Volume: 08 Number: 03 Category: 02 Page: 75-85 ISSN No: 1303-6319 Cilt: 08 Sayı: 03 Kategori: 02 Sayfa: 75-85

GENETIC ALGORITHM BASED VARIABLE SELECTION FOR PARTIAL LEAST SQUARES REGRESSION

Özlem GÜRÜNLÜ ALMA* **Elif BULUT****

ABSTRACT

Partial Least Squares (PLS) regression has been an alternative to ordinary least squares for handling multicollinearity in several areas of scientific research. At the core of the PLS methodology lies a dimension reduction technique coupled with a regression model. In this paper, we investigate the genetic algorithms-partial least square regression (GAPLSR). This technique combines genetic algorithms as powerful optimization methods with PLS as a statistical method for variable selection. Variable importance for projection is a weighted sum of squares of the PLS-weights and thus a summary of the importance of a variable for the modeling of both X and Y (Wold et al., 2001). In this study, comparisons of R^2_{adj} *values of GAPLSR predicting model, PLSR-NIPALS model and significant model PLSR-VIP were established according to the VIP scores of PLSR model to see which one has established a model with less error.*

Keywords: Genetic algorithms, Partial least square regressions, Variable selection, Variable importance for projection.

1. INTRODUCTION

Variable selection is used for improving the model performance and give better predictions. Improvement of statistical properties can also be a reason for doing variable selection. In some situations, the purpose of variable selection is to obtain a model that is easier to understand, for example, by getting rid of all the variables that do not contribute positively to the model. This may not give better predictions. Also, variable selection can be relevant to reduce the risk of overfitting or for computational reasons. A plausible way to do variable selection is to try all combinations of variables and select the best ones. This sounds simple, but is, in practical, impossible for a number of reasons. The selection of the most adequate regression model can be stated as an optimization problem with the objective to select those independent variables that maximize the adequacy of the model according to a statistical criterion (Paterlini and Minevra, 2010).

Partial least squares regression (PLSR) differs from traditional regression tools in the way that regression coefficients estimators are constructed. These are constructed through the use of latent variables which maximize the covariance between predictors and the explanatory variable. This construction follows an iterative procedure to ensure that the latent variables are orthogonal (Vitor et al., 2000).

Today, it has been widely accepted that a feature selection has some advantages. Although PLS is a well-working method to model high dimensional and collinear datasets, the interpretation and understanding of the predictive model and its results are more difficult (Wold et al., 1996).

*Assist. Prof., Department of Statistics, Muğla University, Muğla, e-posta: ozlem.gurunlu@hotmail.com
**Assist. Prof., Department of Business Administration, Ondokuz Mayıs University, Samsun, e-posta: bulut_elif@yahoo.com

Feature selection can also help to built a better predictive PLS model with fewer features (Kubinyu, 1996). The PLS regression combined with the Variable Importance for Projection (VIP) scores is often used when the multicollinearity is present among variables; however, there are few guidelines about its uses as well as its performance (Chong and Jun, 2005) (Gurunlu and Bulut, 2012).

Another approach to select variables is to apply an optimization algorithm such as genetic algorithms, since the problem of variable selection can be formulated as a combinatorial optimization problem. A genetic algorithm (GA) is a technique somewhat inspired by the theory of evolution. It mimics the selection in nature by evaluating models consisting of certain combinations of variables in a number of generations (Andersen and Bro, 2010).

The purpose of this paper is to explore the nature of the spell out (GAPLSR) method and to compare with the PLS regression (called PLS-NIPALS: PLS-Non-Linear Iterative Partial Least Squares method), and the PLSR-VIP methods, and also to investigate the performance of the VIP scores for selecting the relevant process variables which really have an effect on the response. For this purpose, we used computer simulation experiments where some true models are assumed and data sets are generated. We compare the performance of GAPLSR scores with the PLS-NIPALS and the PLSR-VIP methods. The rest of the paper is organized as follows. A brief review of variable selection methods using PLS regression, PLS-VIP and GA-PLSR methods are given in Sections 2 and 3, respectively. Section 4 contains the design of simulation experiments and their results are presented in Section 5. Concluding remarks take place in the last section.

2. PLSR MODEL AND VARIABLE SELECTION CRITERIA FOR PLSR

PLSR is a latent variable based multivariate statistical method, which forms from the combination of PLS and multiple linear regression (Martens and Naes, 1989) (Gurunlu and Bulut, 2012). In PLS regression and in all algorithms $\mathbf{X}_{N\times K}$ represents the data matrix of N observation units on K explanatory variables and $\mathbf{Y}_{N\times M}$ represents the data matrix of N observation units on M response variables (Gurunlu and Bulut, 2012). The intension of PLSR is to form components that capture most of the information in the **X** matrix, which is useful for predicting response variables, while reducing the dimensionality of the regression problem by using fewer components than the number of **X** variables (Garthwaite, 1994) (Gurunlu and Bulut, 2012). In PLS regression analysis, many algorithms are used to obtain the latent variables. The objective of all linear PLSR algorithm is to project the data down onto a number of latent variables ($\mathbf{t_a}$ and $\mathbf{u_a}$), and then, to develop a regression model between these variables. It uses both the variation of **X** and **Y** to construct the latent variables (Gurunlu and Bulut, 2012). Algorithms work with different sets of variables by maximizing the covariance between them. For the convenience of the calculations and not to be time consuming, the choice of algorithm depends on the shape of the matrices. For example, if there are many observations and few variables, it is better to work with a data matrix that dimensions depend on the number of variables. An often used algorithm is the NIPALS (Non-Linear Iterative Partial Least Squares) algorithm, which is often referred to as the classical algorithm. The development was initiated by Jöreskog and Wold (1982); Wold (1966) (Gurunlu and Bulut, 2012). Later it was extended by Lindgren and Rannar (1998), Wold et al. (1983), and Wold et al. (1996).

NIPALS algorithm composes of two loops. The inner loop is used to attain $\mathbf{t_a}$ and $\mathbf{u_a}$ $(a = 1,..., A)$ latent variables, where A is the number of the latent variables. Then, convergence is tested on the change in $\mathbf{u}$. If convergence has been reached, the outer loop is used sequentially to extract $\mathbf{p_a}$, $\mathbf{q_a}$ from $\mathbf{X}$ and $\mathbf{Y}$ matrices. In this algorithm a regression model between latent variables is written as follows (Gurunlu and Bulut, 2012):

$$\mathbf{u}_a = b_a \mathbf{t}_a + \varepsilon_a \quad a=1,\ldots,A \tag{1}$$

where ε_a is vector of errors and b_a is an unknown parameter estimated by $\hat{\mathbf{b}}_a=(\mathbf{t}_a'\mathbf{t}_a)^{-1}\mathbf{t}_a'\mathbf{u}_a$. The latent variables are computed by $\mathbf{t_a} = \mathbf{X_a w_a}$ and $\mathbf{u_a} = \mathbf{Y_a q_a}$, where both $\mathbf{w}_a$ weight vector for $\mathbf{X}$ and $\mathbf{q_a}$ loading vector for $\mathbf{Y}$ have unit lengths and are determined by maximizing the covariance between $\mathbf{t}_a$ and $\mathbf{u_a}$. $\mathbf{X}$ and $\mathbf{Y}$ data matrices are deflated at the end of each iteration as $\mathbf{X_{a+1}} = \mathbf{X_a} - \mathbf{t_a p_a'}$ where $\mathbf{X}_1 = \mathbf{X}$ and $\mathbf{p_a} = \mathbf{X_a't_a}/(\mathbf{t_a't_a})$, $\mathbf{q}_a = \mathbf{Y_a't_a}/(\mathbf{t_a't_a})$ and $\mathbf{Y_{a+1}} = \mathbf{Y_a} - b_a \mathbf{t_a q_a'}$ where $\mathbf{Y}_1 = \mathbf{Y}$ to be used in the next iteration (Gurunlu and Bulut, 2012). Letting $\hat{\mathbf{u}}_\mathbf{a} = \hat{b}_a \mathbf{t_a}$ be the prediction of $\mathbf{u_a}$, the matrices $\mathbf{X}$ and $\mathbf{Y}$ can be decomposed as the following (Li et al., 2002):

$$\mathbf{X}=\sum_{a=1}^{A}\mathbf{t_a p_a'}+\mathbf{E}, \quad \text{and} \quad \mathbf{Y}=\sum_{a=1}^{A}\hat{\mathbf{u}}_\mathbf{a}\mathbf{q_a'}+\mathbf{F}, \tag{2}$$

where $\mathbf{E}$ and $\mathbf{F}$ are the residuals of $\mathbf{X}$ and $\mathbf{Y}$ after extracting the first "A" pairs of latent variables (Gurunlu and Bulut, 2012).

The objective of variable selection is three-fold: improving the prediction performance of the model, providing faster and more cost-effective predictors, and also providing a better understanding of the underlying process that generated the data (Guyon and Elisseef, 2003) (Gurunlu and Bulut, 2012).

The VIP value, which was derived from the PLS, was considered as a variable selection procedure. It is a statistic of summarizing the contribution that a variable makes to a model (Wold, 1994; Wold et al., 2001). It gives the value of each explanatory variable in fitting the PLS model for both explanatory and response variables. The VIP scores and the beta coefficients that are obtained by PLS regression can be used to select the most influential variables (Chong and Jun, 2005) (Gurunlu and Bulut, 2012). The VIP score can be estimated for the j^{th} explanatory variable by the following formula,

$$\mathrm{VIP}_j = \sqrt{K \times \frac{\sum_a w_{ja}^2 b_a^2 \mathbf{t}_a'\mathbf{t}_a}{\sum_a b_a^2 \mathbf{t}_a'\mathbf{t}_a}} \tag{3}$$

where w_{ja} is a weight of the j^{th} X-variable to the a^{th} latent variable which is obtained by NIPALS algorithm (Jun et al., 2009). Weight values can be interpreted as the contribution of the j^{th} explanatory variable to the a^{th} latent variable. The VIP score greater than or equal to one rule is generally used as a criterion for variable selection (Chong and Jun, 2005) (Gurunlu and Bulut, 2012).

3. VARIABLE SELECTION FOR PLSR MODELS USING GENETIC ALGORITHMS

GA is a search technique used to find true or approximate solutions to optimization and search problems. They belong to a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover (Goldberg, 1989) (Gurunlu and Bulut, 2012). Details of the algorithm can be found elsewhere (Leardi et al., 1992; Leardi, 1996).

It is interesting to notice that several authors have published papers about feature selection by GAs, each of them using a different GA structure, sometimes rather far from the standard algorithm. This demonstrates the need to modify the algorithm according to the peculiarities of the problem to be solved. In the case of feature selection, for instance, a chromosome is made by a very high number of genes (as many as the variables), each of them being just 1 bit long (0 = variable absent, 1 = variable present) (Gurunlu and Bulut, 2012). Leardi et al. (1992) used a simulated data set to show that a GA always find the global maximum of a simple problem in a time much shorter than the time required for a full search. Lucasius et al. (1994) showed that a GA generally performs better than simulated annealing and stepwise regression; on the other hand, Hörchner and Kalivas (1995) demonstrated that simulated annealing can give the same results as Leardi (2001 (Gurunlu and Bulut, 2012)).

The performance of the regression model, which is usually represented as the root mean square error (RMSE), is optimized by GA procedure (Gurunlu and Bulut, 2012). It was reported that the GA-based methods could effectively reduce the number of variables and produce predictive models. However, resultant models tend to be not intuitive because variables are selected independently (Masamoto et al., 2011 (Gurunlu and Bulut, 2012)).

In this study, the GA is made up of a number of steps. First, a vector consisting of zeros and ones is constructed with the size corresponding to the number of variables. It is denoted a chromosome. The randomly defined zeros and ones represent the variables that should be included. Details on the algorithm used can be found in Leardi et al. (1992) and Leardi (1996). Each zero or one is a gene and a PLS model constructured with the chosen genes is defined as an individual. Each model, also called a chromosome, is fully described by a binary vector "d", $d = (d_1, ...d_K)$, where d_i=0 indicates that i^{th} explanatory variable is not selected and d_i=1 indicates that i^{th} explanatory variable is selected for the PLSR model, where $i = 1, ..., K$. In this study, the structure of a chromosome is shown in Figure 1.

	d_1	d_2	d_3	…	d_{K-1}	d_K
d=	1	1	0	…	1	0

Figure 1. The Structure of a chromosome (c).

The main characteristics of GAs used in the study can be listed as follows:

- response to be maximized to explained variance (%);
- Regression method: PLSR

- Fitness function**:** Every candidate solution is evaluated with respect to a fitness function. The performance of the GA is measured by comparing, RMSE which is defined as (Leardi and Gonzalez, 1998):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}}$$, where N is the number of objects in the evaluation set.

- Population size: 30 chromosomes.
- Average number of variables selected in the chromosomes of the starting population: as number of latent variables.
- Selection function: stochastic uniform selection function is used in GA. This function lays out a line in which each parent corresponds to a section of the line of length proportional to its scaled value.
- Cross-over method: uniform probability.
- Mutation probability: 1%.
- Population update: one pair of chromosomes of the existing population is selected by a random (biased) selection; after cross-over and mutation, two offsprings are obtained and evaluated; each of them enters the population if it is better than the worst chromosome, which is discarded (the exceptions to this rule are described in the next point); this is the highest possible elitism since the components of the final population are the best chromosomes found; due to the fact that a new generation is composed by just two chromosomes, it is better to refer to the number of chromosomes evaluated rather than to the generations.
- Subset check: chromosome A cannot exist (is discarded) if the variables selected by another chromosome (B) are a subset of the variables selected by chromosome A, and B has a response higher than A.

4. DESIGN OF SIMULATION STUDY

The framework for the simulation models was based on the study of Li et al. (2002); Naes and Martens (1985). It was extended in this paper to the situation where there exists multiple response variables and different number of explanatory variables. In the simulation study, the multivariate regression models were first developed from which data was generated, and then GAPLSR, PLSR-NIPALS, and PLSR-VIP methods were applied. The resulting models were then compared with R^2_{adj} values (Li et al., 2002) (Gurunlu and Bulut, 2012). The **X** and **Y** block data, with sample size N, were generated as:

$$\mathbf{X}=\sum_{i=1}^{A^*}\mathbf{r}_i\boldsymbol{\xi}_i'+\tilde{\mathbf{E}}, \quad (4)$$

$$\mathbf{Y}=\sum_{i=1}^{A^*}\mathbf{z}_i\boldsymbol{\eta}_{A^*i}'+\boldsymbol{\psi}=\sum_{i=1}^{A^*}\mathbf{r}_i\boldsymbol{\eta}_{A^*i}'+\tilde{\mathbf{F}}_{A^*} \quad (5)$$

where $\tilde{E}$ and r_i were generated from mutually independent normal variables. Generation of **X** and **Y** data matrices are just explained for 5*3, which means that the number of predictor variables is 5 and these variables are reduced to a number of 3 latent variables.

$\boldsymbol{\psi}$ was generated from a multivariate normal distribution and generated as Li et al. (2002). $\tilde{\mathbf{F}}$ is a noise matrix, and $\mathbf{Z}$ was constructed as $\mathbf{z_i}=\mathbf{r_i}+\mathbf{f_i}$, $\mathbf{f_i}$ were generated as independent normal variables with zero means and different variances (0.5, 0.25 and 0.1) (Gurunlu and Bulut, 2012). $\{\xi_i\}$ and $\{\boldsymbol{\eta}_{A^*i}\}$ are normalized orthogonal vector series, and $\mathbf{r}_i$ are mutually independent random variables with zero means and variances (15, 7.5 and 3). To carry out simulation runs, it is preceded on different simulations. The dimensions of explanatory variables is extended as $N\times5$, $N\times8$, $N\times10$, and $N\times12$. The dimension of response variables matrix, $\mathbf{Y}$, is chosen as $N\times3$, $N\times4$, and sample sizes are selected as N=50, 100, 250, 500 (Gurunlu and Bulut, 2012). For each combinations, 100 data sets are generated taking the dimension of PLSR models into account and sample sizes so that 16×100 data sets are generated. It is seen that the variance inflation factor (VIF) values for 5*3 design matrix show that there is multicollinearity; the VIF values are calculated by Minitab package program. The relative cumulative variances by the five latent variables for the $\mathbf{X}$ and $\mathbf{Y}$ blocks, averaged over 100 simulation experiments show that the optimum latent variable number is $A^*=3$. That is, first three latent variables capture 100% and 98% of the variances in the $\mathbf{X}$ and $\mathbf{Y}$ data sets, respectively. This verifies the theoretical value of the number of latent variables $A^*=3$. For more information one may refer to Li et al. (2002). GAPLSR, PLSR-NIPALS, and PLSR-VIP methods are applied to these data sets.

5. RESULTS

In this paper, it a simulation study was hold to gain a better understanding of the performances of GAPLSR, PLSR-NIPALS, and PLSR-VIP methods for PLSR model selection. An experimental simulation study was designed to see which one has established a model with less error. In this study, EM (median of R^2_{adj}) and $\overline{R}^2_{adj}$ values are obtained for each N with 100 iterations. Because of many data sets, it is convenient to work with $\overline{R}^2_{adj}$ as a performance criterion. All of these methods have different study structures. GAPLSR finds variables and then develops models on these variables. These models have the minimum RMSE so have the biggest R^2_{adj} values. PLSR-NIPALS works with latent variables. It finds regression coefficients on latent variables and develops regression models on these latent variables. PLSR-VIP also works with latent variables. Firstly, it finds VIP values by the help of PLS and then works with explanatory variables which have VIP values greater than 1 in building the regression models and calculating R^2_{adj} values. All of the methods work approximately with same number of explanatory variables or latent variables. These numbers on the average are equal to A^*. The results show that $\overline{R}^2_{adj}$ values for models with GAPLSR have the biggest values for each of the design matrices and for each N. PLSR-VIP finds bigger numbers for the number of explanatory variables for the higher design matrix. For that reason it has high $\overline{R}^2_{adj}$ values for higher design matrix. This study shows that variable selection with GAPLSR method gives better results than selection with PLSR-VIP. Table 1 shows the results obtained by applying the three selection methods on the simulated datasets. Wilcoxon signed rank test (Hollander and Wolfe, 1973) was used for comparing the results obtained by the three methods. This test statistic is useful for evaluating the differences of paired samples of subjects. The null hypothesis is retained when the median of the differences between predicted values for each sample object by

the two methods is zero, and the alternative hypothesis is retained when the median of the selection method is bigger than the value 0.8. The test statistics z is calculated by dividing the sum of the signed ranks (U) by the square roots of the sum of squares of the signed ranks (S). S is the standard deviation of U. The null hypothesis that the median difference is zero is assessed by comparing the test statistics U/S to critical values of the standard normal distribution. To definitely compare the performance of the two selection methods, the Wilcoxon signed rank test with α=0.05 (95% confidence level) was used to compare R^2_{adj} (Shariati-Rad and Hasani, 2010).

Table 1. Comparion results of gaplsr, plsr-nipals, and plsr-vip methods

			5*3			8*4			10*4			12*4		
		N	p	EM	$\overline{R}^2_{adj}$	p	EM	$\overline{R}^2_{adj}$	p	EM	$\overline{R}^2_{adj}$	p	EM	$\overline{R}^2_{adj}$
GA	R^2_{adj1}	50	0.0	0.96	0.95	0.0	0.97	0.97	0.0	0.97	0.96	0.0	0.97	0.97
		100	0.0	0.96	0.95	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		250	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		500	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
	R^2_{adj2}	50	0.0	0.96	0.95	0.0	0.97	0.97	0.0	0.97	0.96	0.0	0.97	0.97
		100	0.0	0.96	0.95	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		250	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		500	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
	R^2_{adj3}	50	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.96	0.0	0.97	0.97
		100	0.0	0.96	0.93	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		250	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		500	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
	R^2_{adj4}	50				0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		100				0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		250				0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		500				0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
NIPALS-PLSR	R^2_{adj1}	50	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.96	0.97
		100	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		250	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		500	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
	R^2_{adj2}	50	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.96	0.97
		100	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		250	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		500	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
	R^2_{adj3}	50	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.96	0.97
		100	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		250	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		500	0.0	0.96	0.96	0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
	R^2_{adj4}	50				0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.96	0.97
		100				0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		250				0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97
		500				0.0	0.97	0.97	0.0	0.97	0.97	0.0	0.97	0.97

Table 1. Comparion results of gaplsr, plsr-nipals, and plsr-vip methods (Continued)

		50	0.04	0.93	0.94	0.04	0.92	0.92	0.04	0.92	0.91	0.06	0.90	0.91
		100	.04	0.93	0.94	0.04	0.92	0.92	0.05	0.92	0.92	0.05	0.91	0.91
	R^2_{adj1}	250	0.02	0.94	0.94	0.04	0.93	0.92	0.06	0.90	0.92	0.06	0.91	0.91
		500	0.02	0.94	0.94	0.05	0.92	0.92	0.06	0.91	0.91	0.06	0.90	0.91
	R^2_{adj2}	50	0.03	0.93	0.93	0.04	0.92	0.92	0.06	0.92	0.91	0.06	0.90	0.91
		100	0.02	0.93	0.93	0.04	0.92	0.92	0.07	0.91	0.91	0.04	0.92	0.91
		250	0.02	0.94	0.93	0.04	0.93	0.92	0.06	0.91	0.91	0.06	0.91	0.91
		500	0.02	0.94	0.94	0.05	0.92	0.92	0.06	0.90	0.91	0.06	0.90	0.91
VIP														
	R^2_{adj3}	50	0.02	0.93	0.94	0.023	0.93	0.93	0.00	0.92	0.91	0.00	0.92	0.90
		100	0.02	0.94	0.95	0.005	0.93	0.92	0.00	0.92	0.90	0.00	0.91	0.91
		250	0.04	0.93	0.94	0.001	0.92	0.92	0.00	0.90	0.90	0.07	0.90	0.89
		500	0.02	0.94	0.94	0.00	0.93	0.92	0.00	0.91	0.91	0.00	0.91	0.91
	R^2_{adj4}	50				0.00	0.92	0.79	0.00	0.95	0.87	0.00	0.95	0.91
		100				0.00	0.92	0.80	0.00	0.95	0.92	0.00	0.95	0.94
		250				0.00	0.93	0.81	0.00	0.95	0.94	0.00	0.95	0.94
		500				0.00	0.93	0.80	0.00	0.95	0.94	0.00	0.96	0.95

p: p value, EM: estimated Median by Wilcoxon Sign rank test, $\overline{R}^2_{adj}$: Mean of adjusted determination coefficient.

6. CONCLUSIONS

In this paper, a simulation study was hold to gain a better understanding of the performances of GAPLSR, PLSR-NIPALS, and PLSR-VIP methods for PLSR model selection; it was run a designed experimental simulation study to see which one has established a model with less error and have the biggest R^2_{adj} values.

The PLS-VIP method performed excellently in identifying relevant predictors and outperformed the other methods. It was also found that a model with good fitness performance may not guarantee good variable selection performance. Thus, for the purpose of selecting relevant process variables, investigators must be careful when using model performance criteria such as RMSEP, R^2_{adj}, etc. Second, the GAPLSR method was compared with the PLS-VIP and the PLSR-NIPALS method. We found an interesting observation that GAPLSR and PLSR-NIPALS method might be complementary. Hence, if we use a strategy which combines these two methods for selecting relevant predictors, a better variable selection performance could be achieved. Actually, Wold et al. (1993) recommended a combination of PLS-VIP and PLS-Beta for variable selection, which stated that both should be small for a variable to be excluded (Chong and Jun, 2005).

7. REFERENCES

Andersen, C. M., Bro, J. R., 2010. Variable Selection in Regression - A Tutorial. Chemometrics, 24, 728-737.

Chong, Il-G., Jun, C. H., 2005. Performance of Some Variable Selection Methods when Multicollinearity is Present. Chemometrics and Intelligent Laboratory Systems, 78, 103-112.

Garthwaite, P. H., 1994. An Interpretation of Partial Least Squares. Journal of the American Statistical Association, 89, 122-127.

Gurunlu Alma Ö., Bulut E., 2012. Genetic Algorithm Based Variable Selection for Partial Least Squares Regression Using ICOMP Criterion, Asian Journal of Mathematics and Statistics, 5(3), 82-92.

Guyon, I., Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3, 1157-1182.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, USA.

Hollander, M., Wolfe, D. A., 1973. Nonparametric Statistical Methods. John Wiley & Sons: New York, NY.

Hörchner, U., Kalivas, J. H., 1995. Further Investigation on a Comparative Study of Simulated Annealing and Genetic Algorithm for Wavelengths Selection. Analytica Chimica Acta, 311, 1-13.

Jöreskog, K. G., Wold, H., 1982. Systems Under Indirect Observation, Part I, 263-270. Amsterdam, New York, Oxford: North-Holland.

Jun, C. H., Lee, S. H., Park, H. S., Lee, J. H., 2009. Use of Partial Least Squares Regression for Variable Selection and Quality Prediction. Computers & Industrial Engineering. CIE 2009. International Conference on 6-9 July 2009.

Kubinyu H., 1996. Evolutionary Variable Selection in Regression and PLS Analyses. Journal of Chemometrics, 10, 110-133.

Leardi, R., Boggia, R., Terrile, M., 1992. Genetic Algorithms as a Strategy for Feature Selection, Journal of Chemometrics, 6, 267-281.

Leardi, R., 1996. Genetic Algorithms in Feature Selection, in: J. Devillers_Ed., Genetic Algorithms in Molecular Modeling, Academic Press., 67.

Leardi, R., 2001. Genetic Algorithms in Chemometrics and Chemistry: A Review. Journal of Chemometrics, 15, 559-569.

Leardi, R., Gonza´lez, A. L., 1998. Genetic Algorithms Applied to Feature Selection in PLS Regression: How and When to Use Them. Chemometrics and Intelligent Laboratory Systems, 41, 195–207.

Li, B., Morris, J., Martin, E. B., 2002. Model Selection for Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems, 64, 79-89.

Lindgren, F., Rännar S., 1998. Alternative Partial Least-Squares (PLS) Algorithms. Perspectives in Drug Discovery and Design. 12-14, 105-113.

Lucasius, C. B., Beckers, M. L. M., Kateman, G., 1994. Genetic Algorithms in Wavelengths Selection: A Comparative Study. Analytical Chimica Acta, 286, 135–153.

Masamoto, A., Yosuke, Y., Kimito, F., 2011. Genetic Algorithm-Based Wavelength Selection Method for Spectral Calibration. Journal of Chemometrics, 25, 10–19.

Martens H., Naes T., 1989. Multivariate Calibration. John Wiley & Sons.

Naes, T., Martens, H., 1985. Comparison of Prediction Methods for Collinear Data, Communication in Statistics Simulation and Computation, 14, 545-576.

Paterlini, S., Minerva, T., 2010. Regression Model Selection using Genetic Algorithms. Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing, WSEAS Press Stevens Point, Wisconsin, 19-28.

Shariati-Rad, M., Hasani, M., 2010. Selection of Individual Variables versus Intervals of Variables in PLSR. Journal of Chemometrics, 24, 45–56.

Vitor, L., Carla C. P., José C. M. 2000. Evolutionary Programming for Variable Selection in PLSR: Predicting Qualities from a Crude Distillation Unit, Controlo'2000: 4th Portuguese Conference on Automatic Control.

Wold, H., In David, F., 1966. Research papers in statistics. Wiley, New York, 411-444.

Wold, S., Martens, M., Wold, H., 1983. The Multivariate Calibration Problem in Chemistry Solved By The PLS Method. In Ruhe, and Kågstrom, B. (Eds) Matrix Pencils, Springer-Verlag, Hieldelberg, Germany. 286-293.

Wold, S., Ruhe, A., Wold, H., Dunn III, W. J., 1984. The Collinearity Problem in Linear Regression: The Partial Least Squares Approach to Generalized Inverses. Siam J. Sci. Stat. Comput, 5, 735-743.

Wold, S., Johansson, E., Cocchi, M., 1993. 3D QSAR in Drug Design; Theory, Methods and Applications. ESCOM, Leiden, Holland, 523-550.

Wold, S., 1994. PLS For Multivariate Linear Modelling, QSAR: Chemometric Metods in Molecular Design. Methods and Principles in Medicinal Chemistry. (Ed. H. Van de Waterbeemd), Weinheim, Germany: Verlag-Chemie.

Wold, S., Kettaneh, N., Tjessem, K., 1996. Hierarchical Multiblock PLS and PC Models for Easier Model Interpretation and as an Alternative to Variable Selection. Journal of Chemometrics, 10, 463-482.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-Regression: A Basic Tool of Chemometrics. Chemometrics and Intelligent Laboratory Systems, 58, 109-130.

GENETİK ALGORİTMA TABANLI KISMİ EN KÜÇÜK KARELER REGRESYONU İÇİN DEĞİŞKEN SEÇİMİ

ÖZET

Kısmi En Küçük Kareler Regresyonu (KEKKR), bilimsel araştırmaların birçok alanında çoklu doğrusal bağlantı probleminin üstesinden gelmede sıradan en küçük karelere bir alternatif oluşturmaktadır. KEKKR yönteminin temelinde regresyon modeli ile iç içe geçmiş bir boyut indirgeme tekniği yer almaktadır. Bu çalışmada, genetik algoritma-kısmi en küçük kareler regresyonu (GAKEKK) incelenmiştir. Bu yöntemde, değişken seçiminde kullanılan KEKK ile güçlü optimizasyon yöntemleri olan GA birleştirilmiştir. İz düşüm için değişken önemi, KEKK ağırlıklarının ağırlıklandırılmış kareler toplamı olarak isimlendirilmekte ve hem X hem de Y' i modellemede bir değişkenin önemini özetlemektedir (Wold ve arkadaşları, 2001). Bu çalışmada, daha küçük hataya sahip modeli belirlemede GAKEKK tahmin modeli, KEKK-NIPALS modeli ve KEKK-VIP yöntemlerinin performans karşılaştırmaları R^2_{adj} *değerleri kullanılarak incelenmiştir.*

Anahtar Kelimeler: Değişken seçimi, Genetik algoritma, İz düşüm için değişken önemi, Kısmi en küçük kareler regresyonu.