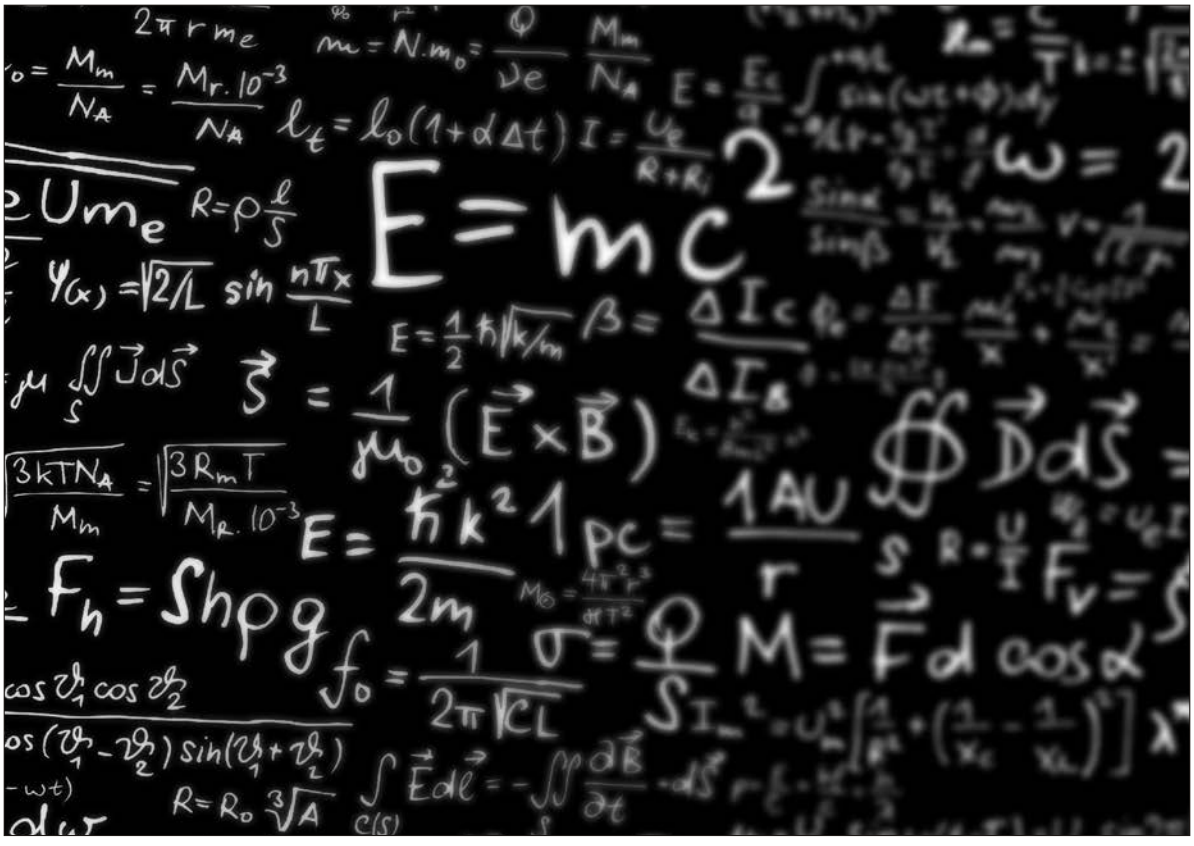

Asian Journal of Mathematics & Statistics

ISSN 1994-5418

science alert

ANSInet
an open access publisher
http://ansinet.com

Asian Journal of Mathematics and Statistics 5 (3): 82-92, 2012
ISSN 1994-5418 / DOI: 10.3923/ajms.2012.82.92

Genetic Algorithm Based Variable Selection for Partial Least Squares Regression Using ICOMP Criterion

[1]Ozlem Gurunlu Alma and [2]Elif Bulut
[1]Department of Statistics, Faculty of Sciences, Mugla University, 48000, Mugla, Turkey
[2]Department of Economics, Ondokuz Mayis University, Samsun, Turkey

Corresponding Author: Ozlem Gurunlu Alma, Department of Statistics, Faculty of Sciences, Mugla University, 48000, Mugla, Turkey

ABSTRACT

Partial least squares regression is a statistical method of modeling relationships between $Y_{N\times M}$ response variable and $X_{N\times K}$ explanatory variables which is particularly well suited for analyzing when explanatory variables are highly correlated. In partial least square part, some model selection criteria are used to obtain the latent variables which are the most relevant variables describing the response variables. In this study, we investigate the performance of Partial Least Squares Regression-the Nonlinear Iterative Partial Least Squares (PLSR-NIPALS), Partial Least Squares Regression-the Variable Importance in the Projection (PLSR-VIP) and the Genetic Algorithms Partial Least Square Regression (GAPLSR) when the fitness function is the Information Complexity Criterion (ICOMP) for model selection. We compared the performance of these methods with real world data and simulation data sets and used the adjusted R square (R^2_{adj}) values to quantify the adequacy of the models.

Key words: Genetic algorithms, ICOMP, partial least square regressions, variable selection, variable importance for projection

INTRODUCTION

The PLSR's goal is to predict or analyze a set of response variables from a set of independent variables or predictors. This prediction is achieved by extracting from the predictors a set of orthogonal factors called latent variables which have the best predictive power. These are constructed through the use of latent variables which maximize the covariance between predictors and explanatory variables. This construction follows an interative procedure to ensure that the latent variables are orthogonal (Lopes *et al*., 2000). In the past, PLS regression was considered to be almost insensitive to noise, therefore, there was a common acceptance that no feature selection was necessary to build a better predictive model (Leardi and Gonzalez, 1998; Lawlor *et al*., 2003; Lawrence *et al*., 2006; Mardikyan and Darcan, 2006; Zamani *et al*., 2011). Today, it has been widely accepted that a feature selection has some advantages. Although, PLS is a well-working method to model highly multidimensional and collinear datasets, the interpretation and understanding of the predictive model and its results are more difficult (Wold *et al*., 1996). Feature selection can also help to build a better predictive PLS model with fewer features (Kubinyi, 1996). The PLS regression combined with the 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; D'Ambra and Sarnacchiaro, 2010). An optimal way to do variable selection

is to try all combinations of variables and select the best ones. This sounds simple, but is, in practice, 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 Minerva, 2010). 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 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 study was to explore the nature of the GAPLSR method and to compare with PLS-NIPALS method and the PLSR-VIP methods through computer simulation experiments and a real data set using the R^2_{adj} values to quantify the adequacy of the models.

PLSR MODEL AND VARIABLE IMPORTANCE FOR PROJECTION

PLSR is a latent variable based multivariate statistical method forms from combination of partial least squares and multiple linear regression. It can be understood from various perspectives a way to compute generalized matrix inverses, a method for system analysis and pattern recognition as well as learning algorithm (Martens and Naes, 1989). The intension of PLSR is to form components that capture most of the information in the X, 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). In PLS regression analysis, many algorithms are used to obtain latent variables. The objective of all linear PLSR algorithm is to project the data down onto a number of latent variables (t_a and u_a) and then to develop a regression model between latent variables. It uses both the variation of X and Y to construct latent variables. 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 to 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 often referred to as the classical algorithm. The development was initiated by Joreskog and Wold (1982) and Wold (1966) and later extended by Lindgren and Rannar (1998), Wold *et al.* (1983) and Wold *et al.* (1984).

In PLS regression and in all algorithms $X_{N\times K}$ represents the data matrix of N observation units on K explanatory variables and $Y_{N\times M}$ the data matrix of N observation units on M response variables. NIPALS algorithm composes of two loops. The inner loop is used to attain t_a and u_a (a = 1,...A) latent variables, where A is the number of the latent variables. Then, a convergence is tested on the change in u. If convergence has been reached, the outer loop is used sequentially to extract p_a, q_a from X and Y matrices. In this algorithm a regression model between latent variables is written as follows:

$$u_a = b_a t_a + e_a \quad a = 1,\ldots, A \tag{1}$$

where, e_a is vector of errors and b_a is an unknown parameter estimated by $\hat{b}_a = (t'_a t_a)^{-1} t'_a u_a$. The latent variables are computed by $t_a = X_a w_a$ and $u_a + Y_a q_a$, where both w_a weight vector for X and q_a loading

vector for Y have unit length and are determined by maximizing the covariance between t_a and u_a. X and Y data matrices are deflated at the end of each iteration as $X_{a+1} = X_a - t_a p'_a$ where $x_1 = x$ and $p_a = X'_a t_a / (t'_a t_a)$ and $Y_{a+1} = Y_a - b_a t_a q'_a$ where $Y_1 = Y$ for to use in the next iteration. Letting $\hat{u}_a = \hat{b}_a t_a$ be prediction of u_a, the matrices X and Y can be decomposed as the following (Li *et al.*, 2002):

$$X = \sum_{a=1}^{A} t_a p'_a + E \text{ and } Y = \sum_{a=1}^{A} \hat{u}_a q'_a + F, \tag{2}$$

where, E and F are the residuals of X and Y after extracting the first "A" pairs of latent variables.

The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors and providing a better understanding of the underlying process that generated the data (Guyon and Elisseeff, 2003). Variable Importance for Projection (VIP) is a variable selection criterion which 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). The VIP value was derived from the partial least squares was considered as a variable selection procedure. It is a statistic of summarizing the contribution that a variable makes to the model (Wold, 1994). 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 are obtained by PLS regression can be used to select the most influential variables (Chong and Jun, 2005). The VIP score can be estimated for jth explanatory variable by:

$$VIP_j = \sqrt{K \times \frac{\sum_a w_{ja}^2 b_a^2 t'_a t_a}{\sum_a b_a^2 t'_a t_a}} \tag{3}$$

where, w_{ja} is a weight of the jth X-variable to the ath latent variable which is obtained by NIPALS algorithm (Chi-Hyuck *et al.*, 2009). Weight values can be interpreted as the contribution of the jth explanatory variable to the ath latent variable. The VIP score equals to 1: Greater than one rule is generally used as a criterion for variable selection (Chong and Jun, 2005).

GENETIC ALGORITHMS BASED VARIABLE SELECTION FOR PLSR USING ICOMP CRITERIA

GA is a search technique used in computing to find true or approximate solutions to optimization and search problems which are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, selection and crossover (Goldberg, 1989). Details on the algorithm used can be found in the literature (Leardi *et al.*, 1992; Leardi, 1996; Xia *et al.*, 2009; Hasheminia and Niaki, 2008; Sinha and Chande, 2010; Noorizadeh and Farmany, 2011). 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). Leardi *et al.* (1992) use a simulated data set to show that a GA can always find the global maximum of a simple problem, in

a time much shorter than the time required by a full search. Lucasius *et al.* (1994) showed that a GA generally performs better than simulated annealing and stepwise regression, on the other hand, Horchner and Kalivas (1995) demonstrated that simulated annealing can give the same results (Leardi, 2001).

The performance of the regression model, which is usually represented as the Root Mean Square Error in Prediction (RMSEP) is optimized by GA procedure. 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 (Arakawa *et al.*, 2011). In this study, the GA is made up by a number of steps. Its main characteristics of GAs are the following:

Step 1: Genetic coding scheme: First, a vector consisting of zeros and ones is made with the size corresponding to the number of variables. It is denoted a chromosome. The randomly defined zeros and ones indicate the variables that should be included. Details on the algorithm used can be found by Leardi *et al.* (1992) and Leardi (1996). Each zero or one is a gene and a PLSR model made 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_1 = 0 indicates no explanatory variable selected and d_1 = 1 indicates an explanatory variable selected for PLSR model, for each i = 1,..., K, (K is number of explanatory variables). In this study, the structure of a chromosome is shown in Fig. 1.

Step 2: Generating the initial population:

- Response to be maximized to explained variance (%)
- Regression method: Partial Least Square Regression
- Population size: 30 chromosomes
- Average number of variables selected in the chromosomes of the starting population: as number of latent variables

Step 3: A fitness function to evaluate model performance: Every candidate solution is evaluated with respect to a fitness function. The fitness function of the GA is selected as ICOMP criterion. Bozdogan (2000) introduced ICOMP (IFIM) under the multivariate normal assumption for the multivariate regression model is defined as:

$$\begin{aligned}
&\text{ICOMP (IFIM)}_{\text{Multivariate}} = nq\log(2\pi) + n\log\left|\hat{\Sigma}\right| + nq + 2C_1(\hat{F}^{-1}(\hat{\theta})),\\
&C_1(\hat{F}^{-1}(\hat{\theta})) = \frac{q(q+p)}{2}\log\left[(tr(\hat{\Sigma})\,tr(X'X)^{-1} + \frac{1}{2n}\left[tr(\hat{\Sigma}^2) + tr^2(\hat{\Sigma}) + 2\sum_{i=1}^{q}\hat{\sigma}_{ii}^2\right])/(q(q+p))\right]\\
&-\frac{1}{2}(p+q+1)\log\left|\hat{\Sigma}\right| - \frac{q}{2}\log\left|(X'X)^{-1}\right| - \frac{q}{2}\log(2)
\end{aligned} \tag{4}$$

and Ta ,Y, M, V (a,Ta) and A was used instead of X, Y, p,∑, q:

$$V(a, T_a) = \min_{T_a}\frac{1}{MN}\sum_{k=1}^{M}\sum_{i=1}^{N}(Y_{ik} - T_{ia}b_{ak})^2 \tag{5}$$

where, N is the number of objects in the evaluation set.

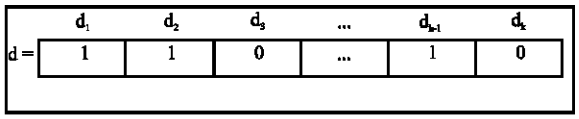

Fig. 1: The structure of a chromosome

Step 4: To select fitter models: It was computed ICOMP score for each subset model in the population. It was used roulette wheel selection method.

Step 5: Producing offspring models: 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 two points), 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. Cross-over method is uniform probability and mutation probability: Is selected as 1% in GAPLSR. General form of algorithm is as shown below:

- Generating simulation data
- Begin GA
- Generating chromosomes and population
- Applying dimension reduction PLSR algorithm on data and
- Repeat until ICOMP has minimum value. (Calculating PLSR model using fitness value as ICOMP)
- Calculate R^2_{adj} values for GAPLSR models
- End GA
- Calculate R^2_{adj} values for PLSR-NIPALS models
- Calculate R^2_{adj} values for PLSR-VIP models

DESIGN OF SIMULATION STUDY AND REAL WORLD DATA

The frame work for the simulation models was based on the study of Li *et al.* (2002), Naes and Martens (1985). It was extended in this study 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). The X and Y block data, with sample size N, were generated as:

$$X = \sum_{i=1}^{A^*} \tau_i \xi'_i + \tilde{E}$$

and:

$$Y = \sum_{i=1}^{A^*} z_i \eta'_{A^*_1} + \psi = \sum_{i=1}^{A^*} \tau_i \eta'_{A^*_1} + \tilde{F}_{A^*} \quad (6)$$

where, $\tilde{E}$ and r_i were generated from mutually independent normal variables. Generating of X and Y data matrices are just explained for 5×3. Ψ was generated from a multivariate normal distribution and generated as (Li *et al.*, 2002), $\tilde{F}$ is a noise matrix and Z was constructed as $z_i = r_i + f_i$, f_i, were generated as independent normal variables with zero means and variances (0.5, 0.25 and 0.1). $\{\xi_i\}$ and $\{\eta_{A\times i}\}$ are normalized orthogonal vector series and r_i are mutually independent random variables with zero means and variances (15, 7.5 and 3). To carry out simulations run, it is proceeded on different simulation. The dimensions of explanatory variables is extended as N×5, N×5, N×8, N×10 and N×2. The dimension of response variables matrix, Y, is chosen N×3, N×4 as and sample sizes are selected as N = 50, 100, 250, 500. For each of the combinations, 100 data sets are generated taking into account the dimension of partial least squares regression models 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 X and Y blocks, averaged the 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 variance in the X and Y data sets, respectively. This verifies the theoretical value of the number of latent variables $A^* = 3$. For more information look the reference (Li *et al.*, 2002). Then these data sets are applied to GAPLSR, PLSR-NIPALS and PLSR-VIP methods.

Real world data example: GAPLSR, PLSR-NIPALS and PLSR-VIP methods have been tested considering a real world dataset: The Body Fat Measurement. In this example we determine the best subset predictors of y = Percent body fat from Siria (1956) equation, using k = 13 predictors, x_1 = Age (years), x_2 = Weight (lbs), x_3 = Height (inches), x_4 = Neck circumference (cm), x_5 = Chest circumference (cm), x_6 = Abdomen 2 circumference (cm), x_7 = Hip circumference (cm), x_8 = Thigh circumference (cm), x_9 = Knee circumference (cm), x_{10} = Ankle circumference (cm), x_{11} = Biceps (extended) circumference (cm), x_{12} = Forearm circumference (cm), x_{13} = Wrist circumference (cm) using the GA with ICOMP as the fitness function. The data contains the estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for n = 252 men. This is a good example to illustrate the versatility and utility of our approach using multiple regression analysis with GA. A variety of popular health books suggest that the readers assess their health, at least in part, by estimating their percentage of body fat. In Bailey (1994), for instance, the reader can estimate body fat from tables using their age and various skin-fold measurements obtained by using a caliper. Other texts give predictive equations for body fat using body circumference measurements (e.g., abdominal circumference) and/or skin-fold measurements. See, for instance, Behnke and Wilmore (1974), Wilmore (1976), or Katch and McArdle (1977). Percentage of body fat for an individual can be estimated once body density has been determined. Siria (1956) assume that the body consists of two components-lean body tissue and fat tissue. Letting:

D = Body Density (g cm^{-3})
D = 1/[(A/a)+(B/b)]
A = Proportion of lean body tissue
B = Proportion of fat tissue (A+B = 1)
B = (1/D)*[ab/(a-b)]-[b/(a-b)]
a = Density of lean body tissue (g cm^{-3})
b = Density of fat tissue (g cm^{-3})

Table 1: Parameters of the GA run for the body fat data

Parameter	Value
No. of runs	100
No. of generations	30
Fitness Value	ICOMP
Population size	30
Probability of crossover	0.8
Elitism	Yes
Probability of mutation	0.01

Table 2: Summary of fit of best subset model chosen by GAPLSR, PLSR-NIPALS and PLSR-VIP methods

Method	No. of latent variables	Selected variables	R^2 Adj.
GAPLSR	7	1- -3 4- -7 8 –10- - 13	0.8095
PLSR-NIPALS	7	-	0.7237
PLSR-VIP	6	- - 3 4 – 6 7 8 9 - - - -	0.6612

Using the estimates a = 1.10 g cm^{-3} and b = 0.90 g cm^{-3} (Katch and McArdle, 1977) or Wilmore (1976), we come up with Siri's equation:

$$\text{Percentage of body fat (i.e., } 100\times B) = 495/D-450$$

Volume and hence body density, can be accurately measured a variety of ways. The technique of underwater weighing computes body volume as the difference between body weight measured in air and weight measured during water submersion. In other words, body volume is equal to the loss of weight in water with the appropriate temperature correction for the water's density (Katch and McArdle, 1977). Using this technique,

$$\text{Body density} = WA/[(WA-WW)/c.f.-LV]$$

where, WA = Weight in air (kg), WW = Weight in water (kg) c.f. = Water correction factor (= 1 at 39.2 deg F as one-gram of water occupies exactly one cm^{-3} at this temperature, = .997 at 76-78 deg F), LV = Residual Lung Volume (liters) (Katch and McArdle, 1977). Other methods of determining body volume are given in Behnke and Wilmore (1974). Parameters of the GA run for the body fat data as seen in Table 1.

Summary of fit of best subset model chosen by GAPLSR, PLSR-NIPALS and PLSR-VIP methods among all possible PLSR models for the body fat data as following in Table 2.

The best model chosen by the GAPLSR with lowest ICOMP (IFIM) value = 1718.091 with the subset {x_1 = Age (years), x_3 = Height (inches), x_4 = Neck circumference (cm), x_7 = Hip circumference (cm), x_8 = Thigh circumference (cm), x_{10} = Ankle circumference (cm), x_{13} = Wrist circumference (cm)} among all possible models for the body fat data. Our GA application to the problem of optimal statistical model selection on the body fat data indicates that the GAPLSR can indeed find the best model having the biggest R^2_{adj} value than other methods.

RESULTS

In this study, it has done a simulation study to gain a better understanding of GAPLSR, PLSR-NIPALS and PLSR-VIP methods performances for PLSR model selection, it was run a designed experimental simulation study for see which one has established a model with less error and have the biggest R^2_{adj} values. Table 3 shows the empirical results.

Table 3: The empirical results of GAPLSR, PLSR-NIPALS and PLSR-VIP methods

	5×3		8×4		10×4		12×4	
Method	R^2_{adj}	No. of LVs	R^2_{adj}	No. of LVs	R^2_{adj}	No. of LVs	R^2_{adj}	No. of LVs
GA								
R^2_{adj1}								
50	0.9842	3	0.9714	4	0.9732	5	0.9680	5
100	0.9569	3	0.9713	4	0.9762	5	0.9707	5
250	0.9696	4	0.9714	4	0.9723	5	0.9805	5
500	0.9568	4	0.9690	4	0.9748	6	0.9758	6
R^2_{adj2}								
50	0.9739	3	0.9770	4	0.9546	5	0.9374	5
100	0.9840	3	0.9742	4	0.9730	5	0.9691	5
250	0.9608	4	0.9714	4	0.9711	5	0.9739	5
500	0.9610	4	0.9713	4	0.9751	6	0.9795	6
R^2_{adj3}								
50	0.9640	3	0.9508	4	0.9779	5	0.9812	5
100	0.9702	3	0.9709	4	0.9792	5	0.9797	5
250	0.9598	4	0.9700	4	0.9752	5	0.9749	5
500	0.9718	4	0.9697	4	0.9776	6	0.9747	6
R^2_{adj4}								
50			0.9712	4	0.9615	5	0.9711	5
100			0.9700	4	0.9725	5	0.9756	5
250			0.9773	4	0.9727	5	0.9742	5
500			0.9753	4	0.9743	6	0.9792	6
NIPALS-PLSR								
R^2_{adj1}								
50	0.9826	3	0.9687	4	0.9708	4	0.9608	4
100	0.9561	3	0.9703	4	0.9737	4	0.9677	4
250	0.9695	3	0.9709	4	0.9713	4	0.9709	4
500	0.9710	3	0.9699	4	0.9776	4	0.9705	4
R^2_{adj2}								
50	0.9721	3	0.9734	4	0.9499	4	0.9259	4
100	0.9854	3	0.9721	4	0.9705	4	0.9657	4
250	0.9606	3	0.9708	4	0.9705	4	0.9692	4
500	0.9626	3	0.9716	4	0.9754	4	0.9732	4
R^2_{adj3}								
50	0.9631	3	0.9446	4	0.9743	4	0.9805	4
100	0.9708	3	0.9709	4	0.9769	4	0.9785	4
250	0.9622	3	0.9694	4	0.9750	4	0.9780	4
500	0.9728	3	0.9699	4	0.9741	4	0.9698	4
R^2_{adj4}								
50			0.9686	4	0.9579	4	0.9669	4
100			0.9687	4	0.9704	4	0.9723	4
250			0.9771	4	0.9723	4	0.9751	4
500			0.9754	4	0.9761	4	0.9725	4
VIP								
R^2_{adj1}								
50	0.9835	3	0.6313	4	0.6722	4	0.7357	4
100	0.9574	3	0.5422	4	0.7045	4	0.7515	4
250	0.9696	3	0.6911	4	0.7331	4	0.7709	4
500	0.9705	3	0.6112	4	0.8797	4	0.9687	4

Table 3: Continued

	5×3		8×4		10×4		12×4	
Method	R^2_{adj}	No. of LVs	R^2_{adj}	No. of LVs	R^2_{adj}	No. of LVs	R^2_{adj}	No. of LVs
R^2_{adj2}								
50	0.8093	3	0.6589	4	0.9008	4	0.6290	4
100	0.8013	3	0.7325	4	0.6789	4	0.7992	4
250	0.7551	3	0.9111	4	0.8210	4	0.8102	4
500	0.7667	3	0.9191	4	0.7014	4	0.9726	4
R^2_{adj3}								
50	0.6312	3	0.8402	4	0.9887	4	0.8086	4
100	0.6321	3	0.8294	4	0.9175	4	0.9758	4
250	0.5241	3	0.8224	4	0.9697	4	0.9026	4
500	0.6987	3	0.9466	4	0.8833	4	0.9699	4
R^2_{adj4}								
50			0.8630	4	0.9580	4	0.9570	
100			0.9187	4	0.9368	4	0.9588	4
250			0.9393	4	0.9663	4	0.9459	4
500			0.9483	4	0.9455	4	0.9726	4

All of these methods have different study structures. GAPLSR finds variables according to minimum value of ICOMP criterion, and then make modeling on these variables. PLSR-NIPALS works with latent variables. It finds regression coefficients on latent variables and makes regression modeling on variables. PLSR-VIP also works with latent variables. It finds VIP values by the help of PLS and then uses explanatory variables which have VIP values bigger than 1 and make regression modeling on these explanatory variables to calculate R^2_{adj}, PLSR-NIPALS and PLSR-VIP methods work with same number of latent variables. These numbers equal the A^*. The results show that R^2_{adj} values for models with GAPLSR have the biggest values for each of the design matrices and for each N. This study shows that variable selection with GAPLSR method gives better results than selection with PLSR-VIP. 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 such as RMSEP, R-squares, 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. So, if we use a strategy which combines these two methods for selecting relevant predictors, better variable selection performance could be achieved. Actually, Wold *et al.* (1993) recommend a combination of PLS-VIP and PLS-Beta for variable selection, which states that both should be small for a variable to be excluded (Chong and Jun, 2005). As a result, GAPLSR set out PLSR model which have a little more latent variables than PLSR-NIPALS and PLSR-VIP methods. However, finding of GAPLSR models have the biggest R^2_{adj} values.

REFERENCES

Andersen, C.M. and R. Bro, 2010. Variable selection in regression: A tutorial. J. Chemometrics, 24: 728-737.

Arakawa, M., Y. Yamashita and K. Funatsu, 2011. Genetic algorithm-based wavelength selection method for spectral calibration. J. Chemometrics, 25: 10-19.

Bailey, C., 1994. Smart Exercise: Burning Fat, Getting Fit. Houghton-Mifflin Co., Boston, pp: 179-186.

Behnke, A.R. and J.H. Wilmore, 1974. Evaluation and Regulation of Body Build and Composition. Prentice-Hall, Englewood Cliffs, NJ., Pages: 236.

Bozdogan, H., 2000. Akaike`s information criterion and recent developments in information complexity. J. Math. Psychol., 44: 62-91.

Chi-Hyuck, J., S.H. Lee, H.S. Park and J.H. Lee, 2009. Use of partial least squares regression for variable selection and quality prediction. Proceedings of the International Conference on Computers and Industrial Engineering, July 6-9, 2009, Troyes, pp: 1302-1307.

Chong, I.G. and C.H. Jun, 2005. Performance of some variable selection methods when multicollinearity is present. Chemometrics Intell. Lab. Syst., 78: 103-112.

D'Ambra, A. and P. Sarnacchiaro, 2010. Some data reduction methods to analyze the dependence with highly collinear variables: A simulation study. Asian J. Math. Stat., 3: 69-81.

Garthwaite, P.H., 1994. An interpretation of partial least squares. J. Am. Stat. Assoc., 89: 122-127.

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

Guyon, I. and A. Ehsseeff, 2003. An introduction to variable and feature selection. J. Mach. Learn. Res., 3: 1157-1182.

Hasheminia, H. and S.T.A. Niaki, 2008. A hybrid method of neural networks and genetic algorithm in econometric modeling and analysis. J. Applied Sci., 8: 2825-2833.

Horchner, U. and J.H. Kalivas, 1995. Further investigation on a comparative study of simulated annealing and genetic algorithm for wavelength selection. Anal. Chim. Acta, 311: 1-13.

Joreskog, K.G. and H. Wold, 1982. System under Indirect Observation. Vol. 1 and 2, North-Holland, Amsterdam, The Netherlands.

Katch, F. and W. McArdle, 1977. Nutrition, Weight Control and Exercise. Houghton-Mifflin Co., Boston.

Kubinyi, H., 1996. Evolutionary variable selection in regression and PLS analyses. J. Chemometr., 10: 119-133.

Lawlor, J.B., E.M. Sheehan, C.M. Delahunty, J.P. Kerry and P.A. Morrissey, 2003. Sensory characteristics and consumer preference for cooked chicken breasts from organic, corn-fed, free-range and conventionally reared animals. Int. J. Poult. Sci., 2: 409-416.

Lawrence, K.C., D.P. Smith, W.R. Windham, G.W. Heitschmidt and B. Park, 2006. Egg embryo development detection with hyperspectral imaging. Int. J. Poult. Sci., 5: 964-969.

Leardi, R. and A.L. Gonzalez, 1998. Genetic algorithms applied to feature selection in PLS regression: How and when to use them. Chemolab, 41: 195-207.

Leardi, R., 1996. Genetic Algorithms in Feature Selection. In: Genetic Algorithms in Molecular Modeling, Devillers, J. (Ed.). Academic Press, London.

Leardi, R., 2001. Genetic algorithms in chemometrics and chemistry: A review. J. Chemometrics, 15: 559-569.

Leardi, R., R. Boggia and M. Terrile, 1992. Genetic algorithms as a strategy for feature selection. J. Chemometrics, 6: 267-281.

Li, B., J. Morris and E.B. Martin, 2002. Model selection for partial least squares regression. Chemometrics Intell. Lab. Syst., 64: 79-89.

Lindgren, F. and S. Rannar, 1998. Alternative partial least-squares (PLS) algorithms. Perspect. Drug Discovery Des., 12-14: 105-113.

Lopes, V.V., C.C. Pinheiro and J.C. Menezes, 2000. Evolutionary programming for variable selection in PLSR: Predicting qualities from a crude distillation unit. Proceedings of the 4th Portuguese Conference on Automatic Control, October 4-6, 2000, Guimaraes, Portugal, pp: 400-406.

Lucasius, C.B., M.L.M. Beckers and G. Kateman, 1994. Genetic algorithms in wavelength selection: A comparative-study. Anal. Chim. Acta, 286: 135-153.

Mardikyan, S. and O.N. Darcan, 2006. A software tool for regression analysis and its assumptions. Inform. Technol. J., 5: 884-891.

Martens, M. and T. Naes, 1989. Multivariate Calibration. J. Wiley and Sons, Ltd., Chichester.

Naes, T. and H. Martens, 1985. Comparison of prediction methods for multicollinear data. Commun. Stat. Simulat. Comput., 14: 545-576.

Noorizadeh, H. and A. Farmany, 2011. Investigation of capacity behaviors by linear and nonlinear models chemometrics. Trends Applied Sci. Res., 6: 1324-1334.

Paterlini, S. and T. Minerva, 2010. Regression model selection using genetic algorithms. Proceedings of the 11th WSEAS International Conference on Nural Networks and Evolutionary Computing and Fuzzy Systems, June 13-15, 2010, G. Enescu University, Iasi, Romania, pp: 19-27.

Sinha, M. and S.V. Chande, 2010. Query optimization using genetic algorithms. Res. J. Inform. Technol., 2: 139-144.

Siria, W.E., 1956. Gross Composition of the Body. In: Advance in Biological and Medical Physics, Lawrence J.H. and C.A. Tobias (Eds.). Academic Press, New York.

Wilmore, J., 1976. Athletic Training and Physical Fitness: Physiological Principlesof the Conditioning Process. Allyn and Bacon, Inc., Boston.

Wold, H., 1966. Non-Linear Estimation by Iterative Least Squares Procedures. In: Research Papers in Statistics, David, F. (Ed.). New York, pp: 411-444.

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

Wold, S., A. Ruhe, H. Wold and W.J. Dunn, 1984. The collinearity problem in linear regression. The Partial Least Squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput., 5: 735-743.

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

Wold, S., H. Marten and H. Wold, 1983. The Multivariate Calibration Problem in Chemistry Solved by the PLS Method. In: Matrix Pencils, Ruhe, A. and B. Kagstrom (Eds.). Springer-Verlag, Heidelberg, pp: 286-293.

Wold, S., M. Sjorn and L. Erikson, 2001. PLS-regression: A basic tool of chemometrics. Chemometrics Intell. Lab. Syst., 58: 109-130.

Wold, S., N. Kettaneh and K. Tjessem, 1996. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. J. Chemometrics, 10: 463-482.

Xia, Z., X. Sun, J. Qin and C. Niu, 2009. Feature selection for image steganalysis using hybrid genetic algorithm. Inform. Technol. J., 8: 811-820.

Zamani, Z., M. Arjmand, M. Tafazzoli, A. Gholizadeh and F. Pourfallah *et al.*, 2011. Early detection of ýmmunization: A study based on an animal model using 1H nuclear magnetic resonance spectroscopy. Pak. J. Biol. Sci., 14: 195-203.