Journal of Mathematics and Statistics

Review

Explaining the Generalized Cross-Validation on Linear Models

[1]Lucas Monteiro Chaves, [2]Laerte Dias de Carvalho, [2]Carlos José dos Reis and [2]Devanil Jaques de Souza

[1]Department of Exact Sciences, Federal University of Lavras, Brazil
[2]Department of Statistics, Federal University of Lavras, Brazil

Article history
Received: 01-07 2019
Revised: 19-09-2019
Accepted: 23-10-2019

Corresponding Author:
Lucas Monteiro Chaves
Department of Exact Sciences,
Federal University of Lavras,
Brazil
Email: lucas@ufla.br

Abstract: Cross-Validation is a model validation method widely used by the scientific community. The Generalized Cross-Validation (GCV) is an invariant version of the usual Cross-Validation method. This generalization was obtained using the non usual theory of circulant complex matrices. In this work we intend to give a clear and complete exposition concerning the linear algebra assumptions required by the theory. The aim was to make this text accessible to a wide audience of statisticians and non-statisticians who use the Cross-Validation method in their research activities. It is also intended to supply the absence of a basic reference on this topic in the literature.

Keywords: Circulant Matrices, PRESS Statistics, Prediction Error

Introduction

A statistical model should, like almost every scientific procedure, have one eye on the past and two eyes on the future. Once a model has been fitted to a data set, certainly it explain well the past. However, it will also describe well the future? This statistical fact is denoted predictive capability of the model, being probably the most important feature of a statistical model. It can be describe as: let $y' = (y_1,..., y_n)$ be a random vector with mean vector $\mu' = (\mu_1,...,\mu_n)$. After a data vector y was observed, some adjustment technique is adopted and then a model is proposed and expressed in the form $\hat{\mu} = m(y)$, where m is a function of $\mathbb{R}^n \to \mathbb{R}^n$. The question is how can we assess the predictive capacity of $m(y)$. If a new vector y^0 is observed, how close to $\hat{\mu} = m(y)$ will this vector be? Such question does not cover the entire prediction problem because there is still the problem that the data vector used in the adjustment was the realization of a random vector. Then, it is necessary that the whole procedure to be randomized:

(i) The data y are observed
(ii) The estimative $\hat{\mu} = m(y)$ is obtained
(iii) New data vector y^0 of the same random phenomenon is observed
(iv) The square of deviation $||y^0-m(y)||^2$ is then calculated

If this process is repeated several times, what is the mean of the sum of the squares of the deviations? It is necessary to formalize this procedure in terms of mathematical expectations.

Since we have two random vectors y and y^0, it is necessary for a proper definition of prediction error to take expectation in relation to each of these random vectors, that is, the double expectation $E[E_0[||y^0-m(y)||^2]]$ where $E[\cdot]$ is the expectation with respect to the vector y and $E_0[\cdot]$ the expectation with respect to y^0 (Efron (2004)).

One of the most used ways to access the predictive capability of a model is Cross-Validation. In its simplest forms can be described by: A model is fitted without using one of the data values. With this model a predicted value is obtained. The square of the difference between the not used value and the predicted one is taken. This procedure is repeated for all data values and the mean of deviations is calculated, being the estimative of the prediction error. This makes the Cross-Validation process essentially a computational procedure. For linear regressions David M. Allen obtained a closed formula for this estimator, named Prediction Sum of Squares (PRESS) (Allen (1971); Allen (1974)). Golub *et al*. (1979) obtained a generalization of this formula for the case of GCV.

To obtain these formulas it is necessary to use very complicated matrix identities. As in the original articles this isn't done in details and what is more important, no one intuitive ideas are presented, in this article we will present in details all the linear algebra needed and try to give, using geometrical arguments, an intuitive approach. We hope that this could be interesting for the statistical and non statistical

audience that uses Cross-Validation in the linear regression problems.

In section 2 is presented the theory of Cross-Validation for linear regression problem. The PRESS statistics are deduced in details. It showed that the PRESS statistics is an almost unbiased estimator of prediction error. In section 3 the theory of circulant matrices is presented and applied to construct the GCV. This section is also showed that these statistics is almost unbiased. In section 4, some computational simulation is done to show that the GCV is a better estimator than PRESS.

Cross-Validation in Linear Regression

Let us consider the simplest situation. A set of data (y_i, x_i) $i = 1,...,n$, where y_i is the response variable and $x'_i = (x_{i1},..., x_{ip})$ a vector of explanatory variables is observed and a model $m(y)$ is fitted. An estimator of the predictive capability of this model is obtained from the following construction: The model is fitted without the *i-th* observation y_i, that is, using the vector $y_{(i)}$ ($y_{(i)}$ is the vector y with observation y_i omitted). Take the square of the difference between the fitted value $\hat{y}_i$ and the observed value y_i.

The mean $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2$ is an estimator of the predictive capability of the model, $E\left[E_0\left[\left(\hat{y}-y^0\right)^2\right]\right]$.

Consider a linear regression model $y_{n\times 1}\beta_{n\times p} + \varepsilon_{n\times 1}$. The ordinary least square estimator is given by $\hat{\beta} = (X'X)^{-1}X'y$. We will make the regression when we delete the *i*-th sample unit, that is, we will not use y_i and the *i*-th row $x'_i = (x_{i1},..., x_{ip})$ of the matrix X. Denote by $X_{(i)}$ the matrix $(n-1)\times p$, where the *i-th* row was deleted. In statistics this can express, for example, that one experimental unit was lost. With this we have a new linear regression $y_{(i)} = X_{(i)}\beta + \varepsilon$, where $y_{(i)}$, $X_{(i)}$ and ε have dimensions $(n-1)\times 1$, $(n-1)\times p$ and $(n-1)\times 1$, respectively.

For this model, the least square estimator is:

$$\hat{\beta}_{(i)} = \left(X'_{(i)}X_{(i)}\right)^{-1}X'_{(i)}y_{(i)}.$$

If we apply the matrix X on $\hat{\beta}_{(i)}$:

$$X\hat{\beta}_{(i)} = \begin{bmatrix} x'_1 \\ \vdots \\ x'_i \\ \vdots \\ x'_n \end{bmatrix}\left[\hat{\beta}_{(i)}\right] = \begin{bmatrix} \vdots \\ x'_i\,\hat{\beta}_{(i)} \\ \vdots \end{bmatrix}.$$

Therefore, $x'_i\,\hat{\beta}_{(i)}$ is the estimated value for the y_i data value, $x'_i\,\hat{\beta}_{(i)} = \hat{y}_i$ $\left(y_i - \hat{y}_i\right)^2 = \left(y_i - x'_i\,\hat{\beta}_{(i)}\right)^2$ is the square of the difference between the value effectively observed yi and the estimated value $\hat{y}_i$.

Now, we want to relate $X'_{(i)}X_{(i)}$ with $X'X$.

Affirmation 1: $X'X - x_i x'_i = X'_{(i)}X_{(i)}$:

$$X_{n\times p} = \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix}, X'_{p\times n} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}.$$

$$X'_{p\times n}X_{n\times p} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}\begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix}.$$

Will be convenient change the notation of the matrix entries of X as:

$$X_{n\times p} = \begin{bmatrix} z_1 & z_2 & \cdots & z_p \end{bmatrix} \text{ and } X'_{p\times n} = \begin{bmatrix} z'_1 \\ z'_2 \\ \vdots \\ z'_n \end{bmatrix}.$$

$$X'_{p\times n}X_{n\times p} = \begin{bmatrix} z'_1 \\ z'_2 \\ \vdots \\ z'_n \end{bmatrix}\begin{bmatrix} z_1 & z_2 & \cdots & z_p \end{bmatrix} = \begin{bmatrix} z'_1 z_1 & z'_1 z_2 & \cdots & z'_1 z_p \\ \vdots & \vdots & \ddots & \vdots \\ z'_p z_1 & z'_p z_2 & \cdots & z'_p z_p \end{bmatrix}.$$

Therefore, the element in the row l and column k of $X'_{p\times n}X_{n\times p}$ is:

$$z'_l\, z_k = \sum_{s=1}^{n} z_{sl}z_{sk} = \left(\sum_{\substack{s=1 \\ s\neq i}}^{n} z_{sl}z_{sk}\right) + z_{il}z_{ik}, \tag{3}$$

where, $z_l = \begin{pmatrix} z_{1l} \\ z_{2l} \\ \vdots \\ z_{nl} \end{pmatrix}$ *and* $z_k = \begin{pmatrix} z_{1k} \\ z_{2k} \\ \vdots \\ z_{nk} \end{pmatrix}$.

Returning to the old notation for the entries of $X_{n\times p}$, we have $x_{mj} = z_{mj}$.

The matrix $x_i x_i' = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{bmatrix} \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{in} \end{bmatrix}$ has in the position *l*-th row and *k*-th column:

$$x_{il} x_{ik} = z_{il} z_{ik}.$$

As:

$$x_{(i)} = \begin{bmatrix} x_1' \\ \vdots \\ x_{i-1}' \\ 0 \\ x_{i+1}' \\ \vdots \\ x_p' \end{bmatrix} \leftarrow i-th\, row,$$

we have that the value in the row l and column *k* of $X'X - x_i x_i'$ is:

$$\left(\sum_{\substack{s=1 \\ s \neq i}}^{n} z_{sl} z_{sk} \right) + z_{il} z_{ik} = \sum_{\substack{s=1 \\ s \neq i}}^{n} z_{sl} z_{sk}. \tag{4}$$

Therefore, the row *i* of $X = \begin{bmatrix} z_1 & z_2 & \cdots & z_p \end{bmatrix}$ was suppressed and therefore we have the identity:

$$X'X - x_i x_i' = X_{(i)}' X_{(i)}. \tag{5}$$

There is a well known formula for the inverse of this sum (Henderson and Searle (1981), Rencher and Schaalje (2008)):

$$\begin{aligned} \left(X_{(i)}' X_{(i)} \right)^{-1} &= \left(X'X - x_i x_i' \right)^{-1} \\ &= \left(X'X \right)^{-1} + \frac{\left(X'X \right)^{-1} x_i x_i' \left(X'X \right)^{-1}}{1 - x_i' \left(X'X \right)^{-1} x_i}. \end{aligned}$$

The number $x_i' (X'X)^{-1} x_i$ admits the following interpretation: $H = X\left(X'X \right)^{-1} X'$, the hat matrix, is an orthogonal projection from $\mathbb{R}^n$ onto the image of application *X* (Im(*X*)). The element in the *i*-th row and *i*-th column of this projection is:

$$\begin{aligned} h_{ii} &= e_i' He_i \\ &= e_i' X \left(X'X \right)^{-1} X' e_i \\ &= \left(e_i' X \right) \left(X'X \right)^{-1} \left(X' e_i \right) \\ &= x_i' \left(X'X \right)^{-1} x_i. \end{aligned} \tag{6}$$

We have:

$$\begin{aligned} & y_i - \hat{y}_i \\ &= y_i - x_i' \hat{\beta}_{(i)} \\ &= y_i - x_i' \left(X_{(i)}' X_{(i)} \right)^{-1} X_{(i)}' y_{(i)} \\ &= y_i - x_i' \left[\left(X'X \right)^{-1} + \frac{\left(X'X \right)^{-1} x_i x_i' \left(X'X \right)^{-1}}{1 - h_{ii}} \right] X_{(i)}' y_{(i)} \\ &= y_i - x_i' \left(X'X \right)^{-1} X_{(i)}' y_{(i)} - \frac{x_i' \left(X'X \right)^{-1} x_i x_i' \left(X'X \right)^{-1} X_{(i)}' y_{(i)}}{1 - h_{ii}} \\ &= \frac{\left(1 - h_{ii} \right) y_i - \left(1 - h_{ii} \right) x_i' \left(X'X \right)^{-1} X_{(i)}' y_{(i)} - h_{ii} x_i' \left(X'X \right)^{-1} X_{(i)}' y_{(i)}}{1 - h_{ii}} \\ &= \frac{\left(1 - h_{ii} \right) y_i - x_i' \left(X'X \right)^{-1} X_{(i)}' y_{(i)}}{1 - h_{ii}}. \end{aligned} \tag{7}$$

But:

$$\begin{aligned} X'y &= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} y \\ &= \left(\begin{bmatrix} x_1 & \cdots & 0 & \cdots & x_n \end{bmatrix} + \begin{bmatrix} 0 & \cdots & x_i & \cdots & 0 \end{bmatrix} \right) \\ &\times \left(\begin{bmatrix} y_1 \\ \vdots \\ 0 \\ \vdots \\ y_n \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ y_i \\ \vdots \\ 0 \end{bmatrix} \right) \\ &= \begin{bmatrix} x_1 & \cdots & 0 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ 0 \\ \vdots \\ y_n \end{bmatrix} + \begin{bmatrix} x_1 & \cdots & 0 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ y_i \\ \vdots \\ 0 \end{bmatrix} \\ &+ \begin{bmatrix} 0 & \cdots & x_i & \cdots & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ 0 \\ \vdots \\ y_n \end{bmatrix} + \begin{bmatrix} 0 & \cdots & x_i & \cdots & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ y_i \\ \vdots \\ 0 \end{bmatrix} \\ &= X_{(i)}' y_{(i)} + 0 + 0 + y_i x_i. \end{aligned}$$

Then, replacing this result in (7) we hav that:

$$\begin{aligned} & \frac{\left(1 - h_{ii} \right) y_i - x_i' \left(X'X \right)^{-1} X_{(i)}' y_{(i)}}{1 - h_{ii}} \\ &= \frac{\left(1 - h_{ii} \right) y_i - x_i' \left(X'X \right)^{-1} \left(X'y - y_i x_i \right)}{1 - h_{ii}} \\ &= \frac{(1 - h_{ii}) y_i + y_i x_i' \left(X'X \right)^{-1} x_i - x_i' \left(X'X \right)^{-1} X'y}{1 - h_{ii}} \\ &= \frac{y_i - h_{ii} y_i + h_{ii} y_i - x_i' \left(X'X \right)^{-1} X'y}{1 - h_{ii}} \\ &= \frac{y_i - x_i' \left(X'X \right)^{-1} X'y}{1 - h_{ii}} \\ &= \frac{y_i - x_i' \hat{\beta}}{1 - h_{ii}}, \end{aligned} \tag{8}$$

where, $\hat{\beta} = (X'X)^{-1} X'y$ is the ordinary least square estimator.

The PRESS statistics is then defined by:

$$PRESS = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(\frac{y_i - x_i' \hat{\beta}}{1 - h_{ii}}\right)^2. \quad (9)$$

As $X\hat{\beta} = \hat{y}$ (y fitted, projection of y on Im(X)), the sum that defines the PRESS statistics can be written in terms of vector norm:

$$PRESS = \left\| B\left(y - X\left(X'X\right)^{-1} X'y\right)\right\|^2, \quad (10)$$

where, B is the diagonal matrix, $B = \begin{bmatrix} \frac{1}{1-h_{11}} & & \\ & \ddots & \\ & & \frac{1}{1-h_{nn}} \end{bmatrix}$.

Therefore, the PRESS statistics is given by the quadratic form:

$$\begin{aligned} PRESS &= \left\| B\left(y - X(X'X)^{-1} X'Y\right)\right\|^2 \\ &= \left\| B\left(y - Hy\right)\right\|^2 \\ &= \left\| B\left(y - H\right)y\right\|^2 \\ &= \left(B\left(I - H\right)y\right)' \left(B\left(I - H\right)y\right) \\ &= y'\left(I - H\right)' B'B\left(I - H\right)y \\ &= y'\left(I - H\right)BB\left(I - H\right)y \\ &= y'\left(I - H\right)B^2\left(I - H\right)y. \end{aligned} \quad (11)$$

Theorem 1 (Rencher and Schaalje (2008)

If y have a normal distribution with mean μ and covariance matrix Σ and if A is a symmetric matrix of constants, then the mean and variance of a quadratic form y'Ay is respectively:

$$E\left[y'Ay\right] = tr\left(A\Sigma\right) + \mu' A\mu \quad (12)$$

and:

$$\text{var}\left[y'Ay\right] 2tr\left[\left(A\Sigma\right)^2\right] + 4\mu' A\Sigma A\mu. \quad (13)$$

The proof of Theorem 1 can be seen in Rencher and Shaalje (2008, p. 107-110).

Thus, the mean of the PRESS statistic is:

$$\begin{aligned} &E\left[y'\left(I - H\right)B^2\left(I - H\right)y\right] \\ &= tr\left[\left(I - H\right)B^2\left(I - H\right)\right] \\ &+ \left(E\left[y\right]\right)' \left(I - H\right)B^2\left(I - H\right)E\left[y\right]. \end{aligned} \quad (14)$$

As $I–H$ is a orthogonal projection on the subspaces Im(X) and E [y] ∈ Im(X) the second term is null. Then:

$$\begin{aligned} E\left[PRESS\right] &= \sigma^2 tr\left[\left(I - H\right)B^2\left(I - H\right)\right] \\ &= \sigma^2 tr\left[B^2\left(I - H\right)\left(I - H\right)\right] \\ &= \sigma^2 tr\left[B^2\left(I - H\right)^2\right] \\ &= \sigma^2 \sum_{i=1}^{n} \frac{1 - h_{ii}}{\left(1 - h_{ii}\right)^2} \\ &= \sigma^2 \sum_{i=1}^{n} \frac{1}{1 - h_{ii}}. \end{aligned} \quad (15)$$

The variance of PRESS statistics is:

$$\begin{aligned} \text{var}\left[PRESS\right] &= 2tr\left[\left(\left(I - H\right)B^2\left(I - H\right)^2\right)\right]\sigma^2 \\ &= 2tr\left[\left(I - H\right)B^2\left(I - H\right)B^2\left(I - H\right)\right]\sigma^2 \\ &= 2tr\left[B^2\left(I - H\right)B^2\left(I - HB^2\right)\right]\sigma^2. \end{aligned} \quad (16)$$

The matrix B^2 is given by $B^2 = diag\left(\frac{1}{\left(1 - h_{ii}\right)^2}\right)$, i = 1,...,n. Thus, $B^2\left(I–H\right) = \left(\frac{1 - h_{ij}}{\left(1 - h_{ii}\right)^2}\right)$, for i, j = 1,...,n and:

$$B^2\left(I - H\right)B^2\left(I - H\right) = \left(\sum_{s=1}^{n}\left(\frac{1 - h_{is}}{\left(1 - h_{ii}\right)^2}\right)\left(\frac{1 - h_{sj}}{\left(1 - h_{ss}\right)^2}\right)\right).$$

Therefore, the variance of the PRESS statistic is:

$$\begin{aligned} \text{var}\left[PRESS\right] &= 2tr\left[B^2\left(I - H\right)B^2\left(I - H\right)\right]\sigma^2 \\ &= 2\sigma^2 \sum_{i=1}^{n}\sum_{s=1}^{n} \frac{\left(1 - h_{is}\right)}{\left(1 - h_{ii}\right)^2} \frac{\left(1 - h_{si}\right)}{\left(1 - h_{ss}\right)^2} \\ &= 2\sigma^2 \sum_{i=1}^{n}\sum_{s=1}^{n} \frac{\left(1 - h_{is}\right)^2}{\left(1 - h_{ii}\right)^2 \left(1 - h_{ss}\right)^2}. \end{aligned} \quad (17)$$

Proposition 1

The PRESS statistic is an almost unbiased estimator of the prediction error.

Proof:

The prediction error of a multiple linear regression $y = X\beta + \varepsilon$ is:

$$\begin{aligned}
&\sum_{i=1}^{n} E\left[E_0\left[\left(\hat{y}_i - y_i^0\right)^2\right]\right] \\
&= \sum_{i=1}^{n} E\left[E_0\left[\left(\hat{\beta}' x_i - y_i^0\right)^2\right]\right] \\
&= \sum_{i=1}^{n}\left(\sigma^2 + \operatorname{var}\left[\hat{\beta}' x_i\right] + \left(E\left[\hat{\beta}' x_i\right] - \beta' x_i\right)^2\right) \\
&= n\sigma^2 + \sigma^2 \sum_{i=1}^{n} x_i' \left(X'X\right)^{-1} x_i \qquad (18) \\
&= n\sigma^2 + \sigma^2 \sum_{i=1}^{n} e_i' X'\left(X'X\right)^{-1} X e_i \\
&= n\sigma^2 + \sigma^2 \sum_{i=1}^{n} h_{ii} \\
&= n\sigma^2 + \sigma^2 tr\left[H\right] \\
&= \sigma^2\left(n + p\right),
\end{aligned}$$

Allen (1971, p. 470).

Using the Taylor approximation of first order, we have that $\frac{1}{1-h_{ii}} \approx 1 + h_{ii}$. Then:

$$\begin{aligned}
E\left[PRESS\right] &= \sigma^2 \sum_{i=1}^{n} \frac{1}{1-h_{ii}} \\
&= \sigma^2\left(1 + h_{ii} + h_{ii}^2 + h_{ii}^3 + \ldots\right) \\
&\approx \sigma^2 \sum_{i=1}^{n}\left(1 + h_{ii}\right) \qquad (19) \\
&= n\sigma^2 + \sigma^2 tr\left[H\right] \\
&= \sigma^2\left(n + p\right).
\end{aligned}$$

Observe that PRESS overestimate the prediction error.

Generalized Cross-Validation

GCV is a rotation-invariant form of ordinary cross-validation. Golub *et al*. (1979) generalize the construction of the PRESS statistics in such way that they obtained a statistic invariant by rotations. The idea is to take the deviations between linear combinations of estimated responses and the correspondent linear combinations of the observed values. For this generalization is necessary to use a especial type of matrix over complex numbers denoted circulants matrices. These matrices are not so much used and therefore we will develop its theory in details (Golub *et al*., 1979).

Circulant Matrices

This subsection is based in Kra and Simanca (2012). Let $v' = (v_0, v_0, \ldots, v_{n-1})$ be a row vector with coordinates given by complex numbers. A circulant matrix defined by v is constructed such that its rows are obtained by clockwise rotations of the components of v:

$$V = circ(v) = \begin{bmatrix} v_0 & v_1 & \cdots & v_{n-1} \\ v_{n-1} & v_0 & \cdots & v_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ v_1 & v_2 & \cdots & v_0 \end{bmatrix}. \qquad (20)$$

Of particular importance are the circulant matrices obtained from the vector defined by the complex roots of unity. $\varepsilon \in D$ is a n-th primitive complex root of the unity if $\varepsilon^k \neq 1$, $k = 1,2,\ldots,n-1$ and $\varepsilon^n = 1$, for example $\varepsilon = e^{\frac{2\pi_i}{n}}$. The great advantage of the circulant matrices is the fact that its eigenvalues and eingvectors are explicit given by: for a n-th primitive complex root of unity ε, if $\lambda_l = v_0 + \varepsilon^l v_1 + \varepsilon^{2l} v_2 + \ldots + \varepsilon^{(n-1)l} v_{n-1}$ then:

$$\begin{bmatrix} v_0 & v_1 & \cdots & v_{n-1} \\ v_{n-1} & v_0 & \cdots & v_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ v_1 & v_2 & \cdots & v_0 \end{bmatrix} \begin{bmatrix} 1 \\ \varepsilon^l \\ \vdots \\ \varepsilon^{(n-1)l} \end{bmatrix} = \lambda_l \begin{bmatrix} 1 \\ \varepsilon^l \\ \vdots \\ \varepsilon^{(n-1)l} \end{bmatrix}.$$

The proof follows by inspection. For example for the second row we have:

$$\begin{aligned}
&v_{n-1}\varepsilon^0 + v_0\varepsilon^l + v_1\varepsilon^{2l} + \ldots + v_{n-2}\varepsilon^{(n-1)l} \\
&= \lambda_l \varepsilon^l \\
&= \left(v_0 + \varepsilon^l v_1 + \varepsilon^{2l} v_2 + \ldots + \varepsilon^{(n-1)l} v_{n-1}\right)\varepsilon^l \\
&= \varepsilon^l v_0 + \varepsilon^{2l} v_1 + \varepsilon^{3l} v_2 + \ldots + v_{n-1}.
\end{aligned}$$

Observe that $\varepsilon^{nl} = (\varepsilon^n)^l = 1$. If *circ* (n) is the set of all circulant matrices, *circ* $(n) = \left\{circ(v); v \in \mathbb{C}^n\right\}$ then the algebraic properties of this set are given in the theorem 1.

Theorem 1

a) *circ* (n) is a commutative subalgebra of the $n \times n$ matrix algebra relative to the usual sum and matrix product
b) Transpose and inverses of circulant matrices are also circulants
c) All matrices in *circ* (n) are simultaneous diagonalized by the same matrix

$$C = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 1 & \varepsilon & \cdots & \varepsilon^{n-2} & \varepsilon^{n-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \varepsilon^{n-2} & \cdots & \varepsilon^{(n-2)^2} & \varepsilon^{(n-2)(n-1)} \\ 1 & \varepsilon^{n-1} & \cdots & \varepsilon^{(n-2)(n-1)} & \varepsilon^{(n-2)^2} \end{bmatrix}.$$

(the k-th column is obtained from the anterior column by multiplication following the rule: the i-th row element in the previous column is multiplied by ε^{i-1}). The elements of the matrix C are given generically by $C=\left(c_{ij}=\varepsilon ij, for\, i, j=0,1,\ldots,n-1\right)$.

We will not present a demonstration of these properties. For this, see Kra and Simanca (2012). We point out only the observations:

1. The matrix C is unitary, that is, $CC^* = C^*C = I$, where C* is the conjugate transpose of C. Indeed:

$$U=CC^*=\left(u_{ij}\right). \tag{21}$$

$$u_{ij}=\frac{1}{n}\sum_{s=1}^{n}\varepsilon^{is}\varepsilon^{-js}=\frac{1}{n}\sum_{s=1}^{n}\varepsilon^{(i-j)s}. \tag{22}$$

   If $i \neq j$, ε^{i-j} is *n-th* complex root of unity then the sum $\sum_{s=1}^{n}\varepsilon^{(i-j)s}=0$ and if $i = j$ then $u_{ii} = 1$.

2. If V = circ (v) then as C diagonalize all matrix in circ (n), we have:

$$C^*VC=D_v, \tag{23}$$

   where D_v is a diagonal matrix with elements $\lambda_0,\lambda_1,\ldots,\lambda_{n-1}$.

In the article Golub *et al*. (1979), the complex root of unity chosen was $\varepsilon=e^{\frac{2\pi}{n}i}$ and the unitary matrix denoted $W=\left(\frac{1}{\sqrt{n}}e^{\left(\frac{2\pi j}{n}k\right)i}\right)$, for $j, k = 1,\ldots,n$.

GCV Formula for Linear Regression

Consider again the linear regression $y = X\beta + \varepsilon$. Let $X_{n\times p} = U_{n\times n}D_{n\times p}V'_{p\times p}$ the singular value decomposition of matrix X (we are supposing X with complete rank, $n > p$) (Golub and Reinsch (1970)).

The singular value decomposition of the matrix $X_{n\times p} = U_{n\times n}D_{n\times p}V'_{p\times p}$ allows us to make a transformation in the data that simplify our linear regression (Fig. 1):

$$\begin{aligned} &y_{n\times 1}=X_{n\times p}\beta_{p\times 1}+\varepsilon_{n\times 1}\\ &\Rightarrow y_{n\times 1}=U_{n\times n}D_{n\times n}V'_{p\times p}\beta_{p\times 1}+\varepsilon_{n\times 1}\\ &\Rightarrow U'_{n\times n}y_{n\times 1}=U'_{n\times n}U_{n\times n}D_{n\times n}V'_{p\times p}\beta_{p\times 1}+U'_{n\times n}\varepsilon_{n\times 1}\end{aligned}$$

As the matrix U is orthogonal, $U'_{n\times n}U_{n\times n}=I_{n\times n}$, we can consider the transformed data $\tilde{y}_{n\times 1}=U'_{n\times n}y_{n\times 1}$. Then we can suppose that the new model on this new data is defined by the matrix $\tilde{X}_{n\times p}=D_{n\times p}V'_{p\times p}$:

$$\tilde{y}_{n\times 1}=\tilde{X}_{n\times p}\beta_{p\times 1}+\varepsilon_{n\times 1}. \tag{24}$$

The least square estimator for β of this new model is the same as in the original model. We also have a relation between the related projections. For this, consider $H = X(X'X)^{-1}X'$ and $\tilde{H}=\tilde{X}\left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'$:

$$\begin{aligned} H&=X\left(X'X\right)^{-1}X'\\ &=UDV'\left(VDU'UDV'\right)^{-1}VDU'\\ &=UDV'\left(VDDV'\right)^{-1}VDU'\\ &=UDV'\left(\left(DV'\right)'DV'\right)^{-1}\left(DV'\right)'U'\\ &=U\tilde{X}\left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'U'\\ &=U\tilde{H}U'.\end{aligned} \tag{25}$$

As the two projections are defined by conjugates matrices they have the same eigenvalues an eingenvectors are related. Indeed, if $Hv = \alpha v$ and $w = U'v$:

$$U\tilde{H}w=U\tilde{H}U'v=Hv=\alpha v. \tag{26}$$

This imply that:

$$U'U\tilde{H}w=U'\alpha v\Rightarrow\tilde{H}w=\alpha U'v\Rightarrow\tilde{H}w=\alpha w. \tag{27}$$

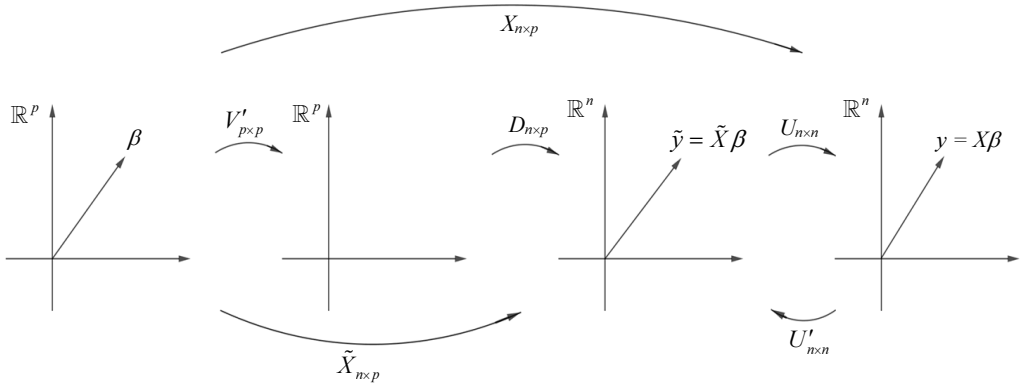

Fig. 1: The singular value decomposition of the matrix $X_{n\times p}$

Also follows that:

$$tr(H) = tr\left(U\tilde{H}U'\right) = tr\left(U'U\tilde{H}\right) = tr\left(\tilde{H}\right). \tag{28}$$

Then we can suppose that the model is always in the simplest form:

$$\tilde{y}_{n\times 1} = D_{n\times p}V'_{p\times p}\beta_{p\times 1} + \varepsilon_{n\times 1} \tag{29}$$

where, $\tilde{y}_{n\times 1} = U'_{n\times n}y_{n\times 1}$ are the modified data. To simplify the notation we will use just y in place of $\tilde{y}$ and $X = DV'$.

The idea is to make this model a complex model by new change of the data by W:

$$W_{n\times n}y_{n\times 1} = W_{n\times n}D_{n\times p}V'_{p\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}. \tag{30}$$

What this means? Let us see a case with $n = 3$:

$$\frac{1}{\sqrt{3}}\begin{bmatrix} 1 & 1 & 1 \\ 1 & \cos(120)+isen(120) & \cos(240)+isen(240) \\ 1 & \cos(240)+isen(240) & \cos(120)+isen(120) \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}.$$

The vector $y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$ is transformed on the vector with linear combinations:

$$\frac{1}{\sqrt{3}}\begin{bmatrix} y_1 + y_2 + y_3 \\ y_1 - \frac{1}{2}y_2 - \frac{1}{2}y_3 + i\left(\frac{\sqrt{3}}{2}y_2 - \frac{\sqrt{3}}{2}y_3\right) \\ y_1 - \frac{1}{2}y_2 - \frac{1}{2}y_3 + i\left(-\frac{\sqrt{3}}{2}y_2 - \frac{\sqrt{3}}{2}y_3\right) \end{bmatrix}.$$

In this way the observed values are transformed in many different linear combinations. The same occurs for the elements in matrix $X = DV'$.

Finally we have the complex model:

$$y = \tilde{X}\beta + \varepsilon, \tag{31}$$

where $\tilde{X}_{n\times p}$ is complex and $\tilde{X}_{n\times p} = W_{n\times n}D_{n\times p}V'_{p\times p} = W_{n\times n}X_{n\times p}$. From now to simplify notation we will drop the indexes for matrix dimensions.

The great advantage of this complexification follows from:

Theorem 2: $\tilde{X}\tilde{X}^*$ is a circulant matrix.

Proof:

Given any $n\times p$ matrix $\tilde{X}$, we may write $\tilde{X} = WDV' = WX$. The matrix D is an $n \times p$ diagonal matrix whose entries are the square roots of the eigenvalues of $X'X$. The number of non-zero entries in D is equal to the rank of X. Then:

$$\tilde{X}\tilde{X}^* = WDV'\left(WDV'\right)^* = WDV'VD'W^* = WDD'W^* \tag{32}$$

where:

$$DD' = \left[\begin{array}{ccc} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \\ \hline 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \end{array}\right]\left[\begin{array}{ccc|ccc} \lambda_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \lambda_p & 0 & \cdots & 0 \end{array}\right]$$

$$= \left[\begin{array}{ccc|c} \lambda_1^2 & \cdots & 0 & \\ \vdots & \ddots & \vdots & 0_{p\times(n-p)} \\ 0 & \cdots & \lambda_1^2 & \\ \hline \multicolumn{3}{c|}{0_{(n-p)\times p}} & 0_{(n-p)\times(n-p)} \end{array}\right].$$

If $DD' = diag\left(\lambda_1^2,...,\lambda_p^2,\lambda_{p+1}^2 = 0,\cdots,\lambda_n^2 = 0\right)$ then $WDD' = \left(\frac{1}{\sqrt{n}}\lambda_k^2 e^{\frac{(2\pi jk)}{n}i}\right)$, with $k = 1,...,n$.

In this way:

$$\begin{aligned} WDD'W^* &= \left(a_{jk}\right) \\ &= \left(\frac{1}{n}\sum_{s=1}^{n}\lambda_s^2 e^{\frac{(2\pi js)}{n}i}e^{-\frac{(2\pi ks)}{n}i}\right) \\ &= \left(\frac{1}{n}\sum_{s=1}^{n}\lambda_s^2 e^{\frac{2\pi s(j-k)}{n}i}\right). \end{aligned} \tag{33}$$

Thus:

$$\begin{aligned} \left(a_{j+1,k+1}\right) &= \frac{1}{n}\sum_{s=1}^{n}\lambda_s^2 e^{\frac{2\pi s(j+1-(k+1))}{n}i} \\ &= \frac{1}{n}\sum_{s=1}^{n}\lambda_s^2 e^{\frac{2\pi s(j-k)}{n}i} \\ &= a_{jk}, \end{aligned} \tag{34}$$

and, therefore, XX^* is a circulant matrix.

Theorem 3: $\tilde{X}\left(\tilde{X}^*\tilde{X}\right)^{-1}\tilde{X}^*$ is a circulant matrix.

Proof:

$$\begin{aligned} \tilde{X}\left(\tilde{X}^*\tilde{X}\right)^{-1}\tilde{X}^* &= WDV'\left(VDW^*WDV'\right)^{-1}VDW^* \\ &= WDV'\left(VD^2V'\right)^{-1}VDW^* \\ &= WD\left(V'\left(VD^2V'\right)^{-1}V\right)DW^* \\ &= WDD^{-2}DW^* \\ &= W\tilde{I}W^*, \end{aligned} \tag{35}$$

where:

$$\tilde{I} = \left[\begin{array}{c|c} I_{p\times p} & 0_{p\times(n-p)} \\ \hline 0_{(n-p)\times p} & 0_{(n-p)\times(n-p)} \end{array}\right].$$

From the proof in previous theorem the result follows. As $\tilde{X}\left(\tilde{X}^*\tilde{X}\right)^{-1}\tilde{X}^*$ is circulant its elements in the diagonal are constants, $a_{ii} = a_{11}$. We now can apply the PRESS formula for the regression $\tilde{y} = \tilde{X}\beta + \varepsilon$. We will denote this statistics as the GCV formula:

$$V = \left\| B\left(I - \tilde{X}\left(\tilde{X}^*\tilde{X}\right)^{-1}\tilde{X}^*\right)\tilde{y}\right\|^2, \tag{36}$$

where:

$$B = \begin{bmatrix} \frac{1}{1-a_{11}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{1-a_{nn}} \end{bmatrix} = \begin{bmatrix} \frac{1}{1-a_{11}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{1-a_{11}} \end{bmatrix}.$$

Therefore:

$$\begin{aligned} V &= \left(\frac{1}{1-a_{11}}\right)^2 \left\|\left(I - \tilde{X}\left(\tilde{X}^*\tilde{X}\right)^{-1}\tilde{X}^*\right)\tilde{y}\right\|^2 \\ &= \frac{1}{\left[\frac{1}{n}tr\left(I - \tilde{X}\left(\tilde{X}^*\tilde{X}\right)^{-1}\tilde{X}^*\right)\right]^2}\left\|\left(I - \tilde{X}\left(\tilde{X}^*\tilde{X}\right)^{-1}\tilde{X}^*\right)\tilde{y}\right\|^2. \end{aligned} \tag{37}$$

Denoting $\tilde{H} = \tilde{X}\left(\tilde{X}^*\tilde{X}\right)^{-1}\tilde{X}^*$. In Golub *et al*. (1979) this matrix is denoted $\tilde{A}$. Then:

$$V = \frac{\left\|\left(I - \tilde{H}\right)\tilde{y}\right\|^2}{\left[\frac{1}{n}tr\left(I - \tilde{H}\right)\right]^2}. \tag{38}$$

Proposition 1: The GCV formula V is a weighted version of PRESS statistics.

Proof:

$$\begin{aligned} \tilde{H} &= \tilde{X}\left(\tilde{X}^*\tilde{X}\right)^{-1}\tilde{X}^* \\ &= WDV'\left(VDW^*WDV'\right)^{-1}VDW^* \\ &= WDV'\left(VDDV'\right)^{-1}VDW^* \\ &= WDV'\left[\left(DV'\right)'DV'\right]^{-1}\left(DV'\right)'W^* \\ &= WX\left(X'X\right)^{-1}X'W^* \\ &= WHW^*. \end{aligned}$$

Thus, follows:

$$\begin{aligned} \left\|\left(I - \tilde{H}\right)\tilde{y}\right\|^2 &= \left\|\left(I - WHW^*\right)Wy\right\|^2 \\ &= \left\|\left(WW^* - WHW^*\right)Wy\right\|^2 \\ &= \left\|\left(W - WH\right)W^*Wy\right\|^2 \\ &= \left\|W\left(I - H\right)y\right\|^2 \\ &= \left\|\left(I - H\right)y\right\|^2. \end{aligned}$$

Note also that:

$$\begin{aligned} I - \tilde{H} &= I - WHW^* \\ &= WW^* - WHW^* \\ &= WIW - WHW^* \\ &= W\left(I - H\right)W^*, \end{aligned}$$

and:

$$\begin{aligned} tr\left(I - \tilde{H}\right) &= tr\left(W\left(I - H\right)W^*\right) \\ &= tr\left(\left(I - H\right)W^*W\right) \\ &= tr\left(I - H\right). \end{aligned}$$

Thus:

$$\begin{aligned} V &= \frac{\left\|\left(I - \tilde{H}\right)\tilde{y}\right\|^2}{\left[\frac{1}{n}\left(tr\left(I - \tilde{H}\right)\right)\right]^2} \\ &= \frac{\left\|\left(I - H\right)y\right\|^2}{\left[\frac{1}{n}\left(tr\left(I - H\right)\right)\right]^2} \\ &= \frac{\left\|\left(I - H\right)y\right\|^2}{\left[\frac{1}{n}\left(n - tr\left(H\right)\right)\right]^2} \\ &= \frac{\left\|\left(I - H\right)y\right\|^2}{\left(1 - \frac{p}{n}\right)^2} \\ &= \frac{1}{\left(1 - \frac{p}{n}\right)^2}\sum_{i=1}^{n}\left(\left(I - H\right)y\right)_i^2 \\ &= \frac{1}{\left(1 - \frac{p}{n}\right)^2}\sum_{i=1}^{n}\left(\frac{1 - h_{ii}}{1 - h_{ii}}\right)^2\left(\left(I - H\right)y\right)_i^2 \\ &= \sum_{i=1}^{n}\left(\frac{\left(I - H\right)y}{1 - h_{ii}}\right)_i^2\left(\frac{1 - h_{ii}}{1 - \frac{p}{n}}\right)^2. \end{aligned} \tag{39}$$

Proposition 2: The GCV statistic is an almost unbiased estimator of the prediction error.

Proof:

$$
\begin{aligned}
E\left[GCV\right] &= E\left[\frac{\left\|\left(I-\tilde{H}\right)\tilde{y}\right\|^2}{\left[\frac{1}{n}tr\left(I-\tilde{H}\right)\right]^2}\right] \\
&= \frac{1}{\left[\frac{1}{n}tr\left(I-\tilde{H}\right)\right]^2}E\left[\tilde{y}'\left(I-\tilde{H}\right)^*\left(I-\tilde{H}\right)\tilde{y}\right] \\
&= \frac{\sigma^2}{\left[\frac{1}{n}tr\left(I-H\right)\right]^2}tr\left[\left(I-H\right)^2\right] \\
&= \frac{\sigma^2}{\left[\frac{1}{n}tr\left(I-H\right)\right]^2}tr\left[\left(I-H\right)\right] \\
&= \frac{\sigma^2}{\left[\frac{1}{n}tr\left(I-H\right)\right]^2}tr\left[\left(I-H\right)\right] \\
&= n^2\sigma^2\frac{1}{tr\left(I-H\right)} \\
&= n^2\sigma^2\frac{1}{n-tr\left(H\right)} \\
&= n^2\sigma^2\frac{1}{n-p} \\
&= n^2\sigma^2\frac{1}{n\left(1-\frac{p}{n}\right)}.
\end{aligned} \tag{40}
$$

Using the Taylor approximation of first order, we have:

$$
\begin{aligned}
E\left[GCV\right] &= n^2\sigma^2\frac{1}{n\left(1-\frac{p}{n}\right)} \\
&= n\sigma^2\left(1+\frac{p}{n}+\frac{p^2}{n^2}+\ldots\right) \\
&\approx n\sigma^2\left(1+\frac{p}{n}\right) \\
&= \sigma^2\left(n+p\right).
\end{aligned} \tag{41}
$$

Note that GCV also overestimate the prediction error. The variance of the GCV statistic is:

$$
\begin{aligned}
\text{var}\left[GCV\right] &= \frac{2\sigma^2}{\left[\frac{1}{n}tr\left(I-H\right)\right]^4}tr\left(I-H\right) \\
&= 2n^4\sigma^2\frac{1}{\left[tr\left(I-H\right)\right]^3} \\
&= 2n^4\sigma^2\frac{1}{\left[n-tr\left(H\right)\right]^3} \\
&= 2n^4\sigma^2\frac{1}{n^3\left[1-\frac{tr\left(H\right)}{n}\right]^3} \\
&= 2n\sigma^2\frac{1}{\left[1-\frac{p}{n}\right]^3}.
\end{aligned} \tag{42}
$$

Some Computational Results

Both PRESS and GCV are estimators of the prediction error. How to choose between them? The one with less variance must be chosen. We wasn't able to proof analytically that the variance of GCV is less than variance of PRESS. Then a computational simulation was done to evaluate of expectations and variances of the PRESS and GCV.

This simulation study was based on the example conceived by Zou *et al*. (2006). The authors considered two variables, which they named "hidden factors". They are:

$$V_1 \sim N\left(0,290\right),\ \ V_2 \sim N\left(0,300\right),$$

where $\varepsilon \sim N\left(0,1\right)$, with V_1, V_2 and ε independents.

Thus, 6 variables were constructed from V_1 and V_2 as follows:

$$
\begin{aligned}
X_i &= V_1 + \varepsilon_i, \varepsilon_i \sim N\left(0,1\right), i = 1,2,3 \\
X_i &= V_2 + \varepsilon_i, \varepsilon_i \sim N\left(0,1\right), i = 4,5,6
\end{aligned}
$$

where $\{\varepsilon_i\}$ are independents, i = 1,...,6.

With these variables we generated N = 1000 matrices of dimensions $n \times 6$, with n equal to 10, 30, 50, 100 and 200. We considered the linear models $y = X\beta + \varepsilon$, $\varepsilon \sim N(0,1)$.

Then, we calculated the bias for each statistics, given by:

$$bias\left(PRESS\right) = \sigma^2\sum_{i=1}^{n}\frac{1}{1-h_{ii}} - \sigma^2\left(n+p\right). \tag{43}$$

$$bias\left(GCV\right) = n\sigma^2\frac{1}{1-\frac{p}{n}} - \sigma^2\left(n+p\right). \tag{44}$$

The variances of PRESS and GCV are respectively:

$$\text{var}\left[PRESS\right] = 2tr\left[B^2\left(I-H\right)B^2\left(I-H\right)\right]\sigma^2. \tag{45}$$

$$\text{var}\left(GCV\right) = 2n\sigma^2\frac{1}{\left(1-\frac{p}{n}\right)^3}. \tag{46}$$

In the Table 1 presents the means of bias and statistical variance in the N = 1000 simulations, considering all values of n.

It can be seen in Table 1 that the bias of the GCV statistic was always lower than the bias of the PRESS statistic for all sample sizes adopted. However, as the sample size was increased, it was found that both statistics tend to have less bias. This result corroborates the theoretical results presented in the present work, that these statistics are almost unbiased. It can also be observed that for each model, the variance of GCV was smaller than the variance of PRESS.

Table 1: Prediction Error (PE), bias and variances from Prediction Sum of Squares (PRESS) and Generalized Cross-Validation (GCV) statistics considering $N = 1000$ simulations for 10, 30, 50 100 and 200 observations.

		PRESS		GCV	
n	PE	Bias	Variance	Bias	Variance
10	16	27,21	11212,84	9,00	312,50
30	36	2,18	128,80	1,50	117,19
50	56	1,14	151,50	0,82	146,74
100	106	0,52	242,62	0,38	240,79
200	206	0,25	439,08	0,19	438,27

Conclusion

The linear algebra theory developed in this article, necessary for the development of the GCV formula for linear models, although not very simple, is accessible to an audience of non-specialists who use cross-validation techniques. This article intends to fill a gap as an accessible reference to the subject. That PRESS and GCV statistics are almost unbiased statistics of prediction error maybe are new results.

Acknowledgment

The authors gratefully acknowledge with thanks the very thoughtful and constructive suggestions and comments of the Editor-in-Chief and the reviewers which resulted in much improved paper.

Author's Contributions

Authors contributed to the same extent to all the process of preparing and developing the manuscript since we operate as a group.

Ethics

The authors declare that there is no conflict of interests regarding the publication of this article.

References

Allen, D.M., 1971. Mean square error of prediction as a criterion for selecting variables. Technometrics, 13: 469-475. DOI: 10.1080/00401706.1971.10488811

Allen, D.M., 1974. The relationship between variable selection and data augmentation and a method for prediction. Technometrics, 16: 125-127. DOI: 10.1080/00401706.1974.10489157

Efron, B., 2004. The estimation of prediction error: Covariance penalties and cross-validation. Theory Meth., 99: 619-642. DOI: 10.1198/016214504000000692

Golub, G.H., M. Heath and G. Wahba, 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics, 21: 215-223. DOI: 10.1080/00401706.1979.10489751

Golub, G.H. and C. Reinsch, 1970. Singular value decomposition and least squares solutions. Numer. Math., 14: 403-420. DOI: 10.1007/BF02163027

Henderson, H.V. and S.R. Searle, 1981. On deriving the inverse of a sum of matrices. Siam Rev., 23: 53-60. DOI: 10.1137/1023004

Kra, I. and S.R. Simanca, 2012. On circulant matrices. Siam Rev., 59: 368-377. DOI: 10.1090/noti804

Rencher, A.C. and G.B. Schaalje, 2008. Linear Models in Statistics. 2nd Edn., John Wiley and Sons. New Jersey, ISBN-10: 9780471754985, pp: 688.

Zou, H., T. Hastie and R. Tibshirani, 2006. Sparse principal component analysis. J. Comput. Grahical Stat., 15: 265-286. DOI: 10.1198/106186006X113430