Int. J. Nonlinear Anal. Appl. 14 (2023) 1, 245–250
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.26998.3470

Penalized least squares optimization problem for high-dimensional data

Mahdi Roozbeh[a,*], Monireh Maanavi[a], Nur Anisah Mohamed[b]

[a] *Department of Statistics, Faculty of Mathematics, Statistics and Computer Sciences, Semnan University, P.O. Box 35195-363, Semnan, Iran*

[b] *Institute of Mathematical Sciences, Faculty of Science, Universiti Malaya, 50603, Kuala Lumpur, Malaysia*

(Communicated by Saman Babaie-Kafaki)

Abstract

In many applications, indexing of high-dimensional data has become increasingly important. High-dimensional data is characterized by multiple dimensions. There can be thousands, if not millions, of dimensions in applications. Classic methods cannot analyse this kind of data set. So, we need the appropriate alternative methods to analyse them. In high-dimensional data sets, since the number of predictors is greater than the sample size, it is generally impossible to apply classical methods to fit a efficient model. A popular method for combating the challenge of the high-dimensionality curse is to solve a penalized least squares optimization problem, which combines the residual sum of squares loss function measuring the goodness of the fitted model to the data sets with some penalization terms that promote the underlying structure. So, the penalized methods can analyse and provide a good fit for the high-dimensional data sets. In this paper, we express some of these approaches and then, apply them to the eye data set for investigating the computational performance of the proposed methods.

Keywords: High-dimensional data, Lasso, Ridge, Elastic Net
2010 MSC: Primary 62J05; Secondary 62J20, 90C11, 90C59

1 Introduction

Linear regression is one of the most popular modeling approaches as it gives often useful and interpretable insight into the data. For the first time, regression was introduced by Francis Galton's efforts in year 1877. In fact linear models were largely developed in the precomputer age of statistics, but even in today's computer era there are still good reasons to study and use them [4]. They are simple and often provide an adequate and interpretable description of how the inputs affect the output. They can solve important problem easily. For prediction purposes they can sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data. Finally, linear methods can be applied to transformations of the inputs and this considerably expands their scope.

Least square colorblue(LS) method is the best way to solve linear regression problems. In 1809 Carl Friedrich Gauss (who was a German mathematician) published method of Least Square [6]. He In 1822, was able to the best

*Corresponding author
Email addresses: mahdi.roozbeh@semnan.ac.ir (Mahdi Roozbeh), m.maanavi95@gmail.com (Monireh Maanavi), nuranisah_mohamed@um.edu.my (Nur Anisah Mohamed)

Received: April 2022 *Accepted:* August 2022

linear unbiased estimator colorblue(BLUE) of the coefficients is the LS estimator [7]. This result is known as the Gauss-Markov theorem. The Gauss-Markov theorem states that, under classic assumption, the ordinary least squares method, in linear regression models, provides BLUE. This theorem gives the most feature of LS method that is very important to users.

In fact, the least square method is the process of finding the best-fitting curve or line of best fit for a set of data points by reducing the sum of the squares of the offsets (residual part) of the points from the curve. In least square method we look for the optimal line. In other words, we need to find the best line with least error to fit.
A linear regression model is generally in the following form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ is a vector of predictor variable, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top \in \mathbb{R}^{n\times p}$ is a matrix of observations on the explanatory variables, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)^\top$ is a vector of unknown regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$ is a vector of error terms with $E(\varepsilon) = \mathbf{0}$ and $E(\varepsilon\varepsilon^\top) = \sigma^2\mathbf{I}_n$, where $\mathbf{I}_n$ is the unit matrix of order n and σ^2 is an unknown constant parameter. The ordinary least-squares estimator (OLSE) of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \mathbf{S}^{-1}\mathbf{X}^\top\mathbf{y} = \arg\min_{\boldsymbol{\beta}} \mathrm{RSS}(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{1.2}$$

where $\mathbf{S}^{-1} = \mathbf{X}^\top\mathbf{X}$.

In recent years, however, the structure of the data sets changed. In fact, the amount of data we are faced with keeps growing. From around the late 1990s we started to see wide data sets, where the number of variables far exceeds the number of observations [2]. This was largely due to our increasing ability to measure a large amount of information automatically. A problem that is happen in this situation is curse of dimensionality, a term usually attributed to Bellman [1]. Roughly speaking, this means that estimation gets harder very quickly as the dimension of the observations increases. There are at least two versions of this curse. The first is the computational curse of dimensionality. This refers to the fact that the computational burden of some methods can increase exponentially with dimension. The second version, which call the statistical curse of dimensionality: if the data have dimension d, then we need a sample size n that grows exponentially with d. While earlier the number of observations, n, usually clearly exceeded the number of explaining variables, p, nowadays often $p \approx n$ or even $p > n$ which is known as high-dimensional data. For example, in genomics we can use a high-throughput experiment to automatically measure the expression of tens of thousands of genes in a sample in a short amount of time. Similarly, sequencing equipment allows us to genotype millions of SNPs (single-nucleotide polymorphism) cheaply and quickly. In document retrieval and modeling, we represent a document by the presence or count of each word in the dictionary. This easily leads to a feature vector with 20,000 components, one for each distinct vocabulary word, although most would be zero for a small document. If we move to bi-grams or higher, the feature space gets really large [2]. When we face with these data sets we need to analysis them somehow. We cannot use LS method because of interpretation [5]. This data set has the number of variables far exceeds the number of observations. In this situation the design matrix will not be full rank, therefore, with a large number of predictors. So $(\mathbf{X}^\top\mathbf{X})^{-1}$ cannot be calculated. In this situation, we often would like to determine a smaller subset that exhibit the strongest effects. In order to get the "big picture" we are willing to sacrifice some of the small details approaches to analysis the high-dimensional data sets.

The remainder of this study is organized as follows: Penalized least squares approach will be proposed in Section 2. Then, the real data analyses will be discussed in Section 3. Finally, conclusions as well as some directions for future research are introduced in Section 4.

2 Penalized least squares

One of the best ways to analysis high-dimensional data sets is penalized least squares. Penalized least squares has a big class of different methods to analysis varied high-dimensional data sets. A large amount of effort has gone into the development of penalized least squares methods for simultaneous variable selection and coefficient estimation. These methods have many applications in various scientific fields and they are really famous that everyone knows them. In this paper we describe a number of approaches. Penalized least squares improves upon least-squares. The basic idea of penalized least squares is constrain or "shrink" parameter estimates. Penalization is also known as regularization. Regularization reduces variance and increases bias. Testing performance of them can be improved by regularizing an appropriate amount (due to bias-variance tradeoff). In penalized regression, we minimize the residual sum squares (RSS) as $\mathrm{RSS}(\boldsymbol{\beta}) + \lambda Penalty(\boldsymbol{\beta})$, where $\lambda \geq 0$ is the regularization parameter and the penalty function can take various forms. Increasing λ will increase bias and decrease variance. Likewise, decreasing λ reduces bias and increases variance. A big part of the building, the best models in LASSO deals with the bias-variance tradeoff. Bias refers to how correct (or incorrect) the model is. There are several ways to choose the optimal λ, such as AIC, BIC, Cp and so on. For this purpose, one of the most popular methods is the cross-validation (CV) method [12]. More generally, in penalized maximum likelihood, we minimize $L(\boldsymbol{\beta}) + \lambda Penalty(\boldsymbol{\beta})$, where $L(.)$ is loss function. The penalty functions have to be singular at the origin to produce sparse solutions (many estimated coefficients are zero), to satisfy certain conditions to produce continuous models (for stability of model selection), and to be bounded by a constant to produce nearly unbiased estimates for large coefficients. A good penalty function should result in an estimator with three properties [3]:

1. Unbiasedness: The resulting estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modeling bias.
2. Sparsity: The resulting estimator is a thresholding rule, which automatically sets small estimated coefficients to zero to reduce model complexity.
3. Continuity: The resulting estimator is continuous in data z to avoid instability in model prediction. Before starting next part we need to define L_p norm.

Definition 2.1. For any arbitrary vector like $\mathbf{x} = (x_1, \ldots, x_n)$, L_p norm is captured by the formula:

$$||\mathbf{x}||_p := (x_1^p, \ldots, x_n^p)^{1/p}$$

2.1 The Ridge method

In practice, even if the sample size is small, a large number of predictors is typically included to mitigate modeling biases [8]. With such a large number of predictors, there might exist problems among explanatory variables, in particular, there could be a problem with multicollinearity (linear correlation between input variables. So, we need to apply a useful way to overcome this situation. The Ridge method can help us. Tikhonov regularization, named for Andrey Tikhonov (who was Russian mathematician), is the most commonly used method of regularization of ill-posed problems. In statistics, the method is known as ridge regression. Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of squares, or we can write

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}||_2,$$

where λ is regularization parameter that some users call it “ridge parameter”. Also $||.||_2$ is L_2 norm. Ridge estimation has continuity property but it is bias.

When we say that ridge is a continuous process, it means that model fitting takes place in a continuous space ($\boldsymbol{\beta}$ is a smooth and continuous function of λ). In the other words, changing λ a little bit, the fitted model will change a little bit. Hence, there are an infinite number of possible models and so, the ridge penalty space is a continuous region. Also ridge estimation is not a sparse method. To end this, consider a regression model through the origin with one independent variable. So, it can be written

$$\hat{\beta}_{Ridge} \quad = \quad \arg\min\nolimits_\beta \Big\{ \overbrace{\sum_{i=1}^n (y_i - x_i\beta)^2 + \lambda\beta^2}^{\phi} \Big\} = \sum_{i=1}^n y_i^2 - 2\beta \sum_{i=1}^n x_i y_i + \beta^2 \sum_{i=1}^n x_i^2 + \lambda\beta^2,$$

now, we have $\frac{\partial\phi}{\partial\beta} = -2\sum_{i=1}^n x_i y_i + 2\beta\sum_{i=1}^n x_i^2 + 2\lambda\beta = 0$, which yields $\hat{\beta}_{Ridge} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2 + \lambda}$, so, it is clear that the denominator of the above fraction never tends to infinity for the finite values of shrinkage parameter.

2.2 The LASSO method

LASSO, short for Least Absolute Shrinkage and Selection Operator, is a statistical formula whose main purpose is the feature selection and regularization of data models. The method was first introduced in 1996 by Statistics Professor Robert Tibshirani [14]. LASSO introduces parameters to the sum of a model, giving it an upper bound that acts as a constraint for the sum to include absolute parameters within an allowable range. The LASSO method regularizes model parameters by shrinking the regression coefficients, reducing some of them to zero. In other word, with a large number of predictors there is often a desire to select a smaller subset that not only fits as well as the full set of variables, but also contains the more important predictors. Such concerns have led to prominent development of least squares regression methods with various penalties to discover relevant explanatory factors and to get higher prediction accuracy in linear regression.

LASSO is an extension of OLS which adds a penalty to the RSS equal to the sum of the absolute values of the non-intercept beta coefficients multiplied by parameter λ that slows or accelerates the penalty. That is, if λ is less than 1, then it slows the penalty while if it is more than 1, it accelerates the penalty. Therefore, the following optimization problem should be solved:

$$\min_{\boldsymbol{\beta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\lambda||\boldsymbol{\beta}||_1,$$

where λ is regularization parameter and $||.||_1$ is L_1 norm. Computation of the solution to above equation is a quadratic programming problem with linear inequality constrain.

2.3 The Elastic Net Method

Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models. The elastic net method improves lasso's limitations, i.e., where lasso takes a few samples for high dimensional data. The elastic net procedure provides the inclusion of number of variables until saturation. If the variables are highly correlated groups, lasso tends to choose one variable from such groups and ignore the rest entirely. This method introduced by Zou and Hastie [15]. The following optimization problem should be solved to find the results of the elastic net method:

$$\min_{\boldsymbol{\beta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\lambda_1||\boldsymbol{\beta}||_1+\lambda_2||\boldsymbol{\beta}||_2, \quad \lambda_1\geq 0,\lambda_2\geq 0$$

As it can be seen, elastic net method has two regularization parameters and is a sparse model.

3 Numerical study

In this section, we analysis eye tissue samples to apply introduced method. In this data set, based on $n = 120$ rates, there exist the expression level of TRIM32 gene and 200 explanatory variables measuring the gene probes [13]. We use R software to analysis our dada. Figure 1 shows CV method to compute optimal value of regularization parameters in Ridge and LASSO Methods. The CV function is used to obtain the optimal regularization parameter which is defined as follows

$$\text{CV}=\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{(-i)}\right)^2,$$

where $\hat{\boldsymbol{\beta}}^{(-i)}$ obtained by omitting the ith pairs $(\mathbf{x}_i, y_i)$. The plots of coefficients in ridge and LASSO are displayed in Figure 2. Figure 3 shows CV method to compute optimal value of one of the regularization parameters in elastic net method. Table 1 shows the number of nonzero coefficients, optimal value of regularization parameter, RSS, RMSE= $\frac{1}{n}\sqrt{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$ and R–squared= $\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$ for ridge, LASSO, elastic net methods. From the table, it can be seen that elastic net method performs better than the other methods in the sense of RMSE and R–squared criteria values in the high-dimensional data set. The best values were denoted in bold font.

Table 1: Results of Ridge, LASSO and elastic net Methods.

Method	Ridge	LASSO	Elastic net
Regularization parameter	1.151633	0.007066311	0.9526316, 0.01069081
Number of nonzero coefficients	200	21	155
RMSE	0.09791446	0.08575985	**0.07641195**
R–squared	0.2794082	0.4832286	**0.5720897**

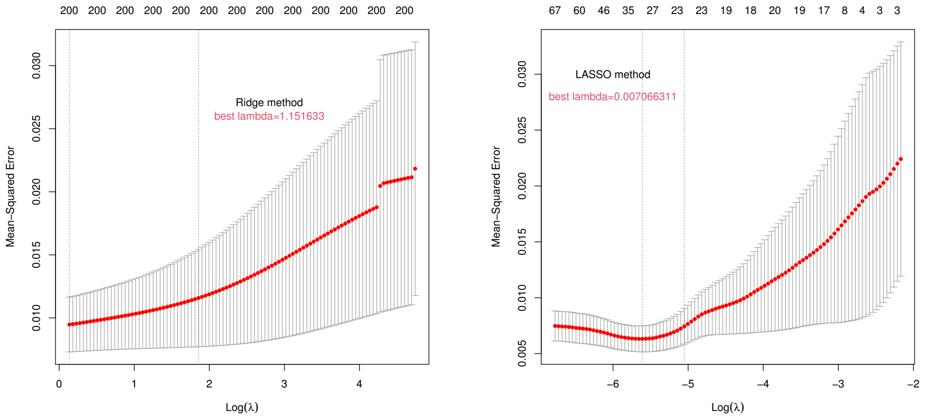

Figure 1: CV method for Ridge and LASSO methods.

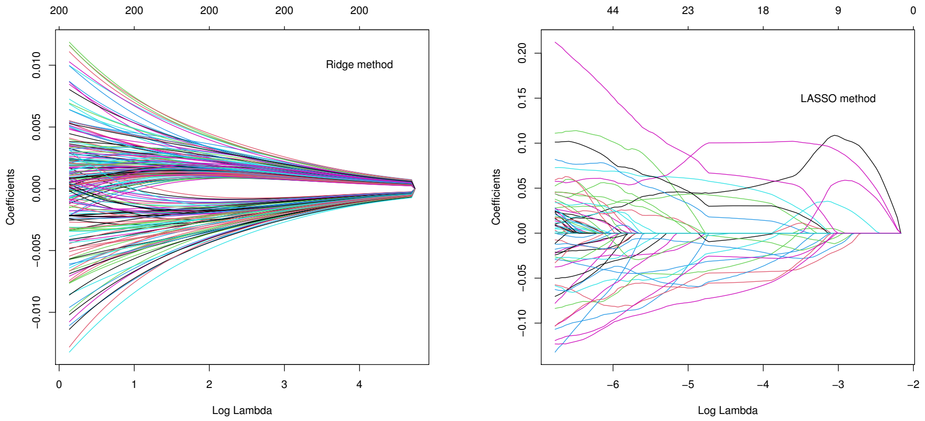

Figure 2: Plot of coefficients in ridge and LASSO.

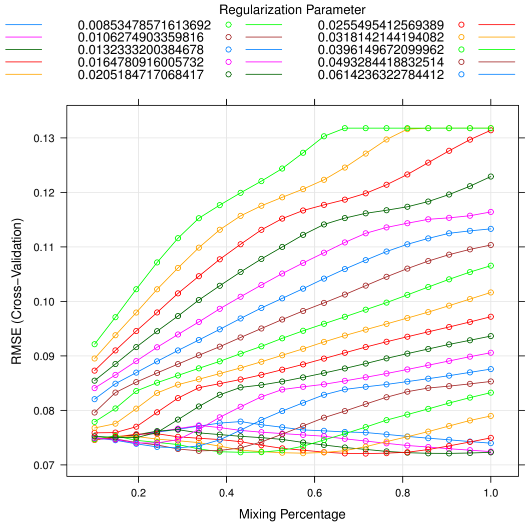

Figure 3: CV method for elastic net.

Acknowledgments

The authors are very grateful to the anonymous reviewers, associate editor and editor-in-Chief for their valuable and constructive comments and suggestions which significantly improved the paper. This research is funded by Universiti Malaya Research Grant (GPF031B–2018).

References

[1] R. Bellman, *On adaptive control processes*, IRE Trans. Automatic Control **4** (1959), 1–9.

[2] B. Efron and T. Hastie, *Computer age statistical inference: Algorithms, evidence and data science*, Springer, Cambridge University Press, 2016.

[3] J. Fan and L. Runze, *Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties*, J. Amer. Statist. Assoc. **96** (2001) 1348–1360.

[4] P. Filzmoser and K. Nordhausen, *Robust linear regression for high-dimensional data: An overview*, Wiley Interdiscip. Rev.: Comput. Statist. **13** (2021), 1–18.

[5] R. Tibshirani, T. Hastie and J.H. Friedman, *The element of statistical learning: Data mining, inference and prediction*, Second edition, Springer, New York, 2017.

[6] C.F. Gauss, *Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore Carolo Friderico Gauss. sumtibus Frid*, Perthes et IH Besser, 1809.

[7] C.F. Gauss, *Theoria combinationis observationum erroribus minimis obnoxiae (Vol. 2)*, H. Dieterich., 1823.

[8] M. Roozbeh, *Robust ridge estimator in restricted semiparametric regression models*, J. Multivar. Anal. **144** (2016) 127–144.

[9] M. Roozbeh, S. Babaie-Kafaki and Z. Aminifard, *Two penalized mixed–integer nonlinear programming approaches to tackle multicollinearity and outliers effects in linear regression models*, J. Ind. Manag. Optim. **17** (2021a), 3475–3491.

[10] M. Roozbeh, S. Babaie-Kafaki and Z. Aminifard, *A nonlinear mixed–integer programming approach for variable selection in linear regression model*, Commun. Statist. Simul. Comput., In press https://doi.org/10.1080/03610918.2021.1990323.

[11] M. Roozbeh, S. Babaie-Kafaki and Z. Aminifard, *Improved high-dimensional regression models with matrix approximations applied to the comparative case studies with support vector machines*, Optim. Meth. Software, In press, https://doi.org/10.1080/10556788.2021.2022144.

[12] M. Roozbeh, M. Maanavi and S. Babaie-Kafaki, *Robust high-dimensional semiparametric regression using optimized differencing method applied to the vitamin B2 production data*, Iran. J. Health Sci. **8** (2020), 9–22.

[13] T.E. Scheetz, K.Y.A. Kim, R.E. Swiderski, A.R. Philp, T.A. Braun, K.L. Knudtson, G.F. DiBona, J. Huang, T.L. Casavant and V.C. Sheffield, *Regulation of gene expression in the mammalian eye and its relevance to eye disease*, Proc. Nat. Acad. Sci. **103** (2006), 14429–14434.

[14] R. Tibshirani, *Regression Shrinkage and Selection via the Lasso*, J. Royal Statist. Soc. Ser. B **58** (1996), 267–288.

[15] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. Royal Statist. Soc. Ser. B **67** (2005), 301–320.