The selective regularization of a linear regression model

V N Lutay¹ and N S Khusainov¹

¹ Institute of Computer Technologies and Information Security, Southern Federal University, Chekhov str., 2, Taganrog, 347922, Russia

Abstract. This paper discusses constructing a linear regression model with regularization of the system matrix of normal equations. In contrast to the conventional ridge regression, where positive parameters are added to all diagonal terms of a matrix, in the method proposed only those matrix diagonal entries that correspond to the data with a high correlation are increased. This leads to a decrease in the matrix conditioning and, therefore, to a decrease in the corresponding coefficients of the regression equation. The selection of the entries to be increased is based on the triangular decomposition of the correlation matrix of the original dataset. The effectiveness of the method is tested on a known dataset, and it is performed not only with a ridge regression, but also with the results of applying the widespread algorithms LARS and Lasso.

1. Introduction

A linear regression model reflects the relationship between the values of observed variables (predictors) and explanatory variables (responses):

\[ y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \epsilon_i, i = 1, n, \quad n \geq m, \]

where \( x_{ij} \) are the observed predictor value \( x_j \), \( y_i \) is the response value, \( \beta_j \) is the regression equation coefficient, \( \epsilon_i \) is the random errors. After centering the observed values \[1\], the linear regression model can be conveniently written down in matrix-vector form

\[ Wb = v, \]

where \( W = (w_{ij})_{i=1,j=1}^{nm} \) is the rectangular matrix of predictor values, \( b = (b_j)_{j=1}^{m} \) and \( v = (v_i)_{i=1}^{n} \) are the vectors of model coefficients and response values, respectively. Conventionally, to obtain values \( b_j \), the least squares (LS) method is used, which is reduced to solving a system called normal equations with a positive definite symmetric matrix:

\[ (W'W)b = W'v, \]

where \( t \) is the transpose symbol. After determining and returning to the observed values, we obtain a linear LS-model:

\[ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_m x_{im}. \]

The efficiency of the model is measured via the sum of the squared deviations of the solution obtained from the response values:
If, after centering, the normalization of the observed variables is performed, then it becomes a matrix of paired correlations of predictors. The diagonal entries of the matrix are equal to 1, all values are between -1 and 1. The values of the off-diagonal terms of the matrix can be used to conclude that the predictors are mutually correlated (collinear). The greater the correlation between the \( i \)-th and the \( j \)-th predictors closer is, will be the value of the off-diagonal entry with the \( ij \) index to 1 and the larger is the matrix condition number and the variance of the response values. If such an entry of the matrix is equal to 1, then the corresponding predictors are linearly dependent and the matrix is singular. The correlation matrix is often used to determine the values of the vector \( \mathbf{b} \).

Algorithms for the analysis and construction of the regression models Ridge, LASSO, LARS are part of machine learning algorithms and are included in the Python, R, Matlab libraries.

In the Ridge model, the diagonal matrix \( \lambda \mathbf{I} \) is added to the matrix \( \mathbf{W}^\top \mathbf{W} \)

\[
(W^\top W + \lambda I) \mathbf{b}_\lambda = W^\top \mathbf{v},
\]

where \( \mathbf{b}_\lambda = [b(\lambda)_{ij}]_{ij=1}^m \), \( \lambda \) is the positive number called a parameter.

Adding a parameter to all diagonal terms of a positive definite symmetric matrix leads to a decrease in its condition number. The values are called the penalty imposed on the coefficients of the model:

\[
P(\lambda) = \lambda \sum_{j=1}^m b_j^2.
\]

The Ridge model takes the following form:

\[
y(\lambda)_i = \beta(\lambda)_i + \sum_{j=1}^m \beta(\lambda)_j x_{ij}, \tag{2}
\]

With an increase in \( \lambda \), the dispersion decreases, but the shift of the regression coefficients \( \beta(\lambda) \) from the values LS increases. The Ridge coefficients can be viewed as linear combinations of coefficients LS.

LASSO [3] consists in the introduction of a constraint the norm of the vector of the coefficients of the model, which leads to the inversion of some of them to 0. The LARS (Least Angle Method) algorithm [4] can reduce the size of the model by analyzing the contribution of predictors to the response similarly to the stepwise regression method, which in [2] is considered to be the best way to build a linear model.

Further, a method for reducing the condition number of a matrix is considered, which differs from the Ridge model in that an increase is only in those diagonal terms of the matrix that cause a significant collinearity of the model, and the added parameters for different terms may be different. This method of regularization will be called selective (SR).

2. Selecting and increasing the diagonal entries of the matrix

We denote the matrix \( \mathbf{W}^\top \mathbf{W} \) by \( \mathbf{A} \) and write down the result of its triangular decomposition by the Cholesky method, which is one of the popular methods for solving LAE systems with a symmetric matrix:

\[
\mathbf{A} = \mathbf{L} \mathbf{L}',
\]

where \( \mathbf{L} \) is the upper triangular matrix.

The condition number of the matrix \( \mathbf{A} \) can be reduced if the condition number of its factors is reduced, since:
cond (A) ≤ cond (L)^2

It was shown in [5] that the lower bound for the condition number of a triangular non-singular matrix has the following form:

\[
\max \left| l_{ii} \right| \leq \frac{ij}{\min |l_{ii}|}.
\]

Thus, to decrease the condition number of the matrix A, one should increase a minimum diagonal element L and should not increase a maximum one. Below are the formulas for calculating the diagonal entries of the triangular matrix L [6]

\[
l_{ii} = \left( a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2 \right)^{1/2}, \quad i = 2, n.
\] (3)

To obtain the coefficients of the system of equations, two systems with triangular matrices are solved

\[
Lz = Wv, \quad L'b = z.
\] (4)

It was shown in [7] that an increase in the i-th diagonal entry of a triangular matrix in the process of a triangular decomposition is equivalent to the fact that the matrix A with an increased a_{ii} undergoes decomposition without changing the other diagonal entries. It follows that if in the process of factorization there was an increase in several diagonal entries of a triangular matrix, then such a decomposition is an exact decomposition of the following complete matrix

\[
LL' = A + S,
\]

where S is the matrix, whose nonzero entries are located on the main diagonal and their indices coincide with the indices of the diagonal entries of the triangular matrix, which increase during the decomposition. The converse is also true: it follows from equation (3) that with an increase in a_{ii} by \( \delta a_{ii} \), the value of \( l_{ii} \) increases by \( \delta l_{ii} \) according to the following relation (all quantities are non-negative):

\[
\delta l_{ii}^2 + 2\delta l_{ii} = \delta a_{ii}.
\]

It follows from the above that increasing \( a_{ii} \) can reduce the number of conditioning A. Let us consider choosing a suitable \( a_{ii} \). For this purpose, we will turn to the result of the decomposition of the matrix according to Cholesky. As an example, below are the matrices corresponding to n=3, and in equation (5) - expressions for the diagonal entries of the triangular matrix

\[
\begin{bmatrix}
0 & 0 & 0 \\
0 & 0 & l_{32} \\
0 & 0 & l_{33}
\end{bmatrix}
= \begin{bmatrix}
a_{11} & a_{12} & a_{13} \\
a_{21} & a_{22} & a_{23} \\
a_{31} & a_{32} & a_{33}
\end{bmatrix},
\]
\[ l_{11} = 1, \]
\[ l_{22}^2 = a_{22}^2 - a_{12}^2, \]
\[ l_{33}^2 = a_{33}^2 - a_{13}^2 - (a_{23} - a_{12}a_{13})^2 / l_{22}^2. \]

It follows from the above expressions that if the diagonal entries of \( A \) are equal to 1 (as in the correlation matrix), then the value \( l_{ii} \) depends on the entries of the matrix \( A \) in the \( i \)-th column (\( a_{12} \) and \( a_{13} \), respectively): if the \( i \)-th column of the correlation matrix contains a large number, then \( l_{ii} \) is small. An increase in this term leads, according to equation (4), to a decrease in the corresponding coefficient of the regression equation and, therefore, to a decrease in the influence of the corresponding predictor on the response.

After increasing one or more diagonal entries of the correlation matrix, selected according to the corresponding triangular matrix, by analogy with equation (1), we can write down
\[
(W'W + S)b_s = W'v.
\]

The regression model in this case looks as follows
\[
y_{ij}(s) = \sum_{j=0}^{m} \beta_{ij}(s)x_{ij}, i = 1, n.
\]

Suppose that the number of diagonal terms in \( S \) is equal to \( k \). Then the penalty is imposed not on all parameters of the model, but only on \( k \) of them:
\[
P(s) = \sum_{j=1}^{k} s_{jj}b_{jj}^2
\]

3. The computational experiments
To analyze the effectiveness of the method proposed, we used Hald's data, analyzed in detail in [2]. In this dataset, the number of observations is 13, the number of predictors is 4. The correlation matrix is as follows (all calculations and graphing are done in Python):
\[
\begin{bmatrix}
0.999 & 0.228 & -0.824 & -0.245 \\
0.228 & 1.0 & -0.139 & -0.972 \\
-0.824 & -0.139 & 0.999 & 0.029 \\
-0.245 & -0.972 & 0.029 & 1.0
\end{bmatrix}
\]

It follows from the form of the matrix that the cross-correlation is high between the 1st and the 3rd predictors and between the 2nd and the 4th predictors. The values of the diagonal entries of the triangular matrix obtained during its decomposition are as follows
\[
l_{11} = 0.999, l_{22} = 0.973, l_{33} = 0.564, l_{44} = 0.059.
\]

Thus, to reduce the condition number of the matrix, one should increase its 3rd and 4th diagonal entries. (When performing calculations, it should be borne in mind that with an increase in the value of \( l_{ii} (i > 1) \) by increasing \( a_{ii} \), the off-diagonal terms \( l_{ii} \) decrease, and the diagonal \( l_{jj} (j > i) \) increases. Therefore, it is advisable to rearrange the predictor columns so that the diagonal entries of the corresponding triangular matrix do not increase. In our case, this condition is fulfilled.)

Table 1 shows the characteristics of the SR algorithm in comparison with the Ridge regression with an increase in the parameter \( \lambda \) and the corresponding diagonal entries of the matrix.
The 2nd column of the table contains the condition numbers of the matrix, the 3rd is the values of the parameter $\lambda$ corresponding to these numbers. Next are the columns of the values added to the 3rd and 4th diagonal entries of the original correlation matrix, and the columns R and SR show the values of the shift of the mathematical expectation for Ridge regression and method proposed.

Figure 1 shows the graphs of changes $\beta_3(s)$ from equation (6) (solid lines) and $\beta_4(\lambda)$ from equation (2) (dashed lines) depending on a decrease in the condition number (the numbers on the axis x correspond to the first column of table 1). The figure shows that $\beta_3(s)$ and $\beta_4(s)$ have lower absolute values than $\beta_3(\lambda)$ and $\beta_4(\lambda)$. The values $\beta_0$, not shown in figure 1 decrease with decreasing the condition number from 55 to 51 for SR and increase from 75 to 89 for the Ridge regression.

![Figure 1](image)

**Figure 1.** Regression coefficients for Ridge regression and SR.

The optimal value of $\lambda$ chosen in [2] for the Hald data is 0.013. In table 1, this value corresponds to point 3; the values of the regression coefficients for Ridge regression and SR are summarized in table 2. It follows from the last column that SR decreases in due to the fact that the fourth predictor is the most correlated one.

---

**Table 1.** Comparative characteristics of selective regularization and Ridge regression.

| № | $\text{cond}(A)$ | $\lambda$ | $\delta a_{33}$ | $\delta a_{44}$ | $\text{Estimate deviation}$ |
|---|------------------|------------|-----------------|-----------------|-----------------------------|
|   |                  |            |                 |                 | R  | SR |
| 1 | 1400             | 0          | 0               | 0               | 0  | 0  |
| 2 | 400              | 0.004      | 0.005           | 0.007           | 0.26 | 0.15 |
| 3 | 140              | 0.015      | 0.01            | 0.03            | 0.33 | 0.21 |
| 4 | 60               | 0.038      | 0.05            | 0.07            | 0.35 | 0.21 |
| 5 | 18               | 0.12       | 0.1             | 0.3             | 0.38 | 0.23 |
| 6 | 8                | 0.35       | 0.5             | 0.8             | 0.42 | 0.23 |
| 7 | 7                | 0.45       | 0.7             | 1               | 0.44 | 0.23 |
| 8 | 7                | 0.55       | 0.8             | 1.1             | 0.45 | 0.23 |

The optimal value of $\lambda$ chosen in [2] for the Hald data is 0.013. In table 1, this value corresponds to point 3; the values of the regression coefficients for Ridge regression and SR are summarized in table 2. It follows from the last column that SR decreases in due to the fact that the fourth predictor is the most correlated one.
**Table 2.** Regression coefficients for $\lambda = 0.013$.

|       | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ |
|-------|-----------|-----------|-----------|-----------|-----------|
| SR    | 50.746    | 1.674     | 0.629     | 0.227     | -0.0268   |
| Ridge | 83.437    | 1.308     | 0.297     | -0.133    | -0.351    |

Figure 2 shows the dependences of the SSE values with a decrease in the condition number of the matrix. The SSE values for the Ridge regression sharply increase with increasing a regularization parameter.

![Figure 2](image_url)

**Figure 2.** The SSE values for the Ridge regression (a dashed line) and for SR (solid).

Figure 3 shows the results of comparing SR with the values obtained using the linear_model.LassoLars function from the Python sklearn library. This function jointly implements the LASSO and LARS algorithms; further in the text this function is called Lars. The units of the axis $x$ in this and in the following figure are the numbers given in table 3. Each number corresponds to the value of the parameter $\alpha$ in Lars and the values of $\delta a_{33}$ and $\delta a_{44}$ in SR. Since their initial value in table 3 is the last value of table 1, the axis $x$ in figures 3 and 4 can be viewed as a continuation of the same axis in figures 1 and 2.

**Table 3.** Correspondence of the Lars parameter and the values of the 3rd and the 4th diagonal entries of the matrix.

| №   | 8   | 9   | 10  | 11  | 12  | 13  | 14  | 15 |
|-----|-----|-----|-----|-----|-----|-----|-----|----|
| $\alpha$ (Lars) | 0.01 | 0.015 | 0.018 | 0.02 | 0.025 | 0.03 | 0.035 | 0.05 |
| $\delta a_{33}$ | 1.1 | 1.2 | 1.5 | 1.8 | 2.2 | 3.0 | 3.3 | 4.0 |

The coefficients $\beta_1$ and $\beta_4$ for Lars, starting from point 10, are equal to 0. In the SR, these coefficients are nonzero, but small in absolute value. With increasing $a_{33}$ and $a_{44}$, their absolute values decrease; at point 14 in comparison with the values $\beta_1$ and $\beta_2$ becoming insignificant and may not be taken into account in the regression equation.
Figure 3. The regression coefficients Lars and SR.

Figure 4. The SSE values for Lars (a dashed line) and SR (solid).

Figure 4 shows the SSE changes for Lars and SR. It follows that SSE at point 14, in which for SR $\beta_3$ and $\beta_4$ were taken equal to 0, is less than for Lars at point 10.

4. Conclusion
The results obtained are as follows:

- it is shown that an increase in several diagonal terms of the correlation matrix or the matrix of normal equations leads to an increase in the stability of the regression equation in the same way as in the Ridge regression with an increase in all diagonal entries;
- unlike Ridge, an increase in diagonal entries leads to a decrease in the values of the regression coefficients corresponding to collinear predictors;
- the efficiency of the method in terms of reducing the size of the model is comparable to the well-known LARS and Lasso algorithms; in this case, the method is much simpler and on the presented experimental data shows a better result in terms of a standard deviation.

References
[1] Yan X and Su X G 2009 Linear Regression Analysis Theory and Computing (World Scientific Publishing Company)
[2] Draper N and Smith H 2017 Applied Regression Analysis 2nd Edition (John Wiley & Sons)
[3] Trevor H, Tibshirani R and Wainwright M 2015 Statistical Learning with Sparsity: The Lasso and Generalizations (London: Chapman and Hall)
[4] Efron B, Hastie T, Johnstone J and Tibshirani R 2004 Least angle regression. *The Annals of Statistics* **32** 407-99

[5] Lawson C and Hanson R 1974 *Solving Least Squares Problems* (Prentice-Hall, Inc., Englewood Cliffs, N. J.).

[6] Voevodin V and Kuznetsov Y 1984 *Matrices and Calculations* (Moscow: Nauka)

[7] Lutay V 2019 Improving the stability of triangular decomposition of ill-conditioned matrices *Numerical Analysis and Applications*. **4** 388-94