to account for the endogeneity of the spatial lag [13].

Following is a brief introduction to Lasso:

Suppose that we have data $\left(\mathrm{x}^{\mathrm{i}}, \mathrm{y}_{\mathrm{i}}\right), \mathrm{i} = 1, \cdots, \mathrm{N}$, where $\mathrm{x}^{\mathrm{i}} = \left(\mathrm{x}_{\mathrm{i}1}, \cdots, \mathrm{x}_{\mathrm{ip}}\right)$ are the predictor variable and y_i are responses. As in the usual regression set up, we assume that either the observations are independent or the y_is are conditionally independent given the x_{ij}s. Assume that the x_{ij} is standardized such that $\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\mathrm{x}_{\mathrm{ij}}}{\mathrm{N}} = 0$, $\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x}_{\mathrm{ij}}^2 = 1$.

Without a loss of generality, we assumed that $\bar{\mathrm{y}} = 0$. Letting $\hat{\beta} = \left(\hat{\beta}_1, \cdots, \hat{\beta}_{\mathrm{p}}\right)^{\mathrm{T}}$, the Lasso estimates $\hat{\beta}$ is defined by

$$\hat{\beta} = \operatorname{argmin} \left\{ \sum_{\mathrm{i}=1}^{\mathrm{N}} \left(\mathrm{y}_{\mathrm{i}} - \sum_{\mathrm{j}=1}^{\mathrm{p}} \beta_{\mathrm{j}} \mathrm{x}_{\mathrm{ij}} \right)^2 \right\}$$
$$\text{subject to } \sum_{\mathrm{j}=1}^{\mathrm{p}} \left|\beta_{\mathrm{j}}\right| \le t \tag{1}$$

Here $\mathrm{t} \ge 0$ is a tuning parameter.

Let $\beta^0 = \left(\hat{\beta}_1^0, \cdots, \hat{\beta}_{\mathrm{p}}^0\right)^{\mathrm{T}}$ be the least squares estimation and X be the $n \times p$ design matrix with ijth entry x_{ij}. Under orthonormal design ($\mathrm{X}^{\mathrm{T}}\mathrm{X} = \mathrm{I}$, I denotes the $\mathrm{p} \times \mathrm{p}$ identity matrix), the soft threshold estimation of Lasso was proposed in 1996, and the expression is as follows:

$$\hat{\beta}_{\mathrm{j}} = \operatorname{sign}\left(\hat{\beta}_{\mathrm{j}}^0\right)\left(\left|\hat{\beta}_{\mathrm{j}}^0\right| - \gamma\right)^+ \tag{2}$$

where γ is determined by the condition $\sum \left|\hat{\beta}_{\mathrm{j}}\right| = \mathrm{t}$.

Based on summarizing previous research, under orthonormal design, and when $\left|\hat{\beta}_1^0\right| < \cdots < \left|\hat{\beta}_{\mathrm{p}}^0\right|$, the author provided the exact solution of Lasso using the Lagrange multiplier method, and the expression is as follows:

$$\begin{cases} \hat{\beta}_1 = \hat{\beta}_1^0, \cdots, \hat{\beta}_{\mathrm{p}} = \hat{\beta}_{\mathrm{p}}^0 \text{ subject to } \mathrm{t} \ge \sum_{\mathrm{i}=1}^{\mathrm{p}} \left|\hat{\beta}_{\mathrm{i}}^0\right| \\ \hat{\beta}_{\mathrm{k}} = \operatorname{sgn}\left(\hat{\beta}_{\mathrm{k}}^0\right)\left(\left|\hat{\beta}_{\mathrm{k}}^0\right| - \dfrac{1}{\mathrm{p}-\mathrm{j}+1}\left(\sum_{\mathrm{k}=\mathrm{j}}^{\mathrm{p}} \left|\hat{\beta}_{\mathrm{k}}^0\right| - \mathrm{t}\right)\right)^+ \\ \text{subject to } \mathrm{t} < \sum_{\mathrm{i}=1}^{\mathrm{p}} \left|\hat{\beta}_{\mathrm{i}}^0\right|, \sum_{\mathrm{k}=\mathrm{j}}^{\mathrm{p}} \left|\hat{\beta}_{\mathrm{k}}\right| = \mathrm{t} \end{cases} \tag{3}$$

The soft threshold solution of Lasso is a numerical solution, and the exact solution of Lasso is an analytical expression. These theoretical studies indicate that Lasso under orthonormal design can be solved by the stepwise selection algorithm.

Under nonorthonormal design ($\mathrm{X}^{\mathrm{T}}\mathrm{X} \ne \mathrm{I}$), letting $\beta = \left(\beta_1, \cdots, \beta_{\mathrm{p}}\right)^{\mathrm{T}}$, equation (1) is equivalent to:

$$\hat{\beta} = \operatorname{argmin}\left\{\left(\beta - \hat{\beta}^0\right)^{\mathrm{T}} \mathrm{X}^{\mathrm{T}}\mathrm{X}\left(\beta - \hat{\beta}^0\right)\right\}$$
$$\text{subjectto } \sum\nolimits_{\mathrm{j}=1}^{\mathrm{p}} \left|\beta_{\mathrm{j}}\right| \le t \tag{4}$$

Because $\mathrm{X}^{\mathrm{T}}\mathrm{X} \ne \mathrm{I}$, equation (4) cannot be solved by the Lagrange multiplier method; we can only find numerical solutions by computer simulation.

The software used in this paper is R software.

Although Lasso has a wide range of applications, there are still many issues with its variable selection. Therefore, the variable selection of Lasso is studied in this paper and being compared with that of the other two method Akaike Information Criterion (AIC) and forward stagewise.

This study is divided into five sections. The first section is the introduction, which highlight the research origin of Lasso's variable selection. The second section introduced the variable selection of Lasso, and compares it with the variable selection of AIC. The third section compares the variable selection of Lasso with that of forward stagewise. The fourth section concludes the paper by pointing out the complexity of Lasso's variable selection and the relationship between Lasso and the large model.

II. VARIABLE SELECTION OF LASSO AND AIC

First, we study the variable selection of Lasso under orthonormal design. We then study the variable selection of Lasso under nonorthonormal design. Finally, we study the variable selection of Lasso on the data set. Meanwhile, the variable selection of Lasso was compared with the variable selection of AIC, and their relationship was determined.

A. VARIABLE SELECTION UNDER ORTHONORMAL DESIGN

Here is an example of Lasso under orthonormal design.

Example 1: The design matrix is

$$\mathrm{X} = \begin{pmatrix} 0.5 & 0 & 0.5 & -0.5 & -0.5 \\ 0.5 & 0 & -0.5 & 0.5 & -0.5 \\ 0.5 & 0 & -0.5 & -0.5 & 0.5 \end{pmatrix}^{\mathrm{T}}.$$

The response variable is $\mathrm{y} = (1, -5, -4, 3, 5)^{\mathrm{T}}$. The variable selection of Lasso and AIC was observed.

The least squares estimation of β is $(-4.5, 1.5, 2.5)^{\mathrm{T}}$ and

$$\mathrm{X}^{\mathrm{T}}\mathrm{X} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

This is the problem of Lasso under orthonormal design.

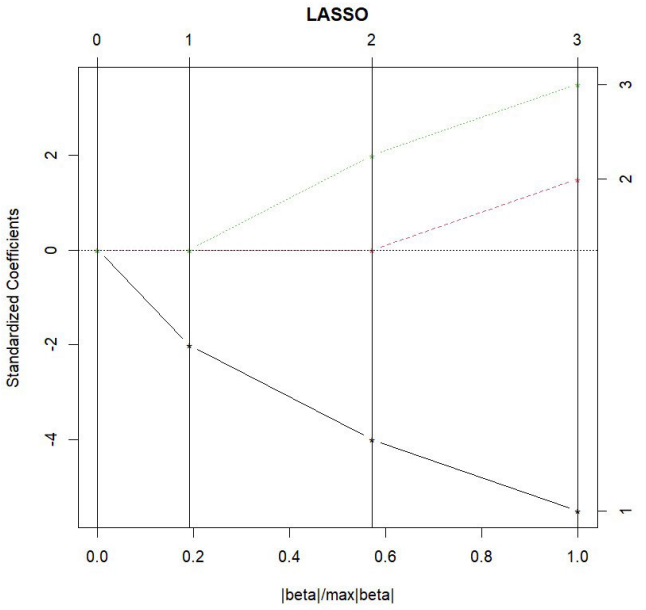

FIGURE 1. Graph1 of Lasso's variable selection.

TABLE 1. Coeffcient change1 of Lasso.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Node 1	-4.5	1.5	2.5
Node 2	-3.0	0.0	1.0
Node 3	-2.0	0.0	0.0

TABLE 2. AIC of two variable.

RV	X2	X3	X1
AIC	15.511	16.075	17.661

TABLE 3. AIC of one variable.

RV	X3	X1
AIC	14.366	15.874

1) VARIABLE SELECTION OF LASSO

Figure 1 shows the variable selection of Lasso in Example 1. As Figure 1 shows, the order of Lasso's variable selection is as follows:

First, the variable X2 is removed, and the absolute value of X2 is the minimum absolute value of the least squares estimators.

Then, the variable X3 is removed, and the absolute value of X3 is the second smallest absolute value of the least squares estimators.

Finally, the variable X1 remains, and the absolute value of X1 is the maximum absolute value of the least squares estimators.

Figure 1 was drawn using the built-in software lars in R. The numbers at the top of Figure 1 represents the node of Lasso's variable selection. Number 3 represents the starting node for Lasso's variable selection, number 2 represents the node when variable X2 is removed, and number 1 represents the node when variable X3 is removed. The numbers on the right side of Figure 1 represent the specific variables. For example, the number 1 represents variable X1, and the corresponding line represents the variable selection process of variable X1.

The variable selection of Lasso is consistent with the absolute value size of the least squares estimators.

Table 1 shows the values of parameter β at each node in Figure 1. As listed in Table 1, the value of parameter β at each node is as follows:

Node 1: The value of parameter β is $(-4.5, 1.5, 2.5)^{\mathrm{T}}$, which is the least squares estimation.

Node 2: The value of parameter β is $(3.0, 0, 1.0)^{\mathrm{T}}$, and variable X2 is removed.

Node 3: The value of parameter β is $(-2.0, 0.0, 0.0)^{\mathrm{T}}$, and variable X3 is removed.

2) VARIABLE SELECTION OF AIC

AIC, also known as the Akaike Information criterion, can improve the goodness of fit and avoid over-fitting. We use AIC to select variables, and the variable selection of AIC in the problem involves two steps:

Step1: Table 2 shows the AIC values of the model after removing one of the three variables. As listed in Table 2, according to the principle of minimizing the AIC value, the variable X2 is removed.

Here, RV denotes Removed variable.

Step2: Table 3 shows the AIC values of the model after removing one of the two variables. As shown in Table 3, according to the principle of minimizing AIC values, variable X3 is removed.

In summary, the variable selection order of AIC is:

First, variable X2 is removed, and the absolute value of variable X2 is the minimum absolute value of the least squares estimators.

Then, variable X3 is removed, and the absolute value of X3 is the second smallest absolute value of the least squares estimators.

Finally, variable X1 remains, and the absolute value of X1 is the maximum absolute value of the least squares estimators.

Therefore, the variable selection of the Lasso is consistent with the variable selection of the AIC in the example.

B. VARIABLE SELECTION UNDER NONORTHONORMAL DESIGN

There is variable disturbance in the variable selection of Lasso under nonorthonormal design, that is, the removed variables will enter again. This means that the stepwise selection algorithm cannot solve Lasso.

Here is an example of nonorthonormal design.

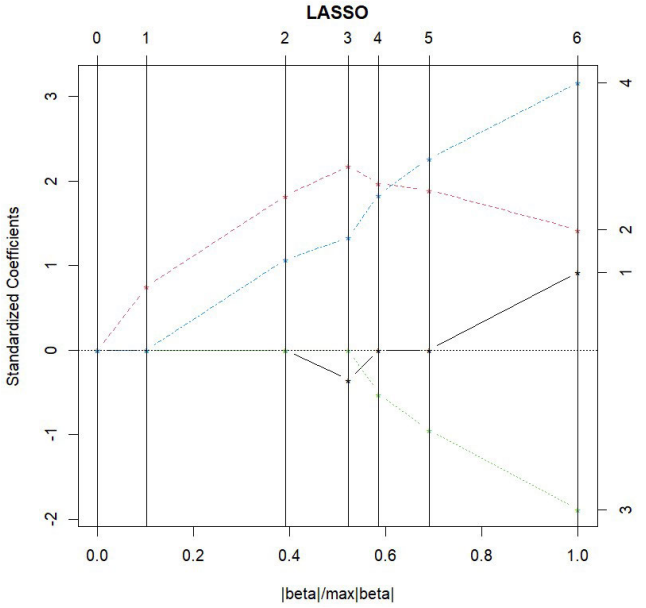

FIGURE 2. Graph2 of Lasso's variable selection.

Example 2: The design matrix is

$$X = \begin{pmatrix} 1/2 & 1/\sqrt{9.5} & 1/\sqrt{6} & 1/\sqrt{14} \\ 0 & 1/\sqrt{38} & 1/\sqrt{6} & 0 \\ 1/2 & -2/\sqrt{9.5} & 0 & -3/\sqrt{14} \\ -1/2 & 2/\sqrt{9.5} & -2/\sqrt{6} & 0 \\ -1/2 & 1/\sqrt{38} & 0 & 2/\sqrt{14} \end{pmatrix}.$$

The response variable is $y = (1, -1, -3, 2, 1)^T$.The variable selection of Lasso and AIC was observed.

The least squares estimation of β is $(0.923, 1.423, -1.884, 3.166)^T$, and

$$X^TX = \begin{pmatrix} 1.00 & -0.41 & 0.61 & -0.53 \\ -0.41 & 1.00 & -0.46 & 0.52 \\ 0.61 & -0.46 & 1.00 & 0.11 \\ -0.53 & 0.52 & 0.11 & 1.00 \end{pmatrix}.$$

This is the problem of Lasso under nonorthonormal design. The example is constructed by the author, the stepwise selection algorithm cannot solve Lasso under nonorthonormal design. When Lasso under nonorthonormal design satisfies the conditions, we can use the stepwise selection algorithm to solve it, which is a complex problem and very valuable. However, further research was not be conducted in this study.

1) VARIABLE SELECTION OF LASSO

Figure 2 shows the variable selection of Lasso in Example 2. As shown in Figure 2, the variable X1 is removed, and, then entered. Variable X1 is the disturbance variable in the variable selection of Lasso.

We observe the coefficient changes in Table 4. According to Table 4, we found that variable X1 was removed at node 2, variable X1 appeared at node 4, and variable X1 was removed again at node 5, variable X1 was the disturbance variable in the variable selection of Lasso. As shown in Table 4, the value of parameter β at each node is as follows:

Node 1: The value of parameter β is $(0.923, 1.423, -1.884, 3.166)^T$.

Node 2: The value of parameter β is $(0.000, 1.894, -0.944, 2.267)^T$.

Node 3: The value of parameter β is $(0.000, 1.972, -0.519, 1.837)^T$.

Node 4: The value of parameter β is $(-0.351, 2.181, 0.000, 1.332)^T$.

Node 5: The value of parameter β is $(0.000, 1.824, 0.000, 1.070)^T$.

Node 6: The value of parameter β is $(0.000, 0.754, 0.000, 0.000)^T$.

Figure 2 was drawn using the built-in software lars in R, where variable X1 is the disturbance variable.

TABLE 4. Coeffcient change2 of Lasso.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Node 1	0.923	1.423	-1.884	3.166
Node 2	0.000	1.894	-0.944	2.267
Node 3	0.000	1.972	-0.519	1.837
Node 4	-0.351	2.181	0.000	1.332
Node 5	0.000	1.824	0.000	1.070
Node 6	0.000	0.754	0.000	0.000

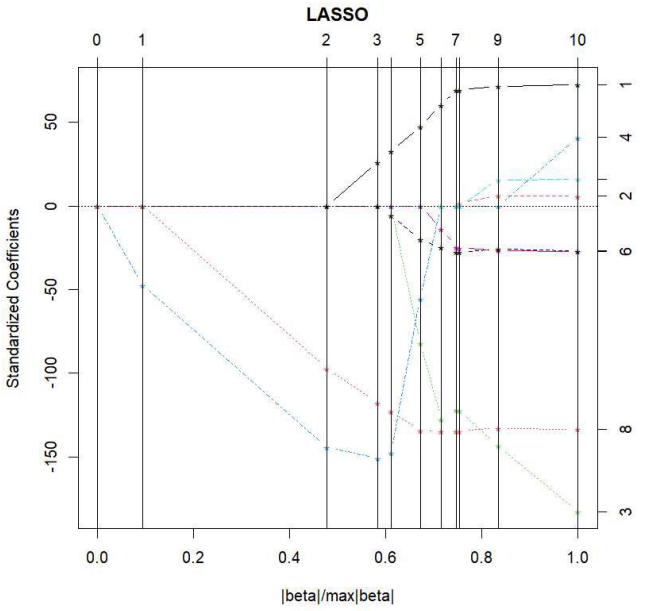

FIGURE 3. Graph3 of Lasso's variable selection.

2) VARIABLE SELECTION OF AIC

The result of computer simulation is "AIC is infinity for this model, so 'step' cannot proceed." That is, AIC cannot be used as a variable selection for this example.

Therefore, the variable selection of Lasso is not consistent with the variable selection of AIC in the example.

C. VARIABLE SELECTION OF DATASET SEATPOS

Example 3: Seatpos is a data set of faraway packages in R, the variable Hipcenter of Seatpos was used as the response variable, and other variables of Seatpos were used as the observed variables, and the variable selection of Lasso and AIC was observed.

The least squares estimation is

$(11.92, 0.94, -30.02, 6.72, 2.63, -4.48, -4.43, -21.92)$.

X^TX is not an identity matrix in the example; this is the problem of Lasso under nonorthonormal design.

1) VARIABLE SELECTION OF LASSO

Figure 3 shows the variable selection of Lasso in Example 3.

Figure 3 is drawn using the built-in software lars in R, and variable X4 is the disturbance variable.

As shown in Figure 3, we found that variable X4(ht) is removed at node 2, and variable X4 appears at node 6, variable X4 is the disturbance variable. As shown in Table 5, the value of parameter β at each node is as follows:

Node 1: The value of parameter β is $(11.92,0.94,-30.02,6.72,2.63,-4.48,-4.43,-21.92)^T$.

Node 2: The value of parameter β is $(11.75,0.99,-23.57,0.00,2.55,-4.37,-4.24,-21.82)^T$.

Node 3: The value of parameter β is $(11.42,0.21,-20.13,0.00,0.00,-4.13,-4.52,-22.13)^T$.

Node 4: The value of parameter β is $(11.34,0.00,-20.01,0.00,0.00,-4.04,-4.52,-22.10)^T$.

Node 5: The value of parameter β is $(9.88,0.00,-20.98,0.00,0.00,-2.29,-4.09,-22.14)^T$.

Node 6: The value of parameter β is $(7.83,0.00,-13.51,-9.11,0.00,0.00,-3.29,-22.04)^T$.

Node 7: The value of parameter β is $(5.37, 0.00, 0.00, -24.21, 0.00, 0.00, -0.94, -22.17)^T$.

Node 8: The value of parameter β is $(4.32, 0.00, 0.00, -24.82, 0.00, 0.00, 0.00, -19.33)^T$.

Node 9: The value of parameter β is $(0.00, 0.00, 0.00, -23.72, 0.00, 0.00, 0.00, -15.95)^T$.

Node 10: The value of parameter β is $(0.00, 0.00, 0.00, -7.77, 0.00, 0.00, 0.00, 0.00)^T$.

2) VARIABLE SELECTION OF AIC

The variable selection of AIC involves seven steps:

Step1: At the beginning, the AIC value was 283.99. After removing variable X8, the AIC value was 283.99, so variable X8 was removed to obtain a model with seven variables.

Step2: As shown in Table 6, the minimum AIC value was 282.00, and variable X2 was removed to obtain a model with six variables.

Step3: As shown in Table 7, the minimum AIC value was 280.01, and variable X4 was removed to obtain a model with five variables.

Step4: As shown in Table 8, the minimum AIC value was 278.10, and variable X7 was removed to obtain a model with four variables.

TABLE 5. Coeffcient change3 of Lasso.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Node 1	11.92	0.94	-30.02	6,72
Node 2	11.75	0.99	-23.57	0.00
Node 3	11.42	0.21	-20.13	0.00
Node 4	11.34	0.00	-20.01	0.00
Node 5	9.88	0.00	-20.98	0.00
Node 6	7.83	0.00	-13.51	-9.11
Node 7	5.37	0.00	0.00	-24.21
Node 8	4.32	0.00	0.00	-24.82
Node 9	0.00	0.00	0.00	-23.72
Nod10	0.00	0.00	0.00	-7.77
	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
Node 1	2.63	-4.48	-4.43	-21.92
Node 2	2.55	-4.37	-4.24	-21.82
Node 3	0.00	-4.13	-4.52	-22.13
Node 4	0.00	-4.04	-4.52	-22.10
Node 5	0.00	-2.29	-4.09	-22.14
Node 6	0.00	0.00	-0.94	-20.04
Node 7	0.00	0.00	-0.94	-20.17
Node 8	0.00	0.00	0.00	-19.33
Node 9	0.00	0.00	0.00	-15.95
Nod10	0.00	0.00	0.00	0.00

Here, Nod 10 denotes Node 10.

TABLE 6. AIC of seven variable.

RV	X2	X4	X7	X3
AIC	282.00	282.01	282.05	282.11
RV	X5	X6	X1	
AIC	282.21	282.57	284.76	

TABLE 7. AIC of six variable.

RV	X4	X7	X3	X5
AIC	280.01	280.05	280.11	280.21
RV	X6	X1		
AIC	280.58	282.86		

TABLE 8. AIC of five variable.

RV	X7	X5	X6	X1	X3
AIC	278.10	278.22	278.65	280.01	282.36

TABLE 9. AIC of four variable.

RV	X5	X6	X1	X3
AIC	276.37	276.82	279.29	281.59

Step5: As shown in Table 9, the minimum AIC value was 276.37, and variable X5 was removed to obtain a model with three variables.

TABLE 10. AIC of three variable.

RV	X6	X1	X3
AIC	275.28	277.41	286.15

TABLE 11. AIC of two variable.

RV	X1	X3
AIC	275.45	312.07

TABLE 12. Coeffcient change1 of forward stagewise.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Node 1	-4.5	0.0	0.0
Node 2	-4.5	0.0	2.5
Node 3	-4.5	1.5	2.5

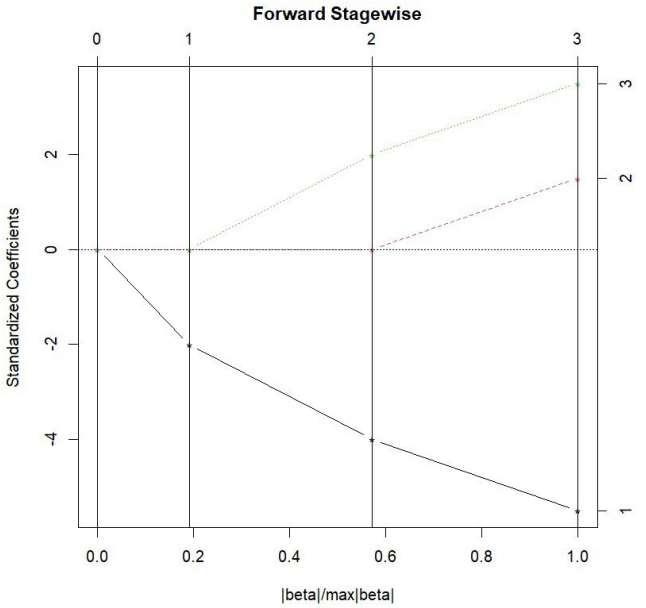

FIGURE 4. Graph1 of forward stagewise.

Step6: As shown in Table 10, the minimum AIC value was 275.28, and variable X6 was removed to obtain a model with two variables.

Step7: As shown in Table 11, The AIC value after removing variables is greater than the original AIC value, and the final model contains two variables.

Therefore, the variable selection of Lasso is not consistent with the variable selection of AIC in the example.

III. VARIABLE SELECTION OF FORWARD STAGEWISE

A. FORWARD STAGEWISE OF EXAMPLE 1

We compare the variable selection of Lasso with the variable selection of the forward stagewise, and point out the complexity of Lasso's variable selection.

We draw forward stagewise's figure of Example 1.

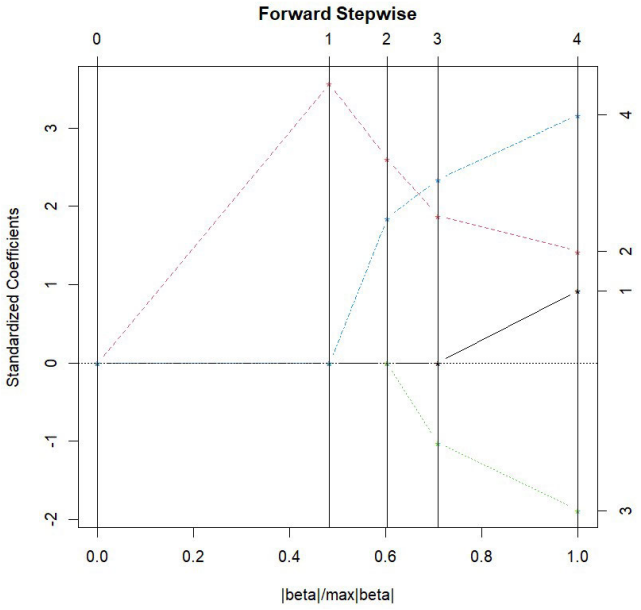

FIGURE 5. Graph2 of forward stagewise.

TABLE 13. Coeffcient change2 of forward stagewise.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Node 1	0.000	3.567	0.000	0.000
Node 2	0.000	2.606	0.000	1.852
Node 3	0.000	1.880	-1.017	2.340
Node 4	0.923	1.423	-1.884	3.166

Table 12 shows the values of parameter β at each node in Figure 4. As shown in Table 12, the value of parameter β at each node is as follows:

Node 1: The value of parameter β is $(-4.5.0, 0.0)^{\mathrm{T}}$.

Node 2: The value of parameter β is $(-4.5, 0, 2.5)^{\mathrm{T}}$.

Node 3: The value of parameter β is $(-4.5, 1.5, 2.5)^{\mathrm{T}}$.

From Figure 4 and 1, the variable selection of forward stagewise is consistent with Lasso under orthonormal design, which shows that the stepwise selection algorithm can solve Lasso.

B. FORWARD STAGEWISE OF EXAMPLE 2

We draw forward stagewise's figure of Example 2.

Table 13 shows the values of parameter β at each node in Figure 5. As shown in Table 13, the value of parameter β at each node is as follows:

Node 1: The value of parameter β is $(0.923, 1.423, -1.884, 3.166)^{\mathrm{T}}$.

Node 2: The value of parameter β is $(0.000, 1.894, -0.944, 2.267)^{\mathrm{T}}$.

Node 3: The value of parameter β is $(0.000, 1.972, -0.519, 1.837)^{\mathrm{T}}$.

Node 4: The value of parameter β is $(-0.351, 2.181, 0.000, 1.332)^{\mathrm{T}}$.

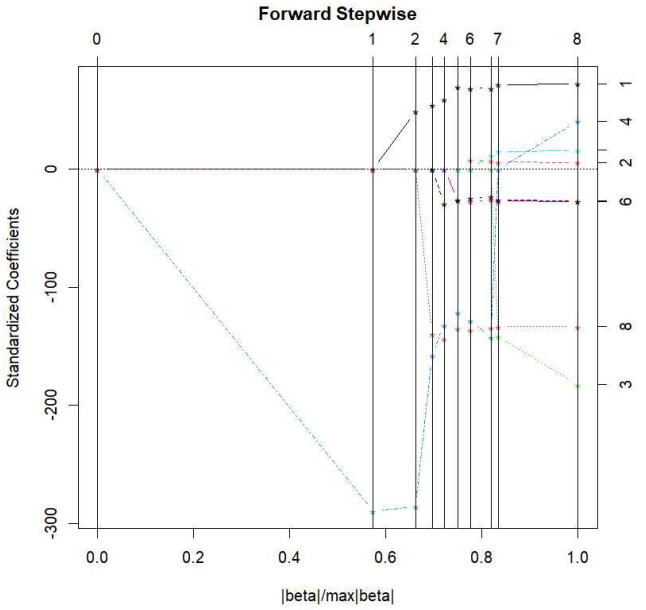

FIGURE 6. Graph3 of forward stagewise.

Node 5: The value of parameter β is $(0.000, 1.824, 0.000, 1.070)^{\mathrm{T}}$.
Node 6: The value of parameter β is $(0.000, 0.754, 0.000, 0.000)^{\mathrm{T}}$.

From Figure 5 and 2, the variable selection of forward stagewise is different from that of Lasso, which shows that the stepwise selection algorithm cannot solve Lasso.

C. FORWARD STAGEWISE OF EXAMPLE 3

We draw the forward stagewise's figure of Example 3.

Table 14 shows the values of parameter β at each node in Figure 6. As listed in Table 14, the value of parameter β at each node is as follows:

Node 1: The value of parameter β is $(0.00, 0.00, 0.00, -47.65, 0.00, 0.00, 0.00, 0.00)^{\mathrm{T}}$.
Node 2: The value of parameter β is $(8.01, 0.00, 0.00, -46.93, 0.00, 0.00, 0.00, 0.00)^{\mathrm{T}}$.
Node 3: The value of parameter β is $(8.93, 0.00, 0.00, -25.98, 0.00, 0.00, 0.00, -22.94)^{\mathrm{T}}$.
Node 4: The value of parameter β is $(9.72, 0.00, 0.00, -21.70, 0.00, 0.00, -4.83, -23.66)^{\mathrm{T}}$.
Node 5: The value of parameter β is $(11.46, 0.00, 0.00, -19.99, 0.00, -4.39, -4.06, -22.36)^{\mathrm{T}}$.
Node 6: The value of parameter β is $(11.26, 1.29, 0.00, -21.04, 0.00, -4.44, -4.06, -22.36)^{\mathrm{T}}$.
Node 7: The value of parameter β is $(11.32, 1.17, 0.00, -23.41, 1.99, -4.28, -3.80, -22.02)^{\mathrm{T}}$.
Node 8: The value of parameter β is $(11.92, 0.94, -30.02, 6.72, 2.63, -4.48, -4.43, -21.92)^{\mathrm{T}}$.

From Figure 6 and 3, the variable selection of forward stagewise is different from that of Lasso, which shows that the stepwise selection algorithm cannot solve Lasso.

TABLE 14. Coeffcient change3 of Lasso.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Node 1	0.00	0.00	0.00	-47.65
Node 2	8.01	0.00	0.00	-46.93
Node 3	8.93	0.00	0.00	-25.98
Node 4	9.74	0.00	0.00	-21.70
Node 5	11.46	0.00	0.00	-19.99
Node 6	11.26	1.29	0.00	-21.04
Node 7	11.32	1.17	0.00	-23.41
Node 8	11.91	0.94	30.02	6.72
	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
Node 1	0.00	0.00	0.00	0.00
Node 2	0.00	0.00	0.00	0.00
Node 3	0.00	0.00	0.00	-22.94
Node 4	0.00	0.00	-4.83	-23.66
Node 5	0.00	-4.39	-4.24	-22.22
Node 6	0.00	-4.44	-4.06	-22.36
Node 7	1.99	-4.28	-3.80	-22.02
Node 8	2.63	-4.48	-4.43	-21.92

The simulation results show that under orthonormal design, the stepwise selection algorithms are consistent with Lasso, and the stepwise selection algorithm can solve Lasso. The variable selection of stepwise selection is inconsistent with Lasso under nonorthogonal design, the stepwise selection algorithm cannot solve Lasso. Therefore, the variable selection of Lasso is very complex.

IV. CONCLUSION

According to the above research, we get the following conclusions.

First: statistical modeling is important. Under orthonormal design, the variable selection of Lasso is consistent with that of AIC and that of forward stagewise, and the stepwise selection algorithms can solve Lasso. Therefore, we attempted to establish a statistical model under orthonormal design to simplify the model we establish easy to calculation. These studies verified the soft threshold estimation of Lasso proposed by Tibshirani in 1996 and the exact solution of Lasso proposed by the author.

Second: the complexity of Lasso's variable selection under nonorthonormal design is highlighted. There was a situation in which the removed variables appeared again in the variable selection of Lasso, but AIC and forward stagewise do not have this property. Computer simulation indicates that the variable selection of Lasso is more complex than that of AIC and forward stagewise. We cannot use the stepwise selection algorithm to solve Lasso under the nonorthonormal design.

Third: the research conclusion of this study can explain the success of ChatGPT. For a specific problem, when the number of parameters is increased, the design matrix becomes sparse. The infinite number of parameters enables the design matrix X to achieve orthonormalization ($XX^{\mathrm{T}} = \mathrm{I}$), so that the stepwise selection algorithm can find the Lasso solution for the large model. This may be the reason for the success of the large model represented by the ChatGPT.

When the parameters and sample size of the large model are both large, the design matrix becomes sparse, and orthogonalization occurs between different column vectors of the large design matrix with a probability of 1.

Let's set the design matrix as:

$$Y = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots \end{pmatrix}_{n\times p}$$

We standardize the elements of the design matrix to obtain a new design matrix that meets the conditions of Lasso. The new design matrix is expressed as follows:

$$X = \begin{pmatrix} \sqrt{\frac{n-1}{n}} & -\sqrt{\frac{1}{n(n-1)}} & -\sqrt{\frac{1}{n(n-1)}} & \cdots \\ -\sqrt{\frac{1}{n(n-1)}} & \sqrt{\frac{n-1}{n}} & -\sqrt{\frac{1}{n(n-1)}} & \cdots \\ -\sqrt{\frac{1}{n(n-1)}} & -\sqrt{\frac{1}{n(n-1)}} & \sqrt{\frac{n-1}{n}} & \cdots \\ -\sqrt{\frac{1}{n(n-1)}} & -\sqrt{\frac{1}{n(n-1)}} & -\sqrt{\frac{1}{n(n-1)}} & \cdots \end{pmatrix}_{n\times p}$$

Based on this new data matrix, it can be concluded that:

$$X^{T}X = \begin{pmatrix} 1 & 1/(n-1) & 1/(n-1) & \cdots \\ 1/(n-1) & 1 & 1/(n-1) & \cdots \\ 1/(n-1) & 1/(n-1) & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ 1/(n-1) & 1/(n-1) & 1/(n-1) & \cdots \end{pmatrix}_{p\times p}$$

$$X^{T}X = I_{p\times p}$$

Therefore, the stepwise selection algorithm can determine the Lasso solution of the large model.

REFERENCES

[1] H. J. Breaux, "On stepwise multiple linear regression," Army Ballistic Res. Lab., Aberdeen Proving Ground, MD, USA, Tech. Rep.

[2] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, Nov. 1995.

[3] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.

[4] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, Apr. 2004.

[5] M. H. R. Khan and J. E. H. Shaw, "Variable selection for survival data with a class of adaptive elastic net techniques," *Statist. Comput.*, vol. 26, no. 3, pp. 725–741, May 2016.

[6] A. Kane and A. Mandal, "A new analysis strategy for designs with complex aliasing," *Amer. Statistician*, vol. 74, no. 3, pp. 274–281, Jul. 2020.

[7] M. Febrero-Bande, W. González-Manteiga, and M. O. D. L. Fuente, "Variable selection in functional additive regression models," *Comput. Statist.*, vol. 34, no. 2, pp. 469–487, Jun. 2019.

[8] H. Xia, "Disturbance of variables in lasso variable selection and influence," in *Proc. ICBDE*, Shanghai, China, Feb. 2022, pp. 434–439.

[9] G. Borboudakis and I. Tsamardinos, "Forward-backward selection with early dropping," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 276–314, 2019.

[10] J. Liang, C. Wang, D. Zhang, Y. Xie, Y. Zeng, T. Li, Z. Zou, J. Ren, and Q. Zhao, "VSOLassoBag: A variable-selection oriented LASSO bagging algorithm for biomarker discovery in omic-based translational research," *J. Genet. Genomics*, vol. 50, no. 3, pp. 151–163, 2023.

[11] L. Wasserman and K. Roeder, "High-dimensional variable selection," *Ann. Stat.*, vol. 37, no. 5A, p. 2718, 2009.

[12] E. Austin, X. Pan, and X. Shen, "Penalized regression and risk prediction in genome-wide association studies," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 6, no. 4, pp. 315–328, 2013.

[13] A. Ahrens and A. Bhattacharjee, "Two-step lasso estimation of the spatial weights matrix," *Econometrics*, vol. 3, no. 1, pp. 128–155, Mar. 2015.

HUIYI XIA received the B.S. and M.S. degrees from the School of Mathematics, Anhui University, Anhui, China, in 1992 and 2006, respectively, and the Ph.D. degree from the School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China, in 2016. He is currently engaged in teaching and scientific research with Chizhou University.