İstatistik Araştırma Dergisi
Journal of Statistical Research
ISSN: 2791-7614

(2022), 12 (2)
Başvuru/Received: 18/10/2022
Kabul/Accepted: 12/12/2022

https://journal.tuik.gov.tr/

ARAŞTIRMA MAKALESİ RESEARCH ARTICLE

A Comparison on Performances of Differential Evolution Algorithm and Genetic Algorithm in Determining the Biasing Parameter *k* of Ridge Regression

Vedide Rezan USLU
Professor at OMÜ, Department of Statistics
rezzanu@omu.edu.tr

Mehmet Arif DEMİRCİ
Lecturer at Muş Alparslan University
Department of Health Management/ Department of Health Systems Management
ma.demirci@alparslan.edu.tr

Abstract

Ridge Regression is a very common way of the remedies for dealing with the “multicollinearity problem” in multiple regression analysis. Although it can provide much more consistent estimates than the ordinary least squares does, there is still a problematic issue in the use of Ridge Regression, which is the choice of biasing parameter k. In this study we propose the use of some Artificial Intelligence Algorithms, such as genetic and differential evolution, for choosing the optimal k value by not allowing to increase too much the mean absolute prediction error while reducing the variation inflation factors and condition number.

Keywords: Ridge Regression; Genetic Algorithm, Differential Evolution Algorithm, MAPE, Variance İnflation Factor, Condition Number

Corresponding Author / Sorumlu Yazar: 1-Vedide Rezan USLU, Ondokuz Mayıs University, Faculty of Art and Sciences, Department of Statistics

2-Mehmet Arif DEMİRCİ, Muş Alparslan University, Department of Health Management

Citation / Atıf: USLU V.R., DEMİRCİ M.A. (2022). A Comparison on Performances of Differential Evolution Algorithm and Genetic Algorithm in Determining the Biasing Parameter *k* of Ridge Regression. İstatistik Araştırma Dergisi, 12 (2), 26-38.

Citation / Atıf: USLU V.R., DEMİRCİ M.A. (2022). A Comparison on Performances of Differential Evolution Algorithm and Genetic Algorithm in Determining the Biasing Parameter *k* of Ridge Regression. İstatistik Araştırma Dergisi, 12 (2), 26-38.

Ridge Regresyonda Yanlılık Parametresi k'nın Belirlenmesinde Genetik ve Differansiyel Gelişim Algoritmalarının Performanslarına Dair Bir Karşılaştırması

Özet

Çoklu regresyonda karşılaşılan "çoklubağlantı" problem için en yaygın olarak önerilen yaklaşım Ridge Regresyondur. Ridge regresyon en küçük kareler yönteminden daha tutarlı tahminler sağlamasına rağmen yanlılık partametresi k'nın belirlenmesi hala çözülmesi gereken bir meseledir. Bu çalışmada optimal k değerini bulmak için Yapay Zeka Tekniklerinden olan Genetik Algoritma ve Diferansiyel Gelişim Algoritması 'nın kullanımı önerilmiştir. Bu yaklaşımların uygulanmasında varyans büyütme faktörü ile şartlı sayı gibi çoklubağlantı probleminin teşhisinde kulanılan göstergeler küçültülmeye çalışılırken ortalama mutlak yüzdelik hatanın çok büyümemesini kontrol altında tutarak algoritmalar geliştirilmiştir.

Anahtar kelimeler: Ridge Regresyon; Genetik Algoritma, Differansiyal Gelişim Algoritması, Ortalama Mutlak Yüzdelik Hata, Varyans Büyütme Faktörü, Şartlı Sayı

1. Introduction

Multiple Regression analysis is a most powerful statistical tool for evaluating the relationship between the dependent variable and the explanatory variables. If it is believed that true relationship between the dependent variable and the independent variables is linear then the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

, which is called a multiple regression model, can be used. This model makes some assumptions on the random error term. These assumptions are that errors have zero mean, $E(\varepsilon_i) = 0$, and constant variance $Var(\varepsilon_i) = \sigma^2$, and are mutually uncorrelated $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$. One another important assumption of the multiple regression model is that there are no any severe or exact linear independencies among the explanatory variables. When you study the real life problems with too many explanatory variables, the linear dependencies among the independent variables can be inevitable. In that case ridge regression can provide biased but much more consistent estimates. This technique is firstly introduced by Hoerl and Kennard (1970 a,b). They provided to decrease the variances of the parameters estimates by adding a positive small number to the diagonal element of the design matrix. Since then many researchers worked on it. As mentioning very briefly these are Hoerl A.E., Kennard R.W., Baldwin K.F. (1975), Hoerl A.E., Kennard R.W. (1976), Vinod H.D. (1976), Gibbons D.G. (1981). There are some other papers which contributes on the choice of k value. For example; Mardikyan S., Çetin E. (2008), Praga-Alejo et al (2008), Ahn, J.J, et al (2012), Khalaf G. And Shukur G.(2005), Kibria B.M.G. (2003), Muniz G. Et all (2012). Uslu V. R et al (2014).

2. Multicollinearity

Multicollinearity is one of the serious problems in multiple regression analysis and depicts a condition in which two or more explanatory variables in the multiple regression model are highly linearly related with one another. Since the case happens, the struggling with this problem is very important. In multiple regression analysis the multicollinearity problem is defined as follows;

Let $X_1, X_2, \ldots, X_p$ be explanatory variables and $a_1, a_2, \ldots, a_p$ scalars which at least one of a_j is not zero. If the relationship

$$a_1 X_1 + a_2 X_2 + \cdots + a_p X_p \cong 0$$

exits, where $\cong$ denotes approximate equality, then the situation is referred as the multicollinearity problem. Multicollinearity implies that $X'X$ is near singular and at least one eigenvalue is very close to zero. In this case $X'X$ can be invertible then the parameter estimates can be found but their standard errors are very large than it should be. High variances of the regression coefficients may drastically reduce the precision of estimates. As a result of this some variables may be excluded from the model because they are not significant in the sample even though they are important in the population. Therefore, the detecting of this problem is very important.

There are some diagnostics to detect it: These are;

a. **Variance Inflation Factors ($\boldsymbol{VIF_j}$):** This measure calculated for each explanatory variables is actually the corresponding diagonal element of the inverse of the correlation matrix $(\boldsymbol{X'X})$ of explanatory variables.

$$VIF_j = 1/\left(1 - R_j^2\right)$$

where R_j^2 is the determination coefficient of the jth explanatory variable regressing on the remaining variables. The general rule of thumb there is a serious multicollinearity problem on the data set if one or some VIF values are greater than 10. (Wooldridge, J. M, 2000)

b. **The eigenvalues of the correlation matrix**$(\boldsymbol{X'X})$: Let $\lambda_1, \lambda_2, \dots \dots, \lambda_p$ be the eigenvalues of the correlation matrix. If there is one or more severe collinearity between the columns of matrix X, this causes some of the eigenvalues to be very close zero. In a very ideal case, which is the orthogonality between columns of X, the sum of invers of eigenvalues is equal to the number of the explanatory variables

$$\sum_{i=1}^{p} \frac{1}{\lambda_i} = p$$

As the sum is going apart from p then the severity of multicollinearity is increasing (Belsley, Kuh and Welsch, 1980).

c. **Condition Number:** It is defined as the ratio of the maximum eigenvalue to the minimum eigenvalue. If it lies between 30 and 100 it signs a moderate multicollinearity and is greater than 100 the data has a severe multicollinearity problem

$$Condition\ Number(CN) = \frac{\lambda_{max}}{\lambda_{min}} < 30$$

d. where λ_{max} and λ_{min} is the maximum and minimum eigenvalues of the correlation matrix, respectively (Belsley et al, 1980).

3. Ridge Regression

The multiple regression model is given by

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{1}$$

In presence of multicollinearity there are several remedies that we can apply, for avoiding from *its* undesirable effects on the estimates. Ridge regression is one of the remedies mostly employed. It was firstly proposed by Hoerl and Kennard (1970 a, b). In this method the estimates of the regression coefficients are obtained with a little bias guaranteed a smaller variance by adding a very small positive number in the diagonal elements of $\boldsymbol{X'X}$. While the least squares estimators of regression coefficients are

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} \tag{2}$$

the ridge estimators are introduced as

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} = (\boldsymbol{X'X} + k\boldsymbol{I})\boldsymbol{X'Y} \tag{3}$$

where $\boldsymbol{k}$ is a very small constant determined by the researcher. Here $(\boldsymbol{X'X})$ is in the correlation form. Gauss Markov Theorem states that under the standard assumptions about errors; such as errors have expectation zero, are uncorrelated and have the equal variances; the least squares estimators of the parameters of the model in (1) are linear, unbiased and have the minimum variances. But there is no guarantee that the variance of $\widehat{\boldsymbol{\beta}}$ will be small. For this purpose the ridge estimator estimates $\boldsymbol{\beta}$ with a bias but a smaller variance than the least squares estimators' one. The mean squared error of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ we can easily see that

$$MSE\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}\right) = E\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} - \boldsymbol{\beta}\right)^2 = Var\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}\right) + \left[E\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}\right) - \boldsymbol{\beta}\right]^2 \tag{4}$$

can be made small than the mean squared error of $\widehat{\boldsymbol{\beta}}$ which is equal to variance of $\widehat{\boldsymbol{\beta}}$ since there is no bias in it.

The ridge estimator is actually the linear transformation of the least squares estimator.

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} &= (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{Y}\\ &= (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-\mathbf{1}}(\boldsymbol{X}'\boldsymbol{X})\widehat{\boldsymbol{\beta}} = \boldsymbol{Z}\widehat{\boldsymbol{\beta}} \qquad (5)\end{aligned}$$

The expected value of ridge estimator tells us that it is also an biased estimator of $\boldsymbol{\beta}$.

$$E(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}) = E(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}) = \boldsymbol{Z}\,\boldsymbol{\beta} \qquad (6)$$

The variance-covariance matrix of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ is

$$\begin{aligned}Var(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}) &= Var(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}) = \boldsymbol{Z}Var(\widehat{\boldsymbol{\beta}})\boldsymbol{Z}'\\ &= \boldsymbol{Z}(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}})\boldsymbol{Z}'\\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{kI})^{-\mathbf{1}}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-\mathbf{1}} \qquad (7)\end{aligned}$$

Furthermore, VIF values based on the ridge estimators are defined as the diagonal elements of the matrix

$$(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{kI})^{-\mathbf{1}}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-\mathbf{1}}$$

On the other side since the least squares estimator is unbiased, the mean squared error will be the variance of the estimator.

$$\begin{aligned}MSE(\widehat{\boldsymbol{\beta}}\,) &= E(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = Tr\ Var(\widehat{\boldsymbol{\beta}}\,)\\ &= \sigma^2 Tr(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2\sum_{j=1}^{p}\frac{1}{\lambda_j} \qquad (8)\end{aligned}$$

where λ_j is the j^th^ eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$. Contrarily the mean squared error of ridge estimator is decomposed into two parts as below.

$$\begin{aligned}MSE(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}) &= Tr\ Var(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}) + Bias^2\\ &= \sigma^2 Tr[\,(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-\mathbf{1}}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-\mathbf{1}}] + k^2\,\boldsymbol{\beta}(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-\mathbf{2}}\boldsymbol{\beta}\\ &= \sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j + k)^2} + k^2\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-\mathbf{2}}\boldsymbol{\beta} \qquad (9)\\ &= \gamma_1(k) + \gamma_2(k)\end{aligned}$$

First part is the sum of variances of all the $\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}$**.** The second part is considered the square of a bias. It is obvious that the total variance decreases as k increases, while the squared bias increases. Therefore the possibility exists that there are admissible nonzero values of k for which

$$MSE(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}) < MSE(\widehat{\boldsymbol{\beta}}\,)$$

If it can be done $Var(\widehat{\boldsymbol{\beta}}\,) > Var(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}})$ can be satisfied. In order to understand the relationship among the variance, bias and k, there will be more informative to have a look at the graph, which is well known by the researchers dealing with ridge regression, presented in Figure 1. All of the related proofs can be accessed from the paper by Hoerl and Kennard (1970a).

The residual sum of squares for the ridge estimator is

$$\begin{aligned}SSE(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}) &= (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}})\\ &= (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} - \widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} - \widehat{\boldsymbol{\beta}}) \qquad (10)\end{aligned}$$

First term in the right side of the equation (10) is the residual sum of squares for the least square estimator and the second term is actually the quadratic form of $(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} - \widehat{\boldsymbol{\beta}})$**.**This implies that as k increases the residual sum of squares of ridge estimator increases and consequently the determination coefficient $\boldsymbol{R}^{\mathbf{2}}$ based on Ridge decreases. Therefore, the ridge estimate will not necessarily give the best fit to the data when we are more interested in obtaining a stable set of parameter estimates.

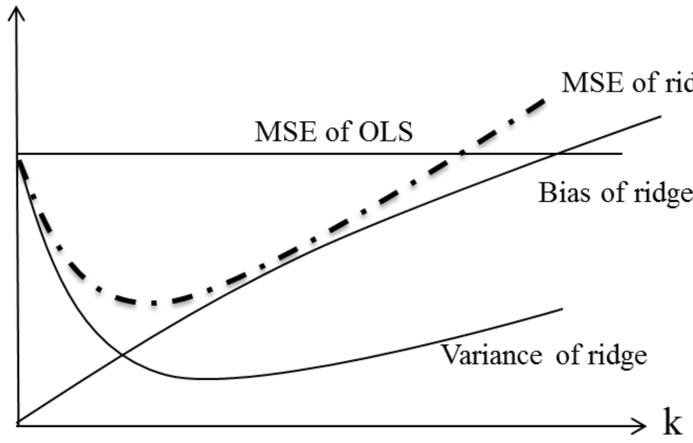

Figure 1: The relationships among variance,bias and k

As we can easily see that the choosing a value of k is a crucial issue in ridge regression. Ridge trace is one of the methods that we can apply. It is a plot of the elements of the ridge estimator versus k usually in the interval (0, 1). From the ridge trace the researchers can see that at a reasonable k value the estimates become stable (Hoerl and Kennard, 1970b). Marquardt and Snee (1975) suggest using only 25 values of k, spaced approximately logarithmically over that interval.

The optimal k value can be found by examining the orthogonal form of the regression model. It is

$$\boldsymbol{Y} = \boldsymbol{X}^{*}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{11}$$

where $\boldsymbol{X}^{*} = \boldsymbol{XD}$ and $\boldsymbol{\alpha} = \boldsymbol{D}'\boldsymbol{\beta}$. The Generalized Ridge Estimators of $\boldsymbol{\alpha}$ is

$$\widehat{\boldsymbol{\alpha}_{\boldsymbol{R}}} = (\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*} + k\boldsymbol{I})^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{Y} \tag{12}$$

The value of k_i which minimizes the MSE ($\widehat{\boldsymbol{\alpha}_{\boldsymbol{R}}}$) is

$$k_{i_{GR}} = \frac{\sigma^2}{\alpha_i^2} \tag{13}$$

where σ^2 is the error variance and α_i is the ith element of $\boldsymbol{\alpha}$ (Hoerl and Kennard, 1970a,b). When the actual values are not known the formula will be

$$k_{i_{GR}} = \frac{\widehat{\sigma^2}}{\hat{\alpha}_i^2} \tag{14}$$

where the estimates are obtained from the least squares. Alternative forms of the formula based on $\hat{\alpha}$ for k can be respectively given as follows;

1. The harmonic mean of $k_{i_{GR}}$; $k_{HR} = \frac{p\hat{\sigma}^2}{\sum \hat{\alpha}_i^2}$(Hoerl et al., 1975) (15)
2. The geometric mean of $k_{i_{GR}}$; $k_{GM} = \frac{\hat{\sigma}^2}{\Pi(\hat{\alpha}_i^2)^{1/p}}$ (Kibria, 2003) (16)
3. The median of $k_{i_{GR}}$;$k_{MED} = Median(k_{i_{GR}}) = Median\left(\frac{\widehat{\sigma^2}}{\hat{\alpha}_i^2}\right)$(Kibria,2003) (17)

Hoerl, Kennard and Baldwin (1975) suggested another method for finding k value which is given as

$$k = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}} \tag{18}$$

where $\hat{\sigma}^2$ and $\hat{\boldsymbol{\beta}}$ are the least squares estimates and this approach is called "*ordinary ridge regression*" (ORR).

Hoerl and Kennard (1976) introduced an iterative method for finding optimal k value, which is called "iterative ridge regression" (IRG). In this method k is calculated as in below;

$$k_{IRG(t)} = \frac{p\hat{\sigma}^2(t-1)}{\hat{\boldsymbol{\beta}}_{(t-1)}{}'\hat{\boldsymbol{\beta}}_{(t-1)}} \tag{19}$$

where $\hat{\sigma}^2(t-1)$ and $\hat{\boldsymbol{\beta}}_{(t-1)}$ are the corresponding residual mean square and the estimate vector of regression coefficients at (t-1)th iteration, respectively. Generally, the initials are chosen the results from the least squares method.

Uslu (2014) proposes to find k value using the particle swarm optimization technique, regarding to the objective function defined as $\min\{MAPE(k) + \emptyset(k)\}$. In this objective function $\emptyset(k)$ is defined by the sum of VIF values providing that VIF's are limited to be less than 10. Praga Alejo et al. (2008) propose to find k by using the genetic algorithm with a problematic objective function and there is no limitation for VIF values. Therefore, the standard errors of the regression coefficients will be able to shrink too much and the bias will increase too much as well, which is the case we don't want.

4. The Proposed Approach for finding k

In this paper we redefine the objective function as in Equation (20) below and propose two approaches based on the genetic algorithm and differential evolution algorithm for finding the best value for k. The objective function is defined as

$$\min\{MAPE(k) + \emptyset(k) + \theta(k)\} \tag{20}$$

with subject to $0 < k < 1$; where $MAPE(k) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad i = 1,2, \ldots..n$

$$\emptyset(k) = \begin{cases} 0, & \forall VIF_j < 10, \qquad j = 1,2,\ldots,p \\ \sum_{j=1}^{p} VIF_j, & otherwise\,, \qquad j = 1,2,\ldots p \end{cases}$$

$$\theta(k) = \begin{cases} 0, & \forall CN_j < 30, \qquad j = 1,2,\ldots,p \\ CN_j, & otherwise, \qquad j = 1,2,\ldots,p \end{cases}$$

By defining this objective function in that way, we are trying to find k value which minimizes the mean absolute prediction error with subject to both VIF values and the condition number in order to be sure that there will be no more multicollinearity problem in the data set used.

In ridge regression as k increases the ill conditioning of $\boldsymbol{X}'\boldsymbol{X}$ is getting well. Therefore, the detectors of multicollinearity such as the condition number and VIF values are getting smaller but the residuals sum of squares increases.

Before we introduce both of the proposed algorithms, it would be good to remind how we generate genes and chromosomes. We have decided the number of genes as 4 because we know that k must be a very small positive number (Hoerl, 1962). Each gene is a number randomly generated from the interval (0, 9) because we want that each chromosome will be corresponded to a 4-digit number. Then this number is converted to the value between (0, 1) for k by using formula given in Equation 22.

Here are the steps of both proposed methods respectively.

The steps of the proposed method based on the differential evolution algorithm.

Step 1: Generate the initial population.

After NP (the size of the population or the # of chromosomes); D (the # of the genes in each chromosomes) are defined the initial population are generated by the equation (21).

$$x_{ji} = x_j^l + round\big(rand_j[0\ 1] * (x^u - x^l)\,\big) \quad i = 1,2,\ldots,NP,\ \ j = 1,2,\ldots.,D \tag{21}$$

where

x_{ji}: j$^{\text{th}}$ gene of i$^{\text{th}}$ chromosome,

x^l, x^u: the lower and upper limits for a gen.

The number of genes D, the lower limit x^l and the upper limit x^u are to be chosen 4, 0 and 9, respectively. The aim of choosing D=4 is to determine the precision of the decimal part.

NP has to be at least 4 for the operators of the differential evolution algorithm to be implemented.

Step 2: Evaluate the values for k from the genes.

Each k value corresponding to the chromosomes are obtained from the following equation.

$$k_i = \frac{1}{10^D}\sum_{j=1}^{D} x_{ji}10^{D-j} \qquad i = 1,2,\ldots,NP,\ \ j = 1,2,\ldots.,D \tag{22}$$

Step 3: Apply the mutation operation.

3 chromosomes are randomly selected apart from the chromosome, which is called *the current chromosome*, symbolized by x, and is actually to be subjected to mutation operation. It is taken the difference between the first two of the 3 randomly selected chromosomes, and then it is multiplied by the scaling factor (F). And finally, it is added to the third chromosome and the final chromosome, which is called *the total chromosome*, is being generated.

With $x_{r1,}x_{r2}, x_{r3}$ randomly chosen chromosomes, and F from the interval (0, 2),

$$n_{ji} = x_{j,r3} + round\left(abs\left(F * \left(x_{j,r1} - x_{j,r2}\right)\right)\right) \tag{23}$$

the total chromosome n is obtained. F is a real and constant factor $\in [0, 2]$ which controls the amplification of the differential variation $\left(x_{j,r1} - x_{j,r2}\right)$ (Storn and Price, 1997). To be able to get the appropriate values for the genes we apply the following adjustment. Then the final total chromosome ${n_{ji}}^*$ is replaced to the total chromosome as follows.

$${n_{ji}}^* = \begin{cases} n_{ji} & x^l \leq n_{ji} \leq x^u \\ n_{ji} = 9, & n_{ji} > x^u \end{cases} \tag{24}$$

$n_{ji} = {n_{ji}}^*$

Step 4: Apply the crossover operation

The nominee chromosome for the new population is generated by using the current chromosome and the total chromosome(n). To do that, first of all random numbers are generated from the interval (0,1) for each genes. Then each random number are compared with the crossover ratio, which should be determined in the first step. With the following rule

$$u_{ji} = \begin{cases} n_{ji}, & rand(0,1) \leq co \ \vee \ j = j_{rand} \\ x_{ji}, & otherwise \end{cases} \tag{25}$$

The nominee chromosome (u) has been generated. With $j = j_{rand}$ at least one gene from the total chromosome has been transferred to the nominee chromosome.

Step 5: Calculate the fitness value and the selection

Fitness value is the value of $MAPE(k) + \emptyset(k) + \theta(k)$. After the mutation and cross over operation the fitness value of the nominee chromosome is calculated. If it is less than the one of the current chromosome the nominee chromosome, instead of the current chromosome, is replaced into the new population, otherwise the current chromosome remains in the next generation.

If we symbolize x_{G+1}, as the chromosome to be involved in the new population, the rule will be the as follows.

$$x_{i,G+1} = \begin{cases} u_{i,} & f(u_i) \geq f(x_i) \\ x_i, & otherwise \end{cases} \tag{26}$$

Step 6: The steps from 3 to 5 is repeated for all chromosome in the population, successively. The new generation has been constructed.

Step 7: The algorithm from Step 2 with the new generation constructed at Step 6 is repeated up to the iteration number (*its*). The best population will be reached at the end of the iteration. The chromosome with the best fitness value in this population will be the best solution of the problem.

The steps of the proposed method based on the genetic algorithm

Step1: Define the parameters of the genetic algorithm.

Let be 'NP' the number of chromosomes which means the size of population, 'D' the number of the gens in the chromosome, 'co' the ratio of crossover, 'mr' the ratio of mutation, 'es' the number of the chromosomes to be eliminated from the population and 'itrs' the number iteration.

Step2: Generate the initial population

$$x_{ji} = x_j^l + integer\big(rand_j[0\ 1] * (x^u - x^l)\big) \quad i = 1,2,\ldots,NP,\ \ j = 1,2,\ldots.,D \qquad (27)$$

where the parameters are defined as in the step 1 of the previous algorithm.

Step 3: Evaluate the values for k from the genes

This step is the same as the step 2 of the previous algorithm. The fitness value based on the objective function is calculated by k values obtained from each chromosome of the initial population.

Step 4: Apply the Natural Selection.

Due to the principle of "the more strong the most possible to survive", the 'es' chromosomes with the worst fitness value, are removed from the population. Then the 'es' chromosomes are regenerated as in the step 2.

Step 5: Apply the crossover operation.

In this step, the chromosomes are randomly paired with each other. The crossover ratio is compared with the randomly generated number from the interval (0, 1) for each pair. The crossover operation is applied to the pair with the random number, which is less than the 'cr'. Then another random number from the integer interval (1, L-1) is also generated to decide which crossover point will be.

Step 6: Apply the mutation operation.

First of all, random numbers are generated from the interval (0, 1) for each chromosome to decide whether the mutation operation is applied, or not. These random numbers are compared to the mutation ratio. When the random number is less than the crossover ratio the crossover operation is applied to the corresponding chromosome. And again, another random number from the integer interval (1, L) is generated to decide which gene is regenerated in that chromosome.

Step 7: The Choice of k.

From Step 4 to Step 6 a new generation has been constructed. In order to reach optimal solution for k, these steps are repeated up to the iteration number. At the end of the iterations the final population has been constructed. The chromosome with the best fitness value will be the optimal solution.

5. Application

Our proposal approaches have been applied to the real data sets which are known as "Import Data" and "Longley Data". We have chosen these data sets since we want to compare the results from our proposed methods with the results from the approaches in literature. The variables are imports (IMPORT-Y), domestic production (DOPROD-X1), stock formation (STOCK-X2) and domestic consumption (CONSUM-X3), all measured in billions of French francs for the years 1949 through 1959 (Chatterjee and Hadi, 2006). Longley's data set is a classic example of the data with the problem multicollinearity (Longley, J.W., 1967).

Our proposal approaches were coded in MATLAB-2015. The programs based on DEA and GA are executed for the parameters as given below for both data sets.

The parameters for DEA;

- ❖ The number of iterations (*its*) as 50, 100, 150, 200.
- ❖ The number of chromosomes (NP) as 30, 40, 50, 60.

- The scaling factor (F) as 0.5, 0.8, 1.2.
- The crossover ration (cr) as 0.4, 0.5, 0.6, 0.7.

The parameters for GA;

- The number of iterations (its) as 50, 100, 150, 200.
- The number of chromosomes (NP) as 30, 40, 50, 60.
- The number of chromosomes to be eliminated (es) as 5, 8, 13, 17.
- The crossover ration (*cr*) as 0.65, 0.75, 0.80, 0.85, 0,95.
- The mutation ration (*mr*) as 0.005, 0.01, 0.05, 0.01.

For Longley Data the best result of MAPE was found as 0.210446 at $\mathrm{k} = 0.1584$ from both DEA and GA. This result from DEA is obtained at $F = 0.5\,,\ cr = 0.6\,,\ NP = 30\ and\ its = 50$ and from GA at $cr = 0.95\,, mr = 0.05\,,\ NP = 40\,,\ es = 8\ and\ its = 50$. For Import Data the best result of MAPE is found as 0.126696 at $\mathrm{k} = 0.0662$ from both DEA and GA. This result is obtained from DEA at $F = 0.5\,,\ cr = 0.6\,,\ NP = 30\ and\ its = 50$ and from GA at $cr = 0.95\,, mr = 0.01\,,\ NP = 40\,,\ es = 8\ and\ its = 100$. Table 1 and 2 provides the comparative results with the other techniques in the literature. Although the other techniques can provide the smaller k value but some VIF values and condition number seem still problematic. In our application k has been found as 0.1584 for Longley Data and as 0.0662 for Import Data from both DEA and GA. It has been observed that at these values of k obtained for both data sets all of VIF's are less than 10 and condition numbers are less than 30.

For the purpose of comparison of two proposed approaches we summarized the results obtained at different combinations of NP and its by fixing the other parameters such as F, *cr*, mr and es. Table 3 and 4 is represented to show these comparisons at $F = 0.5\,,\ cr = 0.6$ for DEA and at $cr = 0.85\,, mr = 0.05\ and\ es = 8$ for GA for Longley Data. Table 5 and 6 is constructed to show the comparisons between DEA and GA for Import data at the same parameter's values. We have picked these parameters values among many trials that give the minimum MAPE, just as an example. From these tables we can conclude that DEA can reach the optimal value more often than GA. We should point out that at the different values for the parameters, which we have fixed at Table 3 and 4, we have reached the same conclusion. We can conclude that DEA has found the minimum MAPE value more often than GA.

Table 1: Longley Data. COEF: Coefficients, SE: Standard Errors of the coefficients.

	k	COEF	SE	VIF	CN	SSE	MAPE
Ordinary Least Squares	0	0.046	0.261	135.53	12220	0.00452	0.0887
		-1.014	0.948	1788.51			
		-0.538	0.130	33.62			
		-0.2047	0.0425	3.59			
		-0.101	0.448	399.15			
		2.48	0.617	758.98			
Harmonic Mean	0.0004	-0.0134	0.2202	87.3167	6242.9	0.0050	0.0753
		-0.2524	0.5122	472.1528			
		-0.4306	0.0780	10.9497			
		-0.1814	0.0400	2.8773			
		-0.2828	0.3164	180.2142			
		1.8797	0.4149	309.8213			
Geometric Mean	0.0021	0.0784	0.1931	46.6209	1824.7	0.0072	0.0938
		0.2467	0.1863	43.3779			
		-0.3449	0.0535	3.5720			
		-0.1511	0.0444	2.4598			
		-0.1801	0.2164	58.5586			
		1.1181	0.2485	77.1886			
Median From Kibria 2003	0.0019	0.0683	0.1949	48.8217	1984.7	0.0070	0.0915
		0.2320	0.1989	50.8888			
		-0.3485	0.0537	3.7031			
		-0.1526	0.0439	2.4802			

		-0.1962	0.2218	63.2613			
		1.1622	0.2577	85.3985			
Ordinary Ridge	0.00036	-0.0134	0.2203	87.3630	6249.9	0.0050	0.0753
		-0.2532	0.5127	473.0805			
		-0.4308	0.0781	10.9656			
		-0.1814	0.0400	2.8779			
		-0.2827	0.3166	180.3958			
		1.8805	0.4151	310.1863			
Iterative Ridge	0.0014	0.0379	0.1999	56.1977	2591.7	0.0064	0.0845
		0.1719	0.2456	84.8510			
		-0.3613	0.0552	4.2905			
		-0.1590	0.0426	2.5478			
		-0.2422	0.2392	80.4556			
		1.3094	0.2876	116.2878			
Eren 2014	0.0172	0.02	0.1168	9.99	262.88	0.0123	0.1378
		0.10	0.0459	1.54			
		-0.37	0.0590	2.55			
		-0.16	0.0516	1.95			
		-0.27	0.1000	7.32			
		1.44	0.0746	4.07			
Proposed Method with GA	**0.1584**	**0.2472**	**0.0341**	**0.4040**	**29.9904**	**0.0259**	**0.2104**
		0.2860	**0.0230**	**0.1845**			
		-0.1377	**0.0557**	**1.0797**			
		-0.0034	**0.0515**	**0.9223**			
		0.2341	**0.0298**	**0.3084**			
		0.2611	**0.0193**	**0.1291**			
Proposed Method with DEA	**0.1584**	**0.2472**	**0.0341**	**0.4040**	**29.9904**	**0.0259**	**0.2104**
		0.2860	**0.0230**	**0.1845**			
		-0.1377	**0.0557**	**1.0797**			
		-0.0034	**0.0515**	**0.9223**			
		0.2341	**0.0298**	**0.3084**			
		0.2611	**0.0193**	**0.1291**			

Table 2. Import Data. COEF: Coefficients, SE: Standard Errors of the coefficients.

	k		COEF	SE	VIF	CN	SSE	MAPE
Ordinary Least Squares	0		-0.3393	0.464	185.9975	742.9346	0.0081	0,1052
			0.2130	0.0343	1.0189			
			1.3027	0.464	186.1100			
Harmonic Mean	0.0016		-0.0297	0.2976	72.0916	462.4151	0.0086	0.1097
			0.2158	0.0351	1.0046			
			0.9922	0.2977	72.1348			
Geometric Mean	0.0035		0.1256	0.2153	34.8947	321.4231	0.0093	0.1235
			0.2169	0.0364	0.9972			
			0.8359	0.2154	34.9153			
Median From Kibria 2003	0.0021		0.0222	0.2704	58.1769	415.3280	0.0088	0.1206
			0.2162	0.0355	1.0022			
			0.9400	0.2705	58.2117			
Ordinary Ridge	0.0016		-0.0340	0.3001	73.3001	466.2798	0.0086	0.1097
			0.2157	0.0351	1.0048			
			0.9965	0.3002	73.3440			
Iterative Ridge	0.0042		0.1594	0,1969	28.5743	290.7252	0.0095	0.1137
			0.2171	0.0368	0.9952			
			0.8018	0.1970	28.5910			
Eren 2014	0.0090		0.2897	0.1212	9.99	171.7709	0.0103	0.1185
			0.2174	0.0380	0.98			
			0.6692	0.1213	10.0			

Proposed Method with GA	**0.0662**		**0.4333**	**0.0305**	**0.5183**	**29.9801**	**0.0126**	**0.1267**
			0.2080	**0.0398**	**0.8803**			
			0.4993	**0.0305**	**0.5181**			
Proposed Method with DEA	**0.0662**		**0.4333**	**0.0305**	**0.5183**	**29.9801**	**0.0126**	**0.1267**
			0.2080	**0.0398**	**0.8803**			
			0.4993	**0.0305**	**0.5181**			

Table 3: For DEA, MAPE/the iteration at which the optimal solution has been achieved for Longley Data.

NP / Iteration	30	40	50	60
50	0,210446/21	0,210784/29	0,210446/34	0,210446/38
100	0,210446/14	0,210446/44	0,210446/51	0,210446/32
150	0,210446/29	0,210446/14	0,210446/14	0,210446/34
200	0,210446/25	0,210446/38	0,210446/62	0,210446/35

The optimal MAPE is obtained at F=0.5 and cr=0.6.

Table 4: For GA, MAPE/the iteration at which the optimal solution has been achieved for Longley Data.

NP / Iteration	30	40	50	60
50	0,211497/16	0,211288/6	0,210657/3	0,210615/19
100	0,212245/3	0,210489/11	0,210784/27	0,210446/2
150	0,210489/18	0,210531/43	0,210446/44	0,210446/46
200	0,210657/153	0,210700/154	0,210489/79	0,210446/111

The optimal MAPE is obtained at cr=0.85, mr=0.05.

Table 5: For DEA, MAPE/the iteration at which the optimal solution has been achieved for IMPORT Data.

NP / Iteration	30	40	50	60
50	0,126696/19	0,126715/42	0,126696/41	0,126696/28
100	0,126734/6	0,126696/32	0,126696/36	0,126696/46
150	0,126696/20	0,126696/21	0,126715/21	0,126696/31
200	0,126696/23	0,126715/16	0,126734/31	0,126696/31

The optimal MAPE is obtained at F=0.5 and cr=0.6.

Table 6: For GA, MAPE/the iteration at which the optimal solution has been achieved for IMPORT Data.

NP Iteration	30	40	50	60
50	0,126940/15	0,127089/20	0,126884/13	0,126977/11
100	0,126696/99	0,126734/80	0,126753/12	0,126828/4
150	0,126772/22	0,126696/62	0,126696/94	0,126734/133
200	0,126696/82	0,126696/85	0,126696/42	0,126696/47

The optimal MAPE is obtained at cr=0.85, mr=0.05.

6. Conclusion

As you can see from Table 1 and 2, at k found from both approaches, the standard errors of the regression coefficients have been decreased and MAPE, and consequently SSE, has not been allowed to increase too much. Also, we can say that the condition number as well as VIF values have been shrank to the desired level. Consequently, this result means that there is no more multicollinearity problem in the data sets.

Since we have taken into account condition number as well as VIF when MAPE was doing minimized, the proposed methods based on GA and DEA could have found the optimal solution for k and consequently minimum MAPE. Speaking of the performances of these two approaches we can say that there is not much difference between them. Both approaches find the same result, however the approach based on DEA finds this result more frequently than GA when we change the values of the parameters of the algorithms such as F, *cr*, mr and es. This result is true for both data sets.

References

Ahn, J.J., Byun, H.W., Oh, K.J., and Kim, T.Y., (2012). "Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting", *Expert Systems with Applications*, 39, 8369–8379.

Belsley, D. A., Kuh, E. and Welsch, R. E., "Regression Diagnostics: Identifying Influential Data and Sources of Collinearity", New York: John Wiley and Sons, 1980.

Chatterjee, S. and Hadi, A., "Regression analysis by example", 4th edition, New York, 2006.

Gibbons, D. G. (1981). "A simulation stady of some ridge estimators", *Journal of the American Statistical Association*, 76, 131-139.

Hoerl, A. E., (1962). "Application of ridge analysis to regression problems", *Chemical Engineering Progress*, 58, 54-59.

Hoerl, A.E., and Kennard, R.W. (1976). "Ridge regression: iterative estimation of the biasing parameter", *Communication in Statistics*, Part A5, 77-88.

Hoerl, A.E., and Kennard, R.W. (1970b). "Ridge regression: applications to non-orthogonal problems", *Technometrics*, 12, 69-82.

Hoerl, A.E., Kennard, R.W., and Baldwin, K.F. (1975). "Ridge regression: some simulation", *Communication in Statistics*, 4, 105-123.

Hoerl, A.E., and Kennard, R.W. (1970a). "Ridge regression: biased estimation for non-orthogonal problems", *Technometrics*, 12, 55-67.

Kibria, B.M.G. (2003). "Performance of Some New Ridge Regression Estimators", *Communications in Statistics - Theory and Methods*, 32, 419-435.

Longley J.W., (1967). "An appraisal of least squrares programs fort the electronic computer from point of view of the user", *Journal of The American Statistical Association*, 62, 819-841.

Mardikyan, S., Cetin, E. (2008). "Efficient Choice of Biasing Constant for Ridge Regression", *International Journal of Contemporary Mathematical Sciences*, Vol: 3, No: 11, pp. 527-536.

Marquardt, D.W., and Snee, R.D., (1975). "Ridge regression in practice", *The American Statisticians*, 29, 3-20.

Muniz G., Kibria, B. M. G., Mansson, K. and Shukur, G., (2012). "On developing ridge regression parameters:a graphical investigation", *Sort-Statistics And Operations Research Transactions*, Volume 36, Issue 2, pages 115-138.

Praga-Alejo, R.J., Torre-Trevino, L.M., and Pina-Monarrez M.R., (2008). "Optimal determination of k constant of ridge regression using a simple genetic algorithm", Electronics robotics and Automotive Mechanics Conference.

Shukur, G. and Khalaf, G. (2005). "Choosing Ridge Parameters for. Regression Problems", *Communication in Statistics – Theory and Methods*, 34, 1177-1182.

Storn, R. and Price K., (1997). "Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces", *Journal of Global Optimization*, 11: 341–359.

Uslu V. R., Eğrioğlu E. and Baş E., (2014). "Finding Optimal Value for the Shrinkage Parameter in Ridge Regression via Particle Swarm Optimization", *American Journal of Intelligent Systems*, Volume 4, Number 4, pages 142-147.

Vinod, H.D., (1976). "Application of ridge regression methods to a study of Bell System scale economics", *Journal of the American Statistical Association*, 71, 835-841.

Wooldridge, J. M., "Introductory Econometrics: A Modern Approach", South Western, 2000.

Wooldridge, J. M., (2000). "A framework for estimating dynamic, unobserved effects panel data models with possible feedback to future explanatory variables", *Economic letters*, 68, 245-250.