On Variable Selection for Data-driven Soft Sensor Development with Application to Industrial Processes

by

Zi Xiu Wang

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
December 8, 2012

Keywords: Variable Selection, Soft Sensor, Data-Driven Models, Process Industry, Model Sampling

Approved by

Jin Wang, Chair, Associate Professor of Chemical Engineering
Qinghua He, Associate Professor of Chemical Engineering, Tuskegee University
Mario R. Eden, Professor of Chemical Engineering

Abstract

In recent years, rapid developments in technology facilitated the collection of vast amount of data from different industrial processes. The data has been utilized in many different areas, such as data-driven soft sensor development and process monitoring, to control and optimize the process. The performance of these data-driven schemes can be greatly improved by selecting only the vital variables that strongly affect the primary variables, rather than all the available process variables. Consequently, variable selection has been one of the most important practical concerns in data-driven approaches. By identifying the irrelevant and redundant variables, variable selection can improve the prediction performance, reduce the computational load and model complexity, obtain better insight into the nature of the process, and lower the cost of measurements [1], [2].

A comprehensive evaluation of different variable selection methods for soft sensor development will be presented in this work. Among all the variable selection methods, seven algorithms are investigated. They are stepwise regression, PLS-BETA, PLS-VIP, UVE-PLS, PLS-SA, CARS-PLS and GA as discussed below. Stepwise regression methods are often used for variable selection in linear regression [3]. The procedure is carried out in such a way that individual predictor/secondary variable is sequentially introduced into the model to observe its relation to the primary variables. Partial Least Squares (PLS) regression is a model parameter based algorithm. Both the regression coefficients estimated by PLS (PLS-BETA) and variable importance in projection (PLS-

VIP) are discussed [4]. Another model parameter based method, called Uninformative Variable Elimination by PLS (UVE-PLS), is also related to regression coefficients. However, instead of looking at the regression coefficients only, the reliability of the coefficients is explored [5]. Variable selection algorithms based on sensitivity analysis, PLS-SA, are also studied. In these approaches, the importance of variables is defined by their sensitivity, which is defined as the change in primary variables by varying the secondary variable in its allowable range [6]. Furthermore, properties of genetic algorithms (GA), which have been recently proposed for variable selection applications [7], are also investigated.

The algorithms of these variable selection methods and their characteristics will be presented. In addition, the strength and limitations when applied for soft sensor development are studied. The soft sensor prediction performance of models developed by these variable selection methods are compared using PLS.

A simple simulation case is used to investigate the properties of the selected variable selection methods. The dataset is generated to mimic the typical characteristics of process data, such as the magnitude of correlations between variables and the magnitude of signal to noise ratio, etc. [4]. In addition, the algorithms are applied to an industrial soft sensor case study. In both cases, independent test sets are used to provide fair comparison and analysis of different algorithms. The final performances are compared to demonstrate the advantages and disadvantages of the different methods in order to provide useful insights to practitioners in the field.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Jin Wang, for her guidance and constant supervision. Her guidance has made this a rewarding and meaningful journey. Her dedication and hard work have set a great example for me to become a better researcher.

I would like to express my deepest appreciation and gratitude to my other research advisor, Dr. Qinghua He. His experience, support, and encouragement have truly made a difference in my journey. Through all the meetings I have had with him, I have always left the office more encouraged and motivated and I am very thankful for that. I would also like to thank Dr. Mario R. Eden for agreeing to be on my committee and for providing me the opportunity to join the Chemical Engineering Department.

Certainly, this journey was not easy, through struggles and success my fellow group members and I shared disappointments and victories together. Thus, I would like to extend my gratitude to my group members: Hector Galicia Escobar, Meng Liang, Min Hea Kim, and Andrew Damiani. Also special thanks to Hector Galicia for his endless guidance, assistance and support. He has been very instrumental in successful completion of this thesis.

Friends have always been important to me, and thankfully during my journey great friendships were always there. I would like to thank my friends: Jimmy Tran, Ari-

anna Tieppo, Pengfei Zhao, Achintya Sujan and Chuan Cai Zou. They have brought laughter, happiness, and comfort to not only this journey but my life.

Last but not least, my love and gratitude go to my parents, Hai Bin Wang and Yan Yun Yu, for raising me to always striving for excellence, to value my education and for putting my success and happiness before their own. I am also grateful for their unconditional love and support wherever I was. I am thankful to my brother, Yu Dong (Jeffrey) Wang, for his support and understanding. Their love and encouragement has always been a constant source of comfort and support to me when times were difficult.

Table of Contents

List of Tables

List of Figures

List of Nomenclature

Symbols	**Descriptions**
A	PLS components
ARS	Adaptive reweighted sampling
BETA	Regression Coefficients
CARS	Competitive Adaptive Reweighted Sampling
EDF	Exponential decreasing function
G	Geometric mean of sensitivity and specificity
GA	Genetic Algorithm
GAVDS	Genetic Algorithm-Based Process Variables and Dynamics Selection
J	Number of variables
K	Number of batches
k	Reciprocal of signal to noise ratio
KRR	Kernel Ridge Regression
MAPE	Mean absolute percentage error
MLR	Multiple Linear Regression
MC	Monte Carlo
MS_E	Mean square error
N	Total number of simulation runs
nb	Normalized regression coefficient

PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares
PLSLDA	Partial Least Squares Linear Discriminant Analysis
RMSEP	Root mean square error of prediction
RV	Random variable added in PLS-SA
SPA	Statistics Pattern Analysis or Successive Projection Algorithm
SR	Stepwise Regression
SS_E	Sum square error
T	Score matrix
USE	Uninformative Sample Elimination
UVE	Uninformative Variable Elimination
v	Cutoff value of VIP score
$var(\cdot)$	Sample variance
VIP	Variable Importance in Projection
W	Weighting matrix
X	Independent variable matrix
XR	Extended matrix with the experimental and random variables in UVE
Y	Dependent variable matrix
α	$1-\alpha$ is the confidence level in statistic testing
Γ	Variance-covariance matrix
ϵ	Normal distributed random noise

ρ	Magnitude of correlation between predictors
Ω	Variable subset
σ	Standard deviation of error, ϵ

Chapter 1. Introduction

In recent years, rapid developments in technology facilitated the collection of vast amount of data from different industrial processes. The data has been utilized in many different areas, such as data-driven soft sensor development and process monitoring, to control and optimize the process. The performance of these data-driven schemes can be greatly improved by selecting only the vital variables that strongly affect the primary variables, rather than all the available process variables. Consequently, variable selection has been one of the most important practical concerns in data-driven approaches. By identifying the irrelevant and redundant variables, variable selection can improve the prediction performance, reduce the computational load and model complexity, obtain better insight into the nature of the process, and lower the cost of measurements [1], [2].

A comprehensive evaluation of different variable selection methods for soft sensor development will be presented in this work. Among all the variable selection methods, seven algorithms are investigated. They are stepwise regression, PLS-BETA, PLS-VIP, UVE-PLS, PLS-SA, CARS-PLS and GA-PLS as discussed below. Stepwise regression methods are often used for variable selection in linear regression [3]. The procedure is carried out in such a way that individual predictor/secondary variable is sequentially introduced into the model to observe its relation to the primary variables. Partial Least Squares (PLS) regression is a model parameter based algorithm. Both the regression coefficients estimated by PLS (PLS-BETA) and variable importance in projection (PLS-VIP) are discussed [4]. Another model parameter based method, called Uninformative

Variable Elimination by PLS (UVE-PLS), is also related to regression coefficients. However, instead of looking at the regression coefficients only, the reliability of the coefficients is explored [5]. Variable selection algorithms based on sensitivity analysis, PLS-SA, are also studied. In these approaches, the importance of variables is defined by their sensitivity, which is defined as the change in primary variables by varying the secondary variable in its allowable range [6]. Furthermore, properties of genetic algorithms (GA), which have been recently proposed for variable selection applications [7], are also investigated.

Stepwise regression has been applied to the selection of predictors for both classification and multivariate calibrations [8], especially in near-infrared (NIR) spectral. Gauchi and Chagnon proposed a stepwise variable selection method based on maximum Q^2 and applied to manufacturing processes in oil, chemical and food industries [9].

Broadhurst et al. applied genetic algorithm to pyrolysis mass spectrometric data and showed that GA is able to determine the optimal subset of variables to provide better or equal prediction performance [10]. Arcos et al. successfully applied GA to a wavelength selection for PLS calibration of mixtures of indomethacin and acemethacin, in spite of the fact that the two compounds have almost identical spectra [11]. A modified genetic algorithm-based wavelength selection method has been proposed by Hiromasa Kaneko and Kimito Funatsu to select process variables and dynamic simultaneously [12]. This method is named as genetic algorithm-based process variables and dynamics selection method, GAVDS. The result of GAVDS, based on its application to a dynamic process of distillation column in Mitsubishi Chemical Corporation, shows its robustness to the presence of nonlinearity and multicollinearity in process data. GA has also been well

recognized in molecular modeling. Jones et al. have shown three application of GA in chemical structure handling and molecular recognition [13].

A modified uninformative variable elimination method based on the principle of Monte Carlo (MC) was applied in quantitative analysis of NIR spectra by Cai et al. [14]. UVE-MC is proven to be capable of selecting important wavelength and making the prediction more robust and accurate in quantitative analysis. Some researchers also suggested to combine UVE with wavelet transform to further simplify the model and to reduce computation time [14], [15]. In the work of Koshoubu et al., they have extended UVE to eliminate uninformative samples (USE) that do not contribute much in the calibration model [16], [17]. They proposed an algorithm where the uninformative wavelengths/variables are eliminated first by UVE-PLS, and then the uninformative samples, which are determined by their standard deviation of prediction error calculated from leave-one-out cross validation, are eliminated from the calibration. Another new method which combined UVE with successive projection algorithm (SPA) has been proposed by [18]. UVE is implemented to remove uninformative variables before application of SPA to improve the efficiency of variable selection by SPA.

Sensitivity analysis has become more popular in selection of optimal variable subsets in recent years. Zamprogna et al. has introduced a novel methodology based on principal component analysis (PCA) to select the most suitable soft sensor inputs [19]. Instead of using the secondary variables directly, the instantaneous sensitivity of each secondary variable to the primary variables are estimated and utilized as the regressor inputs. Li and Shao have proposed a novel method using sensitivity analysis to select the

optimal secondary variables to be used as inputs to kernel ridge regression (KRR) to implement online soft sensing of distillation column compositions [20].

Competitive adaptive reweighted sampling (CARS) method has been proposed by Li et al. [21]. CARS is model independent. In other words, CARS can be combined with any regression or classification models. In [22], [23], CARS has been applied in combination with partial least squares linear discriminant analysis (PLSLDA) to effectively classify two classes of samples in colorectal cancer data.

Variable importance in the projection (VIP) and regression coefficients (BETA) have been broadly adapted as a criterion in partial least squares modeling paradigm for variable selection. Both PLS-VIP and PLS-BETA are model based variable selection methods. Mehmood et al. presented an algorithm that balances the parsimony and predictive ability of model using variables selection based on PLS-VIP [24]. It is shown that the proposed method increases the understandability and consistency of the model and reduces the classification error. Lindgren et al. also implemented PLS-VIP on a benchmark data for variable selection, Selwood dataset [25]. In their study, PLS-VIP is combined with permutation test to extensively investigate the technique. A bootstrap-PLS-VIP has been implemented as a wavelength interval selection method in spectral imaging applications by Gosselin et al. [26]. Their result demonstrates its ability to identify relevant spectral intervals and its simplicity and relatively low computational cost. PLS-VIP and PLS-BETA have also been seen in food science. Andersen and Bro applied PLS-VIP and PLS-BETA to NIR spectral of beer sample and obtained useful insight of the process [27]. A variable selection algorithm based on the standardized regression coefficients are pro-

posed in [28]. The developed models are optimized by the leave-one-out Q^2 values and validated by an external testing set.

The algorithms of these variable selection methods and their characteristics will be presented. In addition, the strength and limitations when applied for soft sensor development are studied. The soft sensor prediction performance of models developed by these variable selection methods are compared using PLS.

A simple simulation case is used to investigate the properties of the selected variable selection methods. The dataset is generated to mimic the typical characteristics of process data, such as the magnitude of correlations between variables and the magnitude of signal to noise ratio, etc. [4]. In addition, the algorithms are applied to an industrial soft sensor case study. In both cases, independent test sets are used to provide fair comparison and analysis of different algorithms. The final performances are compared to demonstrate the advantages and disadvantages of the different methods in order to provide useful insights to practitioners in the field.

This work is structured as follows. In Chapter 2, a brief review of the multivariate statistical techniques is presented, which will be required for further discussion on variables selection methods. Chapter 3 provides detail descriptions of algorithms of different variable selection methods covered in this work: Stepwise Regression (SR), Genetic Algorithm with Partial Least Squares (GA-PLS), Uninformative Variables Elimination by Partial Least Squares (UVE-PLS), Partial Least Squares with Sensitivity Analysis (PLS-SA), Competitive Adaptive Reweighted Sampling with Partial Least Squares (CARS-PLS), Partial Least Squares with Variable Importance in Projection (PLS-VIP), and Partial Least Squares with regression coefficients (PLS-BETA). In Chapter 4, application of

all seven variable selection methods on simulated case study and industrial case study will be investigated. The simulation case is generated to mimic the typical characteristics of industrial data by considering four factors: proportion of relevant predictors, magnitude of correlation between predictors, structure of regression coefficients, and magnitude of signal to noise ratio. A detailed description of data generation will be provided. The industrial case study is focused on the process data of polyester resin production plant. A brief specification of the plant will be included, followed by discussion of characteristics of batch process. The results and comparison of variable selections on both simulated and industrial case studies will be investigated. Chapter 5 will conclude this work with major discussion and contributions. Furthermore, suggestions on future works will be provided.

Chapter 2. Soft Sensor Development

Soft sensors have been developed and implemented decades ago, where predictive models have been built based on large amount of data being measured stored in process industries [29], [30]. Soft sensors can be classified into two categories: model-driven and data-driven. The model-driven soft sensors are based on the first principle models that describe the physical and chemical characteristics of the process. Data-driven soft sensors are based on the data measured and collected within the plants[29–32]. The most popular soft sensor techniques include principal component analysis (PCA) [33] and partial least squares (PLS) [34], artificial neural networks [35], neuro-fuzzy systems [36] and support vector machines [37]. In our work, only the linear models are considered.

2.1 Multiple Linear Regression

The goal of multiple linear regression (MLR) is to establish a linear relationship between the secondary variables and primary variables in the form of Equation (2.1), where x_j is the secondary variable, y is the primary variable, β_j is the sensitivity, and ϵ is the residuals.

$$y = \sum_{j=1}^{p} \beta_j x_j + \epsilon \tag{2.1}$$

The above linear relationship can also be written in matrix form as:

$$Y = XB + E \tag{2.2}$$

2.2 Principal Component Analysis

Principal component analysis (PCA) is linear technique that transforms the original data matrix X into a smaller set of uncorrelated variables T that would capture most of the information in the original space. This linear transformation can be expressed as in Equation (2.7), where T is the score matrix and P is the loading matrix.

$$X = TP' + E \tag{2.3}$$

2.3 Principal Component Regression (PCR)

Principal component regression (PCR) is a combination of PCA and MLR. MLR can be written in the form of score matrix, which has better properties than the original data matrix.

$$Y = TB + E \tag{2.4}$$

2.4 Partial Least Squares Regression

Partial least squares (PLS) regression has established itself as a valuable alternative for analyzing secondary variables that are highly correlated, with high measurement noise, and of high dimensionality. PLS model is built based on the properties of NIPALS algorithms by letting the score matrix represent the data matrix [38]. In PLS, the decomposition of matrix X and Y are done in such a way that the covariance is maximized. The algorithm of PLS were developed by Wold et al. [39]. The decomposition of data matrix X is done by Equation (2.3). And the decomposition of Y can also be done in a similar way by Equation (2.7), where U and Q is the score and loading matrices of Y, respectively, and F is the residual.

$$Y = UQ' + F \tag{2.5}$$

The objective of PLS is to describe maximum amount of variation in Y and get a useful relation between X and Y simultaneously. This can be done by introducing a linear model between the score matrices of X and Y.

$$U = TB \tag{2.6}$$

Consequently, matrix Y can be estimated as in Equation (2.7), F is to be minimized. The detail algorithm of PLS can be found in [38–40].

$$\hat{Y} = TBQ' + F \tag{2.7}$$

Chapter 3. Variable Selection Theory and Algorithm

Due to prompt development of technology, thousands of process measurements are collected by process computers every day. Researchers have been utilizing these data to build soft sensor, which is also known as data-driven soft sensor. By correlating the secondary variables with the primary variables, soft sensors can provide information on those immeasurable, but important variables. Furthermore, soft sensors can provide prediction on infrequently measured variable so that control actions can be taken to prevent process failure. It has been proved by many studies that the performance of soft sensor can be tremendously improved if only the few vital variables are included in soft sensor development. Consequently, variable selection has been one of the most important practical concerns in data-driven approaches. By identifying the relevant variables, variable selection can improve the prediction performance of soft sensor, reduce the computation load and model complexity, provide better insight into the nature of the process, and lower the measurement cost [2], [27].

Seven variable selection methods are explored in this work. They are selected based on their popularity, implementation practicability, complexity, and artificial based criterion. They can be categorized into four groups: iterative methods (stepwise regression and genetic algorithm combined with PLS), methods based on artificial standard (uninformative variable elimination method combined with PLS and PLS based on sensitivity analysis), enforced variable elimination methods (competitive adaptive reweighted

sampling method with PLS), and methods based on predictive properties (PLS based on variable importance in projection and PLS based on regression coefficients).

Stepwise regression has been applied to the selection of predictors for both classification and multivariate calibrations [8], especially in near-infrared (NIR) spectral. Gauchi and Chagnon proposed a stepwise variable selection method based on maximum Q^2 and applied to manufacturing processes in oil, chemical and food industries [9].

Broadhurst et al. applied genetic algorithm to pyrolysis mass spectrometric data and showed that GA is able to determine the optimal subset of variables to provide better or equal prediction performance [10]. Arcos et al. successfully applied GA to a wavelength selection for PLS calibration of mixtures of indomethacin and acemethacin, in spite of the fact that the two compounds have almost identical spectra [11]. A modified genetic algorithm-based wavelength selection method has been proposed by Hiromasa Kaneko and Kimito Funatsu to select process variables and dynamic simultaneously [12]. This method is named as genetic algorithm-based process variables and dynamics selection method, GAVDS. The result of GAVDS, based on its application to a dynamic process of distillation column in Mitsubishi Chemical Corporation, shows its robustness to the presence of nonlinearity and multicollinearity in process data. GA has also been well recognized in molecular modeling. Jones et al. have shown three application of GA in chemical structure handling and molecular recognition [13].

A modified uninformative variable elimination method based on the principle of Monte Carlo (MC) was applied in quantitative analysis of NIR spectra by Cai et al. [14]. UVE-MC is proven to be capable of selecting important wavelength and making the prediction more robust and accurate in quantitative analysis. Some researchers also suggest-

ed to combine UVE with wavelet transform to further simplify the model and to reduce computation time [14], [15]. In the work of Koshoubu et al., they have extended UVE to eliminate uninformative samples (USE) that do not contribute much in the calibration model [16], [17]. They proposed an algorithm where the uninformative wavelengths/variables are eliminated first by UVE-PLS, and then the uninformative samples, which are determined by their standard deviation of prediction error calculated from leave-one-out cross validation, are eliminated from the calibration. Another new method which combined UVE with successive projection algorithm (SPA) has been proposed by [18]. UVE is implemented to remove uninformative variables before application of SPA to improve the efficiency of variable selection by SPA.

Sensitivity analysis has become more popular in selection of optimal variable subsets in recent years. Zamprogna et al. has introduced a novel methodology based on principal component analysis (PCA) to select the most suitable soft sensor inputs [19]. Instead of using the secondary variables directly, the instantaneous sensitivity of each secondary variable to the primary variables are estimated and utilized as the regressor inputs. Li and Shao have proposed a novel method using sensitivity analysis to select the optimal secondary variables to be used as inputs to kernel ridge regression (KRR) to implement online soft sensing of distillation column compositions [20].

Competitive adaptive reweighted sampling (CARS) method has been proposed by Li et al. [21]. CARS is model independent. In other words, CARS can be combined with any regression or classification models. In [22], [23], CARS has been applied in combination with partial least squares linear discriminant analysis (PLSLDA) to effectively classify two classes of samples in colorectal cancer data.

Variable importance in the projection (VIP) and regression coefficients (BETA) have been broadly adapted as a criterion in partial least squares modeling paradigm for variable selection. Both PLS-VIP and PLS-BETA are model based variable selection methods. Mehmood et al. presented an algorithm that balances the parsimony and predictive ability of model using variables selection based on PLS-VIP [24]. It is shown that the proposed method increases the understandability and consistency of the model and reduces the classification error. Lindgren et al. also implemented PLS-VIP on a benchmark data for variable selection, Selwood dataset [25]. In their study, PLS-VIP is combined with permutation test to extensively investigate the technique. A bootstrap-PLS-VIP has been implemented as a wavelength interval selection method in spectral imaging applications by Gosselin et al. [26]. Their result demonstrates its ability to identify relevant spectral intervals and its simplicity and relatively low computational cost. PLS-VIP and PLS-BETA have also been seen in food science. Andersen and Bro applied PLS-VIP and PLS-BETA to NIR spectral of beer sample and obtained useful insight of the process [27]. A variable selection algorithm based on the standardized regression coefficients are proposed in [28]. The developed models are optimized by the leave-one-out Q^2 values and validated by an external testing set.

3.1 Stepwise Regression

Stepwise regression has been widely used for variable selection in linear regression [3]. Stepwise regression is a combination of forward selection and backward elimination methods [9]. Both are well known methods for variable selection in multiple regressions. The forward selection and backward elimination methods are done by introduction or elimination of the variables one-by-one according to the specific thresholds. In

stepwise regression, a sequence of regression models is constructed iteratively by adding or removing variables. The variables are selected according to their statistical significance in a regression [8]. Partial F-test or t-test is used for determination of its significance.

The standard stepwise regression procedure is illustrated in Figure 3.1and summarized as follows:

1. Define thresholds of probability of incorrectly rejecting the true null hypothesis, which is also known as Type I error. The threshold for adding a variable to a model is 0.05, $\alpha_{in} = 0.05$, and the threshold for removing a variable from the model is 0.1, $\alpha_{out} = 0.1$.
2. Assume the total number of variables is p, and $\Omega_1 = \{x_1, \quad x_2, \quad \dots, \quad x_k\}$ is a subset of variables included in linear regression model. The unselected variables are examined by calculating their partial F-statistic using equations (3.1) and (3.2), where SS_R is the residual sum of squares due to regression, and MS_E is the mean square error. The variable with maximum F-statistic among all the unselected ones is added to the model, provided that $F_j > F_{in}$.

$$F_j = \frac{SS_R\left(x_j \middle| x_1, x_2, \dots, x_k\right)}{MS_E\left(x_j, x_1, x_2, \dots, x_k\right)} \tag{3.1}$$

$$SS_R\left(x_j \middle| x_1, \dots, x_k\right) = SS_R\left(x_j, x_1, \dots, x_k\right) - SS_R(x_1, \dots, x_k) \tag{3.2}$$

3. Once a new subset of variables is determined, the same procedure is carried out to check if any of these variables inside the model should be re-

moved. The variable with the smallest F-statistic is removed, provided that $F_j < F_{out}$. Otherwise, the variable is retained in the model.

4. Repeat Step 2 and Step 3 until no other variables can be added into or removed from the model.

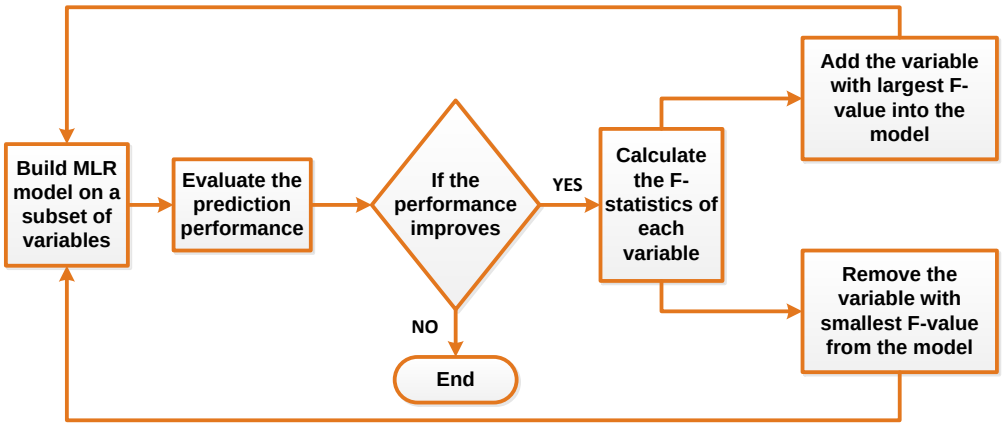

Figure 3.1 Stepwise Regression Algorithm

3.2 Genetic Algorithm

Genetic algorithm has been used widely in solving complex problems of optimization and search problems [41]. More recently, GA has been used to find the optimum subset of regressor variables for a given modeling method based on the results of cost function evaluations for all candidate genetic chromosomes [42].

The original algorithm can be found in [43–45]. Generally speaking, there are five steps in GA: coding of variables, initiation of population, evaluation of the responses, reproductions, and mutations [46]. The last three steps are implemented iteratively until a termination criterion is reached. In our work, GA combined with PLS regression model is studied. These following terms must be defined:

1. Initiation of population. Percentage of variables included in the initial population (30% -50%).

2. Population size. This value is dependent on the total number of variables. There is a tradeoff between the initial coverage of the original space and computation load.
3. Maximum number of generations (50-500). This could be used as one of the termination criterion.
4. Percentage of the population retained after each generation (50% -80%). This number defines the top percentage of populations to be kept in each generation. In other words, only the remaining populations will go through reproduction.
5. Breeding crossover rule (single or double crossover). It is analogous to reproduction. It is a genetic operator used to vary programming of chromosomes from one generation to the next.
6. Mutation rate (0.001-0.01). Chance of alternation of genes after crossover.

An initial population is generated by randomly choosing 30% of the total variables. This is repeated multiple times depending on the population size. A PLS model is built for each population/chromosomes. Populations are then sorted in descending order by its cross validation metrics. Only the top percentages of the populations are remained unchanged, and the rest will undergo crossover/reproduction. A new generation of chromosomes is then produced. This is done iteratively until a termination criterion is reached. This termination criterion can be based on the maximum number of generations or prediction improvement deficiency. The algorithm is also shown in Figure 3.2.

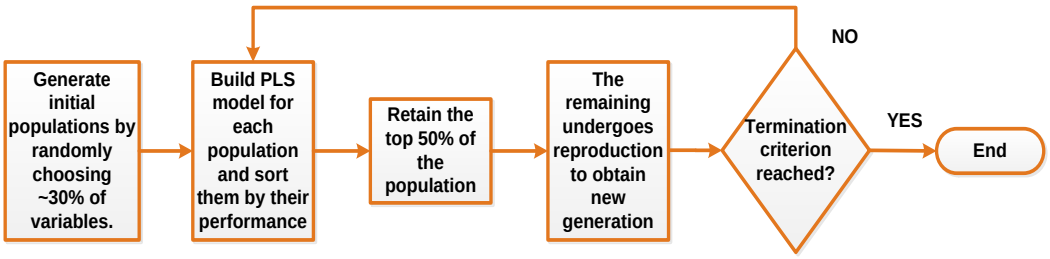

Figure 3.2 Genetic Algorithm with PLS

3.3 Uninformative Variable Elimination

A method for eliminating uninformative variables by comparing with artificial variables was proposed by V. Center, et al. [5]. Models are built using both experimental and artificial variables. The analysis is based on the regression coefficients from the model.

In our work, uninformative variable elimination by Partial Least Squares (UVE-PLS) will be studied. The procedure is illustrated in Figure 3.3 and summarized as follows:

1. For a given set of experimental variables, $X \in \mathbb{R}^{n \times p}$, generate an artificial random variable matrix with very small magnitude and same dimension as the experimental variables. This results in a matrix with dimension of n by $2p$, $XR = [X \quad R]$.
2. Build PLS model for XR based on leave-one-out procedure. This will yield a regression coefficient matrix, $B \in \mathbb{R}^{n \times 2p}$.
3. Calculate the reliability index of each variable j using Equation (3.3), where b_j and $s(b_j)$ are the mean and standard deviation of variable j obtained from leave-one-out procedure.

$$c_j = \frac{b_j}{s(b_j)} \tag{3.3}$$

$$b_j = \frac{\sum_{i=1}^{n} b_{ij}}{n} \tag{3.4}$$

$$s(b_j) = \left(\frac{\sum_{i=1}^{n} (b_{ij} - b_j)^2}{n-1} \right)^{1/2} \tag{3.5}$$

4. Determine the maximum absolute reliability index of the artificial variables, $\text{abs}(\max(c_{artif}))$. The experimental variables with absolute reliability index less than $\text{abs}(\max(c_{artif}))$ are eliminated, i.e., $abs(c_j) < \text{abs}(\max(c_{artif}))$.
5. A new PLS model is built using only the remaining variables.

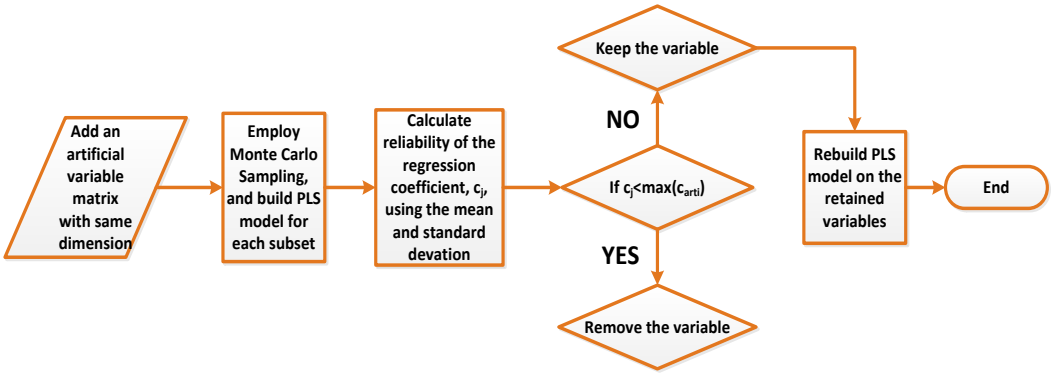

Figure 3.3 Procedure of Uninformative Variable Elimination with PLS

3.4 Partial Least Squares with Sensitivity Analysis

A novel variable selection algorithm proposed in [6] that combines Partial Least Squares with sensitivity analysis [47], PLS-SA, is investigated. The sensitivity of each variable is often expressed in terms of its regression coefficient in linear regression models. In PLS models, the coefficients calculated are a mixture of the original variables. Hence, an alternative measure of sensitivity of individual variable is proposed. In

Rueda's work, the sensitivity of each variable is defined as the absolute maximum change in the PLS prediction (maximum minus the minimum values), when the value of x_j is varied in its allowable range and all other variable are kept constant at their mean/median value [6]. This measurement is referred as $\frac{\Delta \hat{y}}{\Delta x_j}$.

In PLS-SA, the value of $\frac{\Delta \hat{y}}{\Delta x_j}$ is only computed over the training set; the remaining samples are used to measure the predictive power of the model. The relevance of each variable is determined by comparing its sensitivity to that of a random variable, RV. The effect of RV to the response variable should be insignificant since it is random. In sensitivity analysis, a RV is added to the original dataset, and then its sensitivity, $\frac{\Delta \hat{y}}{\Delta RV}$, is computed.

To balance the comparison fairness and computational load, the extended data is divided into several subsets. PLS models are built for each subset. The sensitivity values along with their averages and standard deviations are also computed. A variable x_j is found significant if inequality (3.6) is satisfied. The process is carried out in an iterative manner. In every iteration, the variables with sensitivity values below the sensitivity of random variable are eliminated. The predictive power of the new set of variables is determined. The variables are eliminated permanently only if the predictive power is improved. The process stops when no more variables can be dropped from the model. The stepwise algorithm of PLS-SA is illustrated in Figure 3.4.

$$\overline{\frac{\Delta \hat{y}}{\Delta x_j}} - StDev_{\frac{\Delta \hat{y}}{\Delta x_j}} > \overline{\frac{\Delta \hat{y}}{\Delta RV}} + StDev_{\frac{\Delta \hat{y}}{\Delta RV}} \tag{3.6}$$

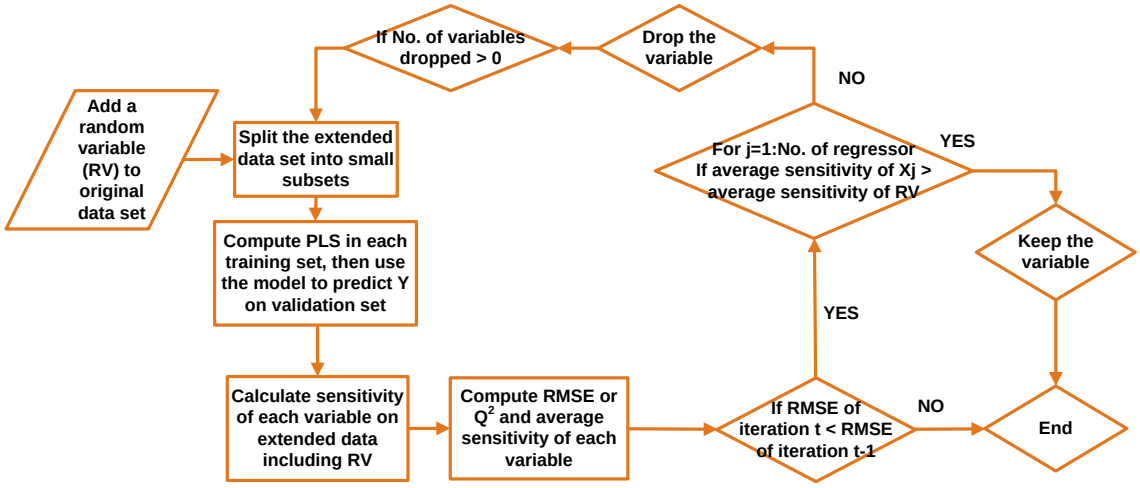

Figure 3.4 PLS-SA Algorithm

3.5 Competitive Adaptive Reweighted Sampling with Partial Least Squares

Hongdong Li et al. have proposed a novel strategy based on the principle 'survival of the fittest', named competitive adaptive reweighted sampling (CARS) [21], [22]. This method utilizes the absolute values of the regression coefficients to evaluate variables' importance. In an iterative manner, N subsets of variables are selected by CARS from N Monte Carlo (MC) sampling runs. At the end, cross validation is employed to evaluate each subset. The general procedure can be described as follows and shown in Figure 3.7:

1. In each MC sampling run, a PLS model is built using 80-90% of the randomly selected samples. The regression coefficients are normalized using Equation (3.7), where p is the total number of variables.

$$nb_j = \frac{b_j}{\sum_{j=1}^{p} b_j} \tag{3.7}$$

2. In CARS, an exponentially decreasing function (EDF) is introduced as in Equation (3.8). EDF is utilized to eliminate variables with relatively small

absolute regression coefficients by force. The ratio of variables to be retained in the i^{th} sampling run is calculated by Equation (3.8) to (3.10), where

$$r_i = de^{-hi} \tag{3.8}$$

$$d = \left(\frac{p}{2}\right)^{\frac{1}{N-1}} \tag{3.9}$$

$$h = \frac{\ln(p-1)}{N-1} \tag{3.10}$$

3. Adaptive reweighted sampling (ARS) is followed by EDF-based reduction to further eliminate variables in a competitive way. In other words, variables with larger regression coefficients will be selected with higher frequency.

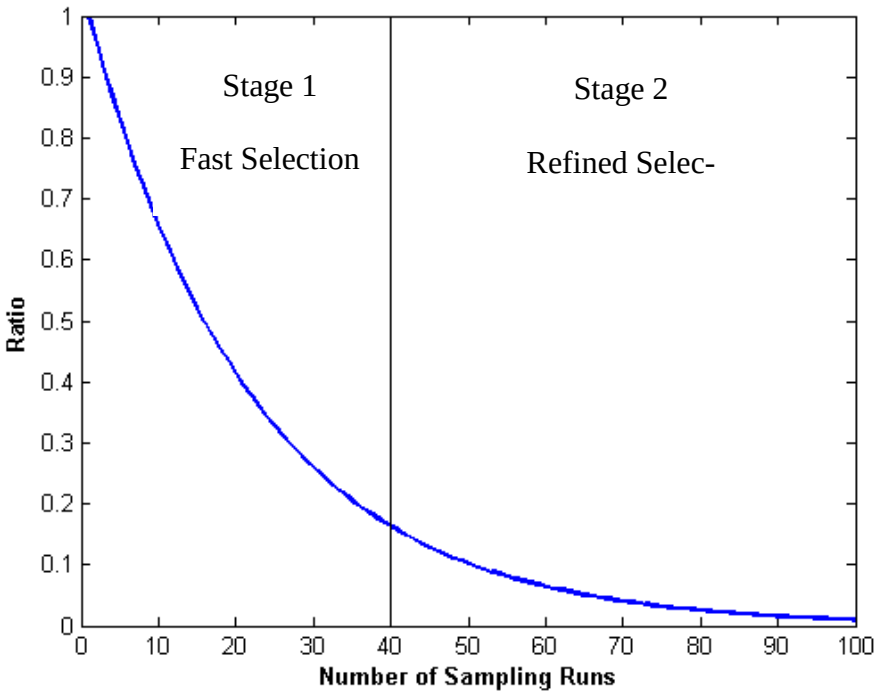

Figure 3.5 Graphical illustration of the exponentially decreasing function

The EDF process in Step 2 is roughly divided into two stages. In the first stage, the variables are eliminated rapidly, so it is called fast selection. In the second stage, the variables are eliminated in a much slower fashion, thus it is called refined selection. An example of EDF is shown in Figure 3.5. Hence, EDF becomes a very efficient algorithm for removing the variables with little information.

The ARS in Step 3 mimics 'survival of fittest' principle. The idea of ARS is illustrated in Figure 3.6. Three scenarios are considered, equal weight, little weight difference, and large weight difference.

	Weights of Variables						Sampled Variable				
	1	2	3	4	5						
Case 1:	0.20	0.20	0.20	0.20	0.20	→	2	1	3	4	5
Case 2:	0.30	0.30	0.20	0.10	0.10	→	1	1	2	3	2
Case 3:	0.40	0.05	0.40	0.10	0.05	→	1	3	3	3	1

Figure 3.6 Illustration of adaptive reweighted sampling technique using five variables in three cases as an example. The variables with larger weights will be selected with higher frequency.

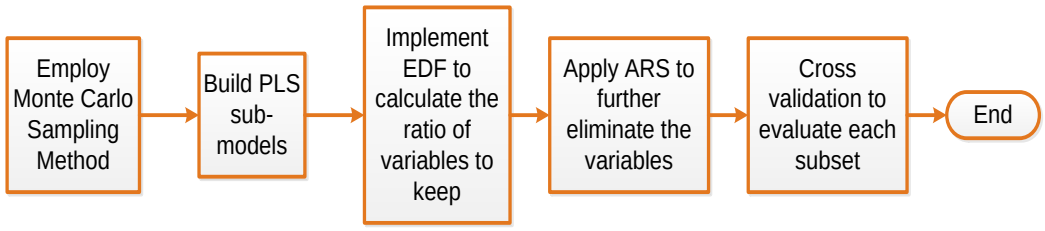

Figure 3.7 General Procedure of CAR-PLS

3.6 Partial Least Squares with Variable Important in Projection

Variable importance in the projection (VIP) score estimates the importance of each variable in the projection used in a PLS model. It was first published in [48]. The VIP score for the j^{th} variable can be calculated using Equation (3.11), where $SS(b_a t_a) = b_a^2 t_a^t t_a$. t_a is the a^{th} column vector of score matrix T. b_a is the a^{th} element of regression

coefficient vector b. w_a is the a^{th} column vector of weighting matrix W. It gives the weighted variability of j^{th} variable in the retained dimensions. VIP score calculates the contribution of each variable according to variance explained by each PLS component [26]. The expression $w_{ja}/\|w_a\|$ represents the importance of j^{th} variable in the a^{th} PLS component. The $SS(b_a t_a)$ is the variance of y explained by the a^{th} PLS component. And the summation of $SS(b_a t_a)$, denominator term, is the total variance explained by the PLS model with A components.

$$VIP_j = \sqrt{p \sum_{a=1}^{A} \left(SS(b_a t_a) \left(\frac{w_{ja}}{\|w_a\|} \right)^2 \right) / \sum_{a=1}^{A} SS(b_a t_a)} \qquad (3.11)$$

A variable selection method based on VIP scores estimated by PLS regression model is known as PLS-VIP. In general, 'greater than one rule' is used as criterion for variable selection. In other words, only variables with VIP values greater than one are considered significant. However, it has been suggested by Il-Gyo Chong et al. that the proper cutoff value for VIP can be utilized to increase the performance of PLS-VIP [4]. This value is defined by the following equation:

$$v_i^* = \frac{\left\{ Min\left(\underset{v \in \{0.01, 0.02, \ldots, 3\}}{\arg max} \; G(v) \right) + Max\left(\underset{v \in \{0.01, 0.02, \ldots, 3\}}{\arg max} \; G(v) \right) \right\}}{2} \qquad (3.12)$$

where v varies from 0.01 to 3 with increments of 0.01. And G, the geometric mean of sensitivity and specificity, is a function of v defined by Equation (3.13). Sensitivity is defined as proportion of selected relevant predictors among relevant predictors. Specificity is the proportion of unselected irrelevant predictors among irrelevant predictors. They are both calculated from the confusion matrix shown in Table 3.1. The every v value

chosen, the elements in the confusion matrix will change. Therefore, the sensitivity, specificity, and G will change as well. The value of G ranges between 0 and 1, where 1 indicates all the predictors are classified correctly. For every run/replication, the v values that maximize G are identified. The optimal cutoff value for VIP, v^*, is obtained by taking the average of the identified v's using Equation (3.12).

$$G = (Sensitivity \times Specificity)^{1/2} \tag{3.13}$$

$$Sensitivity = d/(c+d) \tag{3.14}$$

$$Specificity = a/(a+b) \tag{3.15}$$

Table 3.1 Confusion Matrix and Descriptions of Its Entries

		Predicted classes	
		Irrelevant predictor (IR)	Relevant predictor (R)
True classes	Irrelevant predictor (IR)	a: the number of irrelevant predictors classified correctly	b: the number of irrelevant predictors classified incorrectly
	Relevant predictor (R)	c: the number of relevant predictors classified incorrectly	d: the number of relevant predictors classified correctly

Overall PLS-VIP procedure can be described as follows and presented in Figure 3.8:

1. Build PLS model using all the variables. Apply cross validation to determine the optimal number of PC's.
2. Calculate VIP score for each variable using Equation (3.11).
3. Select variables with VIP scores greater than the cutoff value.

4. Calculate G and the proper cutoff value using Equations (3.13) and (3.12). Repeat Step 3 with the new cutoff value found. (Note: This step is only assessable for simulated case study.)
5. Rebuild PLS model with only the retained variables.
6. Evaluate the model performance using different indexes.

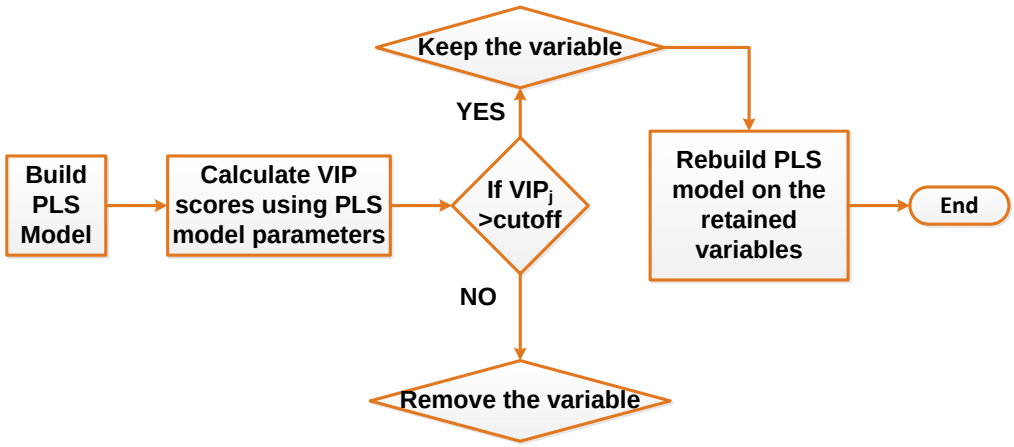

Figure 3.8 Procedure of PLS-VIP

3.7 Partial Least Squares with Regression Coefficients

Partial least squares with regression coefficients is a variable selection method that is very similar to PLS-VIP. It is also known as PLS-BETA. The only difference is PLS-BETA utilizes the regression coefficients estimated by PLS regression instead of VIP scores. The significant variables are selected according to the magnitude of the absolute values of the regression coefficients. The procedure of PLS-BETA is illustrated in Figure 3.9.

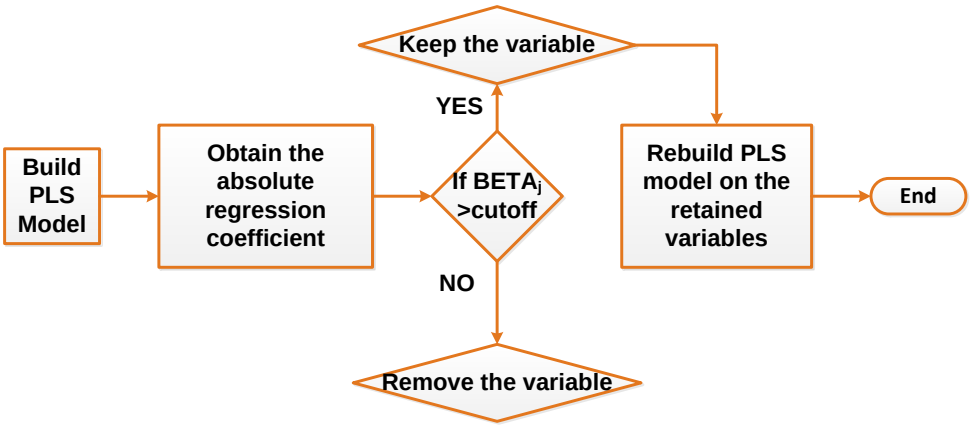

Figure 3.9 Procedure of PLS-BETA

Chapter 4. Variable Selection Method with Its Application to Simulated and Industrial Dataset

4.1 Introduction

The characteristics of these seven variable selection methods will be illustrated using a simulated case study and an industrial case study. The results presented here are based on the rules of thumb of each method to choose the model parameters, for the purpose of just comparing the base line of the methods studied. Further tuning of the parameters can be done to optimize the performance of each model.

The simulation case is generated to mimic the typical characteristics of industrial data by considering four factors: proportion of relevant predictors, magnitude of correlation between predictors, structure of regression coefficients, and magnitude of signal to noise ratio. A detailed description of data generation will be provided. The industrial case study is focused on the process data of polyester resin production plant. A brief specification of the plant will be included, followed by discussion of characteristics of batch process and their necessary preprocessing steps. The results and comparison of variable selections on both simulated and industrial case studies will be investigated. Two aspects are studied: correctly identify all the variables and prediction performance. The former one is only applicable in simulated case study, where the ground truth of the data is known. It is evaluated by the geometric mean of sensitivity and specificity discussed in Chapter 3. In this aspect, we also look at the consistency of the models produced by each variable selection method. In other words, the robustness of each method to data selection is explored. For both case studies, the data are permuted 100 times to generate different

combinations of training and validation sets. Frequency plots of selection of each variable are generated to assess the consistency of the models. The second aspect is also one of the most important factors in soft sensor development, since the optimal goal is to improve the prediction performance of soft sensor schemes. Two performance metrics are considered to evaluate the prediction performance: root mean square error (RMSE) and mean absolute percentage error (MAPE).

$$RMSE = \sqrt{\frac{\sum_i^n (y_i - \hat{y}_i)^2}{n}} \tag{4.1}$$

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \% \tag{4.2}$$

4.2 Simulated Case Study

Four factors are considered in data generation to mimic the characteristic of industrial data. They are proportion of the number of relevant predictors, the magnitude of correlations between predictors, the structure of regression coefficients, and the magnitude of signal to noise ratio. The dataset is generated following a linear model as in(4.3),

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \tag{4.3}$$

where ε_i is normal distributed random error with zero mean and specified standard deviation (described below). The data matrix X of 500 sample points is generated considering the four factors.

- For convenience, the number of relevant predictors is set to be 10. The total number of predictors, p, in data matrix can be varied, which would yield different

proportion of relevant predictors. The total number of predictors, p, in data matrix are varied in three levels, 20, 40 and 100.

$$proportion = \frac{10}{p} \tag{4.4}$$

- Data matrix X is generated from multivariate normal distribution with zero mean vector and variance-covariance matrix of Γ. The elements of matrix Γ are function of the magnitude of correlations between predictors, ρ. The magnitude of correlations between predictors, ρ, is also varied in three levels, 0.5, 0.7 and 0.9.

$$\Gamma_{ij} = \rho^{|i-j|}, \qquad (i, j = 1, \quad 2, \quad \dots \quad p) \tag{4.5}$$

Two types of equal and unequal coefficients are compared. Each type has two levels according to their locations of relevant predictors: in the middle of the range and at the extremes. All the irrelevant predictors have zero coefficients in both types. For the case with 10 relevant predictors, the regression coefficients are generated as follows:

- Equal coefficients in the middle of range

$$\beta_j = 1, \qquad \left(j = \frac{p}{2} - 4, \frac{p}{2} - 3, \dots, \frac{p}{2} + 5\right) \tag{4.6}$$

- Equal coefficients at the extreme

$$\beta_j = 1, \qquad (j = 1,2, \dots, 5; p - 4, p - 3, \dots, p) \tag{4.7}$$

- Unequal coefficients in the middle of range

$$\beta_j = (5.5 - |j - 0.5(p + 1)|)^2, \qquad \left(j = \frac{p}{2} - 4, \frac{p}{2} - 3, \dots \frac{p}{2} + 5\right) \tag{4.8}$$

- Unequal coefficients at the extreme

$$\beta_j = \left(|j - 0.5(p+1)| - 0.5(p-11)\right)^2, \quad (j = 1,2,\dots,5; p-4, p-3, \dots, p) \tag{4.9}$$

- The magnitude of signal to noise ratio is introduced by manipulating the standard deviation of error terms in y, where k is the reciprocal of signal to noise ratio. The magnitude of reciprocal of signal to noise ratio, k, is varied in three levels as well, 0.33, 0.74 and 1.22.

$$\sigma = k\sqrt{var(X\beta)} \tag{4.10}$$

4.2.1 Results

All seven variable selection methods are implemented with the simulated case study. The four parameters are varied one at a time while holding the others constant. The sensitivity results of these fours parameters considered are summarized in Table 4.1 to Table 4.4. The individual result is compared with model before and after variable selection. PLS-BETA performs the best in the sensitivity of proportion of relevant predictors. It outperforms other variables selection methods in all three levels of total number of predictors. The improvement in MAPE of the validation set runs from 1% to 9%. PLS-VIP performs the best when the correlation is at its highest level. However, the improvement is not significant. PLS-BETA yields best performance for correlation at the lower two levels, by 2-3% in MAPE. UVE-PLS gives best performance in the case with unequal regression coefficients with improvement of 3% in MAPE. Surprisingly, CARS-PLS shows performance improvement of 7% in terms of MAPE when the signal to noise ratio is at its lowest.

To visualize the result, they are also demonstrated in Figure 4.1 to Figure 4.20. The performances are evaluated using the geometric mean of sensitivity and specificity (G), root mean square error (RMSE), and mean absolute percentage error (MAPE). The values plotted in Figure 4.5 to Figure 4.20 are the improvement compared to the full models. The ones with values higher than zero indicate improvement of reduced models compared to the full model; and the ones with values lower than zero imply performance deterioration of the reduced models.

From Figure 4.1 to Figure 4.4, one can see that all the variables selection methods yields relatively high G values except PLS-SA and CARSPLS. The sensitivities of prediction performance of each variable selection method to different data generation parameters are illustrated in Figure 4.5 to Figure 4.20. The results are presented in percentage improved in the average performance metrics. The percentage improvement is calculated by comparing the reduced models to their corresponding full model of each case. The results of calibration models shown in Figure 4.5 and Figure 4.12 indicate no improvement from the models produced by the variable selection methods compared with the full models. Especially for PLS-SA, the performance deteriorates by 75% in RMSE and 77% in MAPE. From the prediction performance of validation models shown in Figure 4.13 to Figure 4.20, the prediction performance of the reduced models are improved compared with the full models, with exception of PLS-SA. Performance of PLS-BETA worsens significantly when the regression coefficients are unequal. PLS-BETA only selects the variables with larger regression coefficients. In other words, even if the variable is relevant to the primary variable, it is not selected by PLS-BETA since its coefficient is relatively small.

Table 4.1 Comparison of Sensitivity of Different Variable Selection Methods to Proportion of Relevant Predictors

				Training		Validation	
Model	p	No. Sel	G	RMSE	MAPE	RMSE	MAPE
Full	20	20	--	1.5357	3.0418	1.5724	3.1222
	40	40	--	1.4802	3.0229	1.6177	3.3173
	100	100	--	1.3914	2.8289	1.7140	3.4910
SR	20	10+/-1	0.9767	1.5463	3.0641	1.5613	3.1014
	40	11+/-1	0.9784	1.5131	3.0884	1.5828	3.2434
	100	14+/-2	0.9765	1.4920	3.0349	1.6052	3.2716
GA	20	12+/-1	0.8721	1.5858	3.1446	1.6110	3.2032
	40	16+/-2	0.8746	1.5510	3.1687	1.6417	3.3622
	100	26+/-5	0.9008	1.4840	3.0181	1.6446	3.3517
UVE	20	12+/-1	0.9989	1.5465	3.0639	1.5612	3.1021
	40	12+/-1	0.9856	1.5199	3.1008	1.5758	3.2303
	100	12+/-1	0.9947	1.5234	3.0991	1.5723	3.2036
SA	20	14+/-2	0.4580	2.1686	4.3078	2.2198	4.4048
	40	23+/-3	0.4720	2.5455	5.2190	2.6748	5.4793
	100	55+/-6	0.4991	2.4337	4.9977	2.7491	5.6427
CARS	20	18+/-4	0.1722	1.5381	3.0471	1.5700	3.1177
	40	20+/-13	0.7057	1.5065	3.0738	1.5900	3.2584
	100	17+/-17	0.9384	1.5024	3.0560	1.5933	3.2448
VIP	20	10+/-0	0.9918	1.5807	3.1335	1.5930	3.1667
	40	11+/-1	0.9874	1.5230	3.1075	1.5721	3.2219
	100	13+/-1	0.9849	1.5247	3.1023	1.5710	3.2007
BETA	20	10+/0	1	1.5504	3.0732	**1.5571**	**3.0936**
	40	10+/-0	1	1.5241	3.1097	**1.5712**	**3.2204**
	100	10+/-0	1	1.5287	3.1097	**1.5659**	**3.1893**

Table 4.2 Comparison of Sensitivity of Different Variable Selection Methods to Magnitude of Correlation between Predictors

				Training		Validation	
Model	ρ	No. Sel	G	RMSE	MAPE	RMSE	MAPE
Full	0.5	40	--	1.4802	3.0229	1.6177	3.3173
	0.7	40	--	1.7618	3.0547	1.8912	3.2944
	0.9	40	--	2.1383	3.1542	2.2216	3.2872
SR	0.5	11+/-1	0.9784	1.5131	3.0884	1.5828	3.2434
	0.7	11+/-1	0.9836	1.7854	3.0947	1.8581	3.2348
	0.9	10+/-1	0.9526	2.1409	3.1561	2.2252	3.2908
GA	0.5	16+/-2	0.8746	1.5510	3.1687	1.6417	3.3622
	0.7	15+/-3	0.8873	1.7979	3.1205	1.8937	3.2956
	0.9	15+/-3	0.8344	2.1432	3.1594	2.2484	3.3246
UVE	0.5	12+/-1	0.9856	1.5199	3.1008	1.5758	3.2303
	0.7	14+/-2	0.9636	1.7882	3.0994	1.8568	3.2331
	0.9	28+/-3	0.8452	2.1383	3.1511	2.2220	3.2879
SA	0.5	23+/-3	0.4720	2.5455	5.2190	2.6748	5.4793
	0.7	28+/-3	0.4566	2.2565	3.9257	2.3950	4.1815
	0.9	36+/-3	0.2399	2.1798	3.2195	2.2625	3.3511
CARS	0.5	20+/-13	0.7057	1.5065	3.0738	1.5900	3.2584
	0.7	20+/-13	0.6731	1.7800	3.0856	1.8697	3.2553
	0.9	18+/-12	0.7439	2.1441	3.1603	2.2220	3.2850
VIP	0.5	11+/-1	0.9874	1.5230	3.1075	1.5721	3.2219
	0.7	13+/-1	0.9536	1.7913	3.1040	1.8532	3.2277
	0.9	16+/-1	0.9000	2.1412	3.1543	**2.2087**	**3.2693**
BETA	0.5	10+/-0	1	1.5241	3.1097	**1.5712**	**3.2204**
	0.7	10+/-0	0.9995	1.7959	3.1132	**1.8492**	**3.2198**
	0.9	10+/-0	0.9851	2.1552	3.1772	2.2162	3.2759

Table 4.3 Comparison of Sensitivity of Different Variable Selection Methods to Regression Coefficient Structure

Model	*s*	No. Sel	G	Training		Validation	
				RMSE	MAPE	RMSE	MAPE
Full	EM	40	--	1.6078	2.9768	1.7471	3.2427
	EE	40	--	1.4802	3.0229	1.6177	3.3173
	UM	40	--	21.80	3.0353	23.54	3.2896
	UE	40	--	18.50	3.0014	20.19	3.2909
SR	EM	11+/-1	0.9779	1.6442	3.0432	1.7093	3.1696
	EE	11+/-1	0.9784	1.5131	3.0884	1.5828	3.2434
	UM	9+/-1	0.8830	22.34	3.1090	23.11	3.2238
	UE	9+/-1	0.8861	18.93	3.0704	19.84	3.2307
GA	EM	16+/-3	0.8865	1.6558	3.0641	1.7377	3.2224
	EE	16+/-2	0.8746	1.5510	3.1687	1.6417	3.3622
	UM	14+/-3	0.8065	22.52	3.1327	23.54	3.2850
	UE	14+/-3	0.8091	19.20	3.1129	20.28	3.3039
UVE	EM	12+/-1	0.9928	1.6502	3.0546	1.7021	3.1576
	EE	12+/-1	0.9856	1.5199	3.1008	1.5758	3.2303
	UM	9+/-2	0.9281	22.46	3.1262	**22.95**	**3.2012**
	UE	9+/-1	0.8944	19.06	3.0897	**19.70**	**3.2098**
SA	EM	24+/-3	0.4866	2.5117	4.6580	2.6385	4.8940
	EE	23+/-3	0.4720	2.5455	5.2190	2.6748	5.4793
	UM	22+/-3	0.4950	36.44	5.1019	38.03	5.3456
	UE	22+/-3	0.4797	35.04	5.7135	36.74	6.0067
CARS	EM	19+/-12	0.7196	1.6373	3.0301	1.7154	3.1844
	EE	20+/-13	0.7057	1.5065	3.0738	1.5900	3.2584
	UM	16+/-11	0.6912	22.40	3.1169	23.23	3.2568
	UE	17+/-12	0.6584	19.00	3.0799	20.03	3.2644
VIP	EM	10+/-1	0.9948	1.6557	3.0650	1.6963	3.1466
	EE	11+/-1	0.9874	1.5230	3.1075	1.5721	3.2219
	UM	7+/-1	0.8618	22.74	3.1654	23.06	3.2189
	UE	8+/-1	0.9095	19.13	3.1034	19.71	3.2112
BETA	EM	10+/-0	1	1.6563	3.0658	**1.6957**	**3.1458**
	EE	10+/-0	1	1.5241	3.1097	**1.5712**	**3.2204**
	UM	5+/-1	0.6977	25.09	3.4904	25.67	3.5808
	UE	5+/-1	0.6932	22.23	3.6071	22.85	3.7205

Table 4.4 Comparison of Sensitivity of Different Variable Selection Methods to Magnitude of Signal to Noise Ratio

				Training		Validation	
Model	*k*	No. Sel	G	RMSE	MAPE	RMSE	MAPE
Full	0.33	40	--	1.6078	2.9768	1.7471	3.2427
	0.74	40	--	3.6020	5.6126	3.9212	6.1351
	1.22	40	--	5.9373	7.3677	6.4675	8.0715
SR	0.33	11+/-1	0.9779	1.6442	3.0432	1.7093	3.1696
	0.74	11+/-1	0.9757	3.6872	5.7435	3.8349	5.9951
	1.22	9+/-2	0.8775	6.1020	7.5759	6.3888	7.9648
GA	0.33	16+/-3	0.8865	1.6558	3.0641	1.7377	3.2224
	0.74	16+/-2	0.8781	3.6853	5.7405	3.8758	6.0561
	1.22	15+/-3	0.8435	6.0636	7.5197	6.3973	7.9827
UVE	0.33	12+/-1	0.9928	1.6502	3.0546	1.7021	3.1576
	0.74	11+/-1	0.9928	3.6988	5.7612	3.8195	5.9728
	1.22	11+/-2	0.9928	6.1032	7.5697	6.3022	7.8635
SA	0.33	24+/-3	0.4866	2.5117	4.6580	2.6385	4.8940
	0.74	22+/-3	0.4837	3.6020	5.6126	3.9212	6.1351
	1.22	21+/-3	0.4939	5.9373	7.3677	6.4675	8.0715
CARS	0.33	19+/-12	0.7196	1.6373	3.0301	1.7154	3.1844
	0.74	24+/-13	0.5429	3.6630	5.7062	3.9015	6.1003
	1.22	20+/-13	0.5948	6.0540	7.5116	6.4418	**7.5116**
VIP	0.33	10+/-1	0.9948	1.6557	3.0650	1.6963	3.1466
	0.74	10+/-1	0.9936	3.7122	5.7838	**3.8046**	**5.9502**
	1.22	10+/-1	0.9902	6.1185	7.5934	**6.2758**	7.8297
BETA	0.33	10+/-0	1	1.6563	3.0658	**1.6957**	**3.1458**
	0.74	9+/-1	0.9596	3.7381	5.8259	3.8548	6.0291
	1.22	8+/-1	0.8462	6.1565	7.6452	6.4054	7.9873

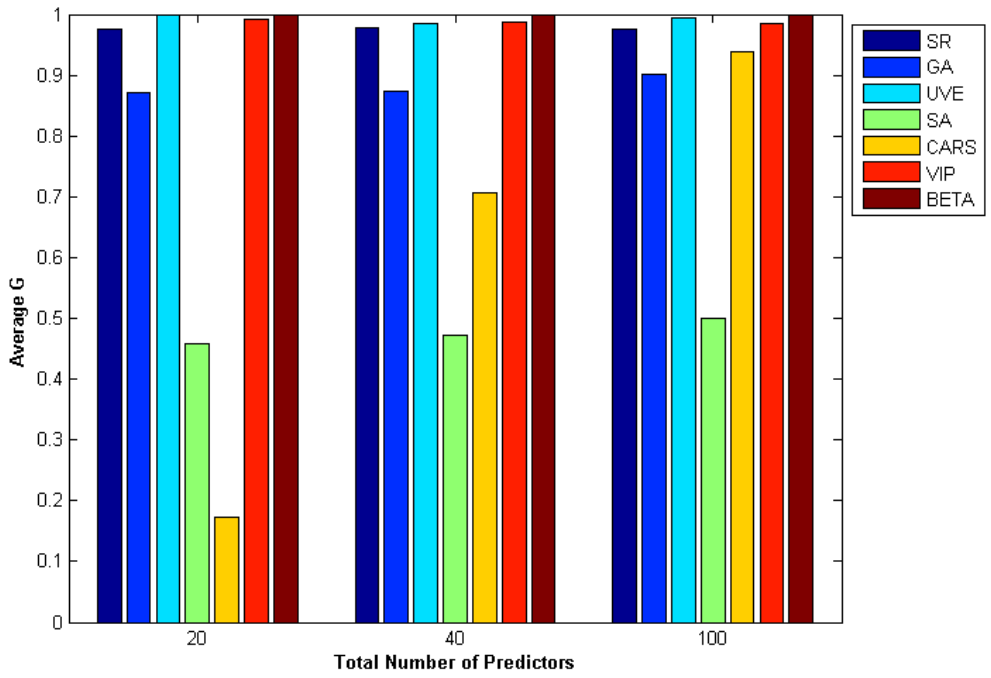

Figure 4.1 Sensitivity of Proportion of Relevant Predictors in Terms of Average G

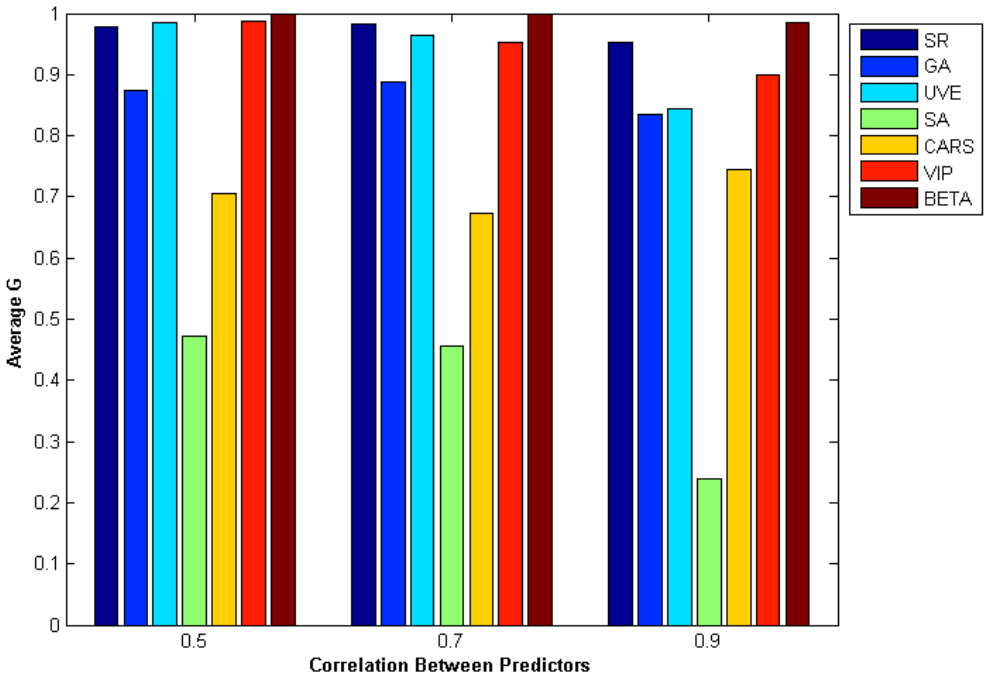

Figure 4.2 Sensitivity of Magnitude of Correlation between Predictors in Terms of Average G

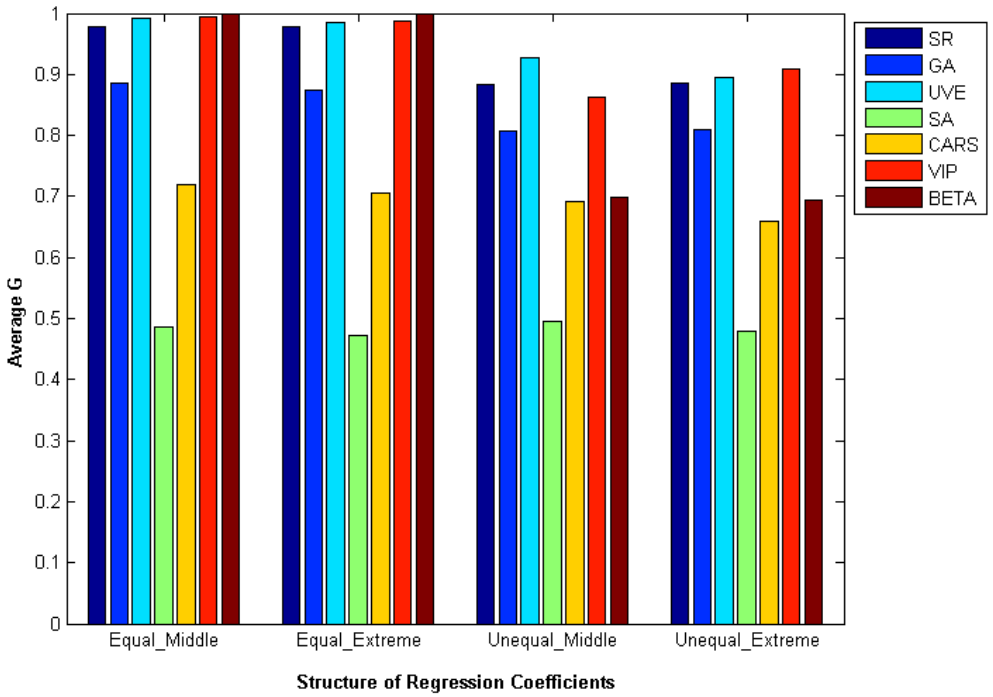

Figure 4.3 Sensitivity of Regression Coefficient Structure in Terms of Average G

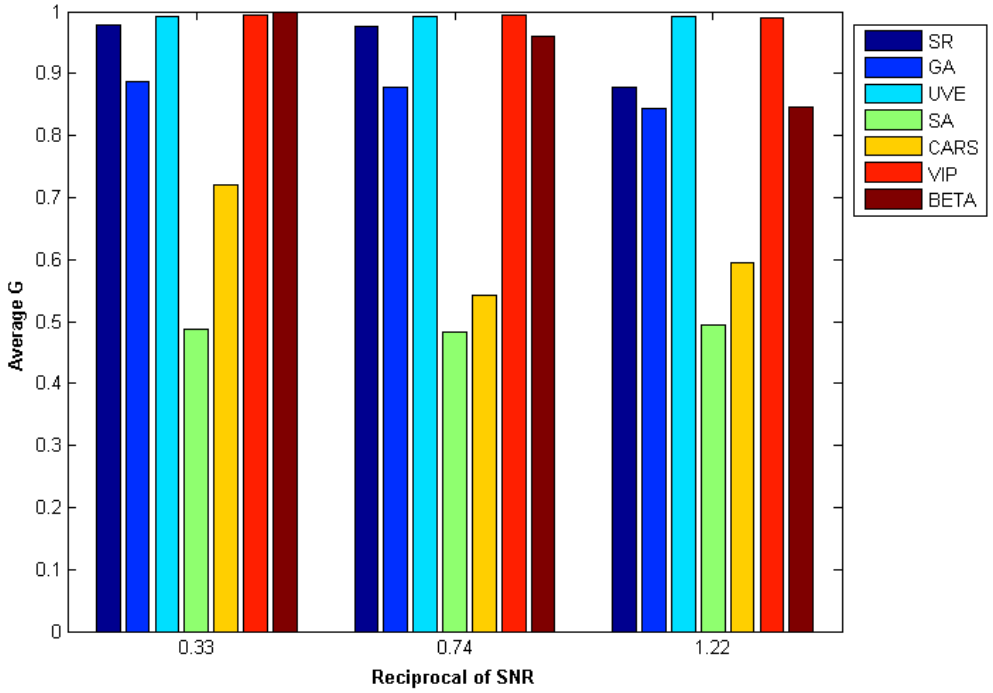

Figure 4.4 Sensitivity of Magnitude of Signal to Noise Ratio in Terms of Average G

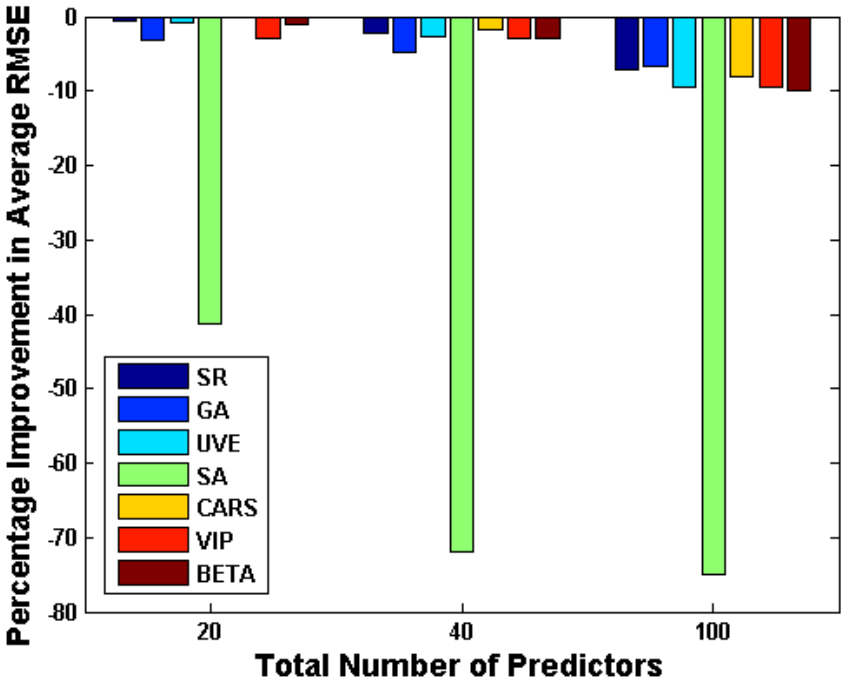

Figure 4.5 Sensitivity of Proportion of Relevant Predictors in Terms of Average RMSE in Training Set

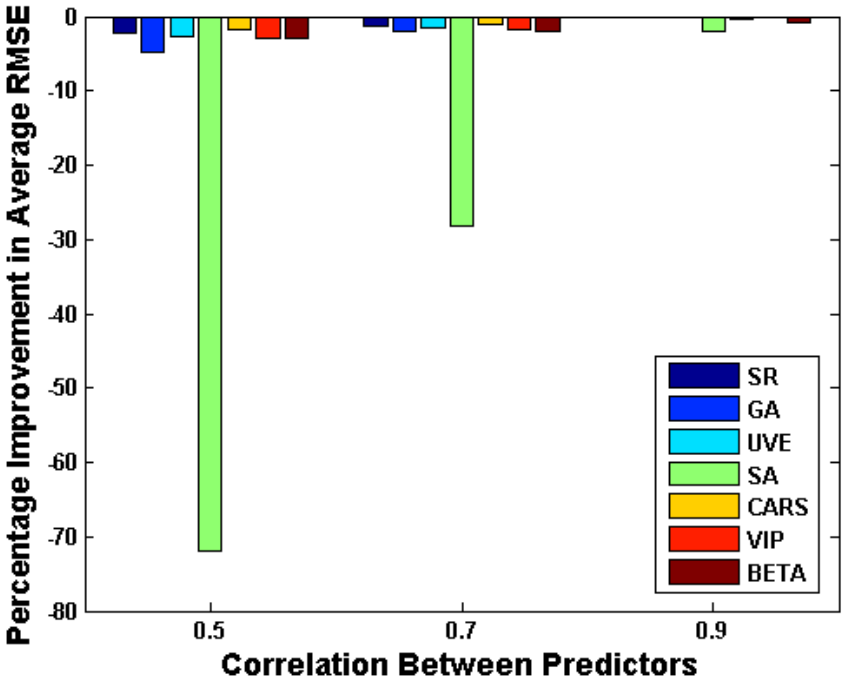

Figure 4.6 Sensitivity of Magnitude of Correlation between Predictors in Terms of Average RMSE in Training Set

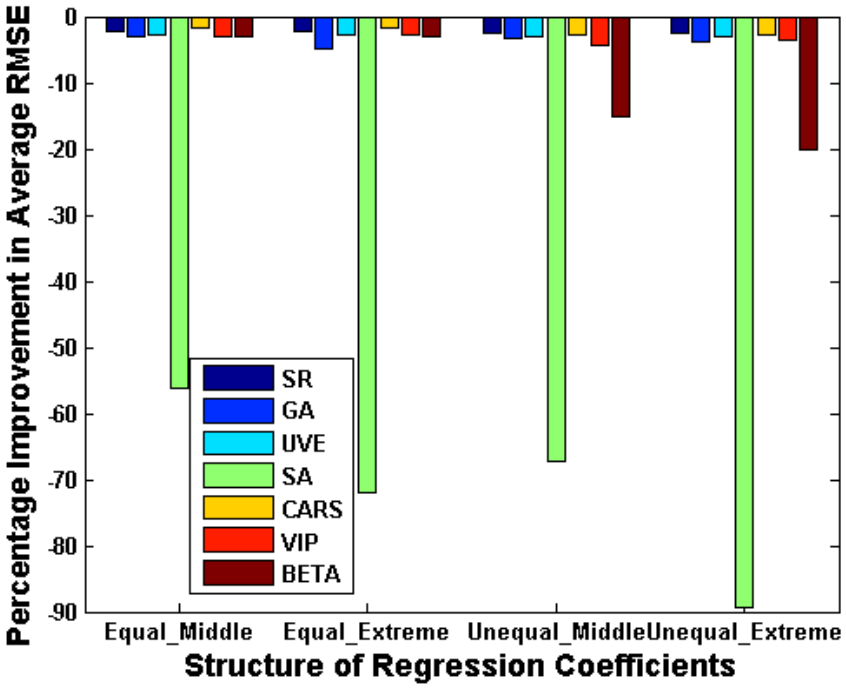

Figure 4.7 Sensitivity of Regression Coefficient Structure in Terms of Average RMSE in Training Set

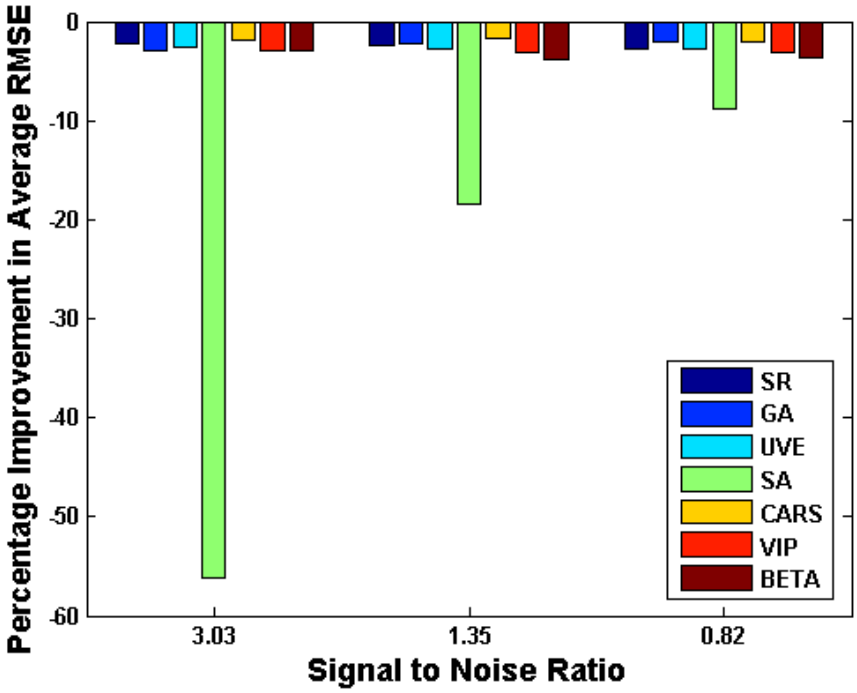

Figure 4.8 Sensitivity of Magnitude of Signal to Noise Ratio in Terms of Average RMSE In Training Set

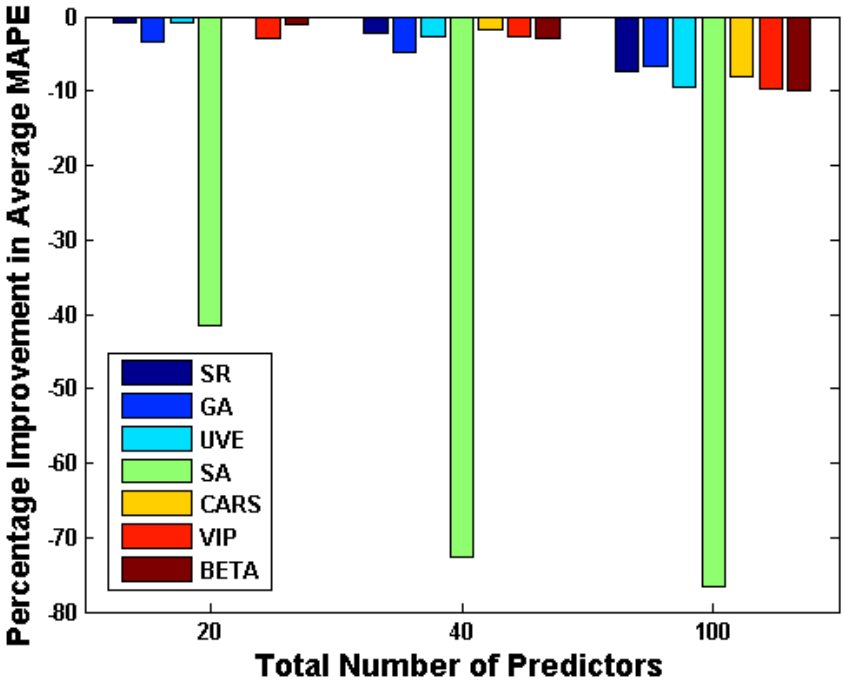

Figure 4.9 Sensitivity of Proportion of Relevant Predictors in Terms of Average MAPE in Training Set

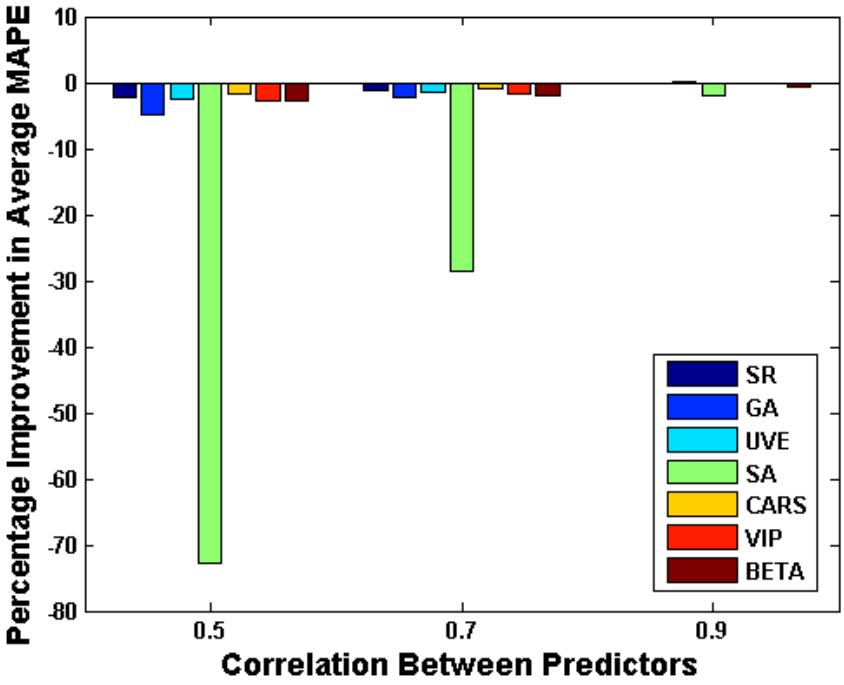

Figure 4.10 Sensitivity of Magnitude of Correlation between Predictors in Terms of Average MAPE in Training Set

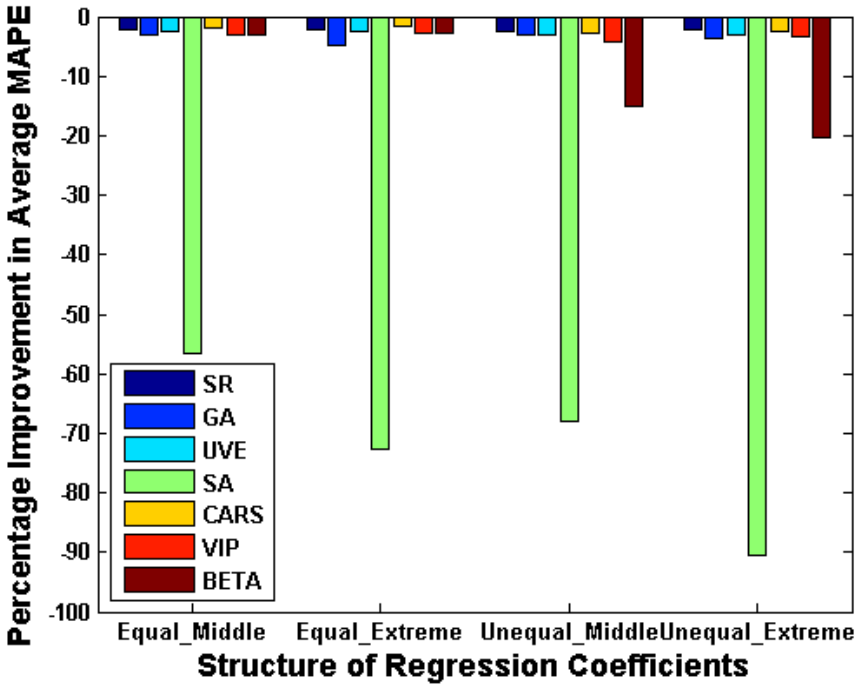

Figure 4.11 Sensitivity of Regression Coefficient Structure in Terms of Average RMSE in Training Set

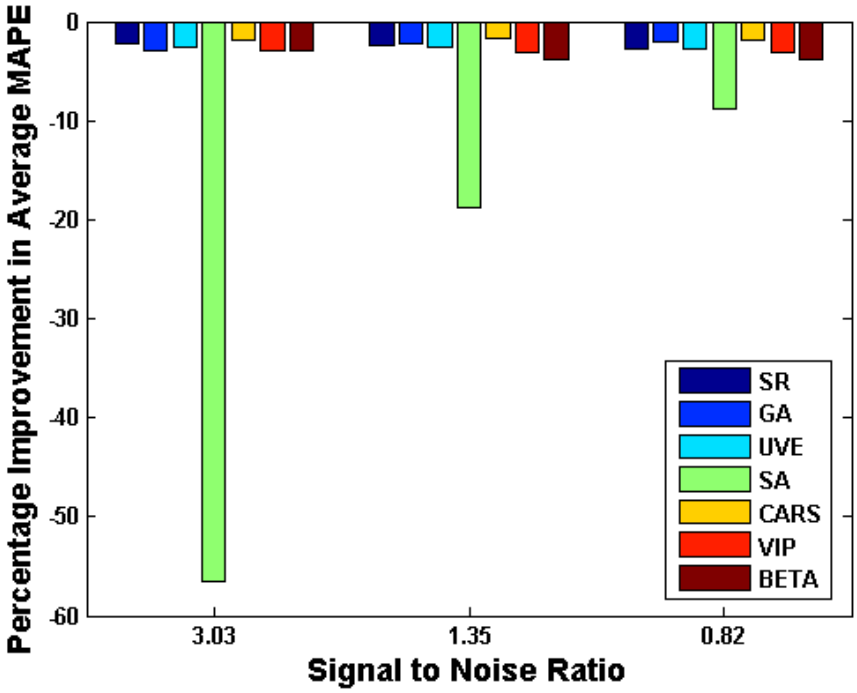

Figure 4.12 Sensitivity of Magnitude of Signal to Noise Ratio in Terms of Average MAPE in Training Set

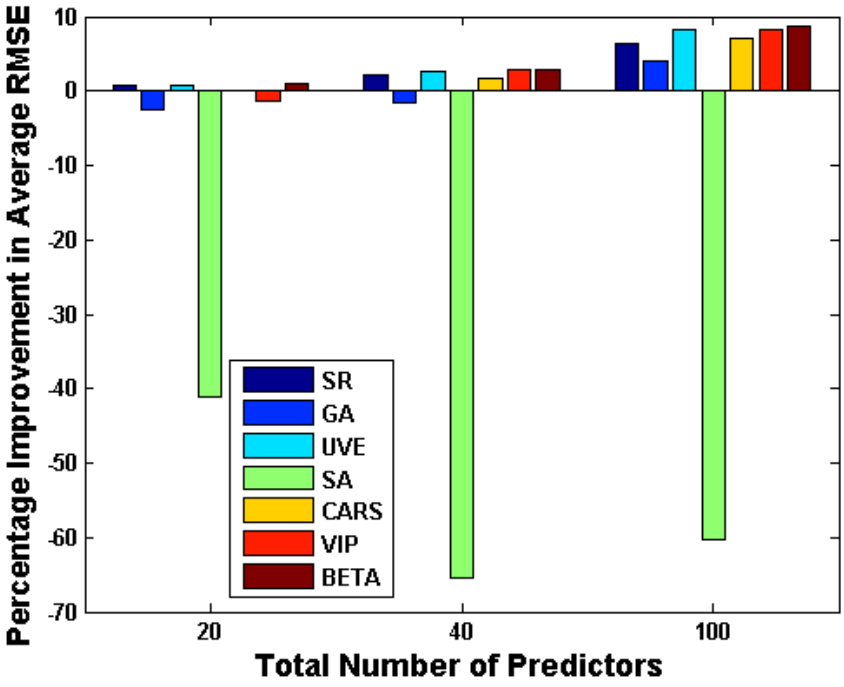

Figure 4.13 Sensitivity of Proportion of Relevant Predictors in Terms of Average RMSE in Validation Set

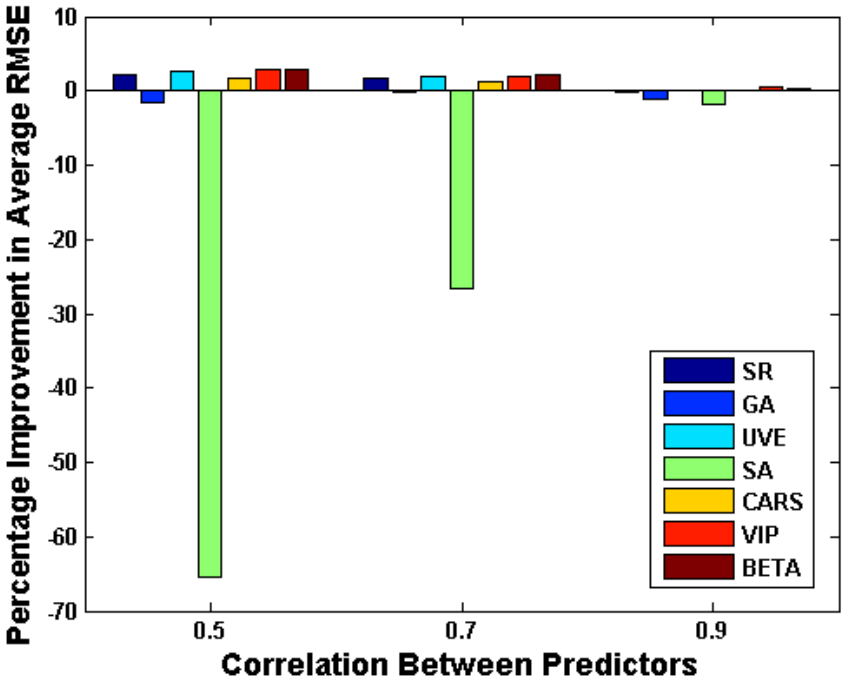

Figure 4.14 Sensitivity of Magnitude of Correlation between Predictors in Terms of Average RMSE in Validation Set

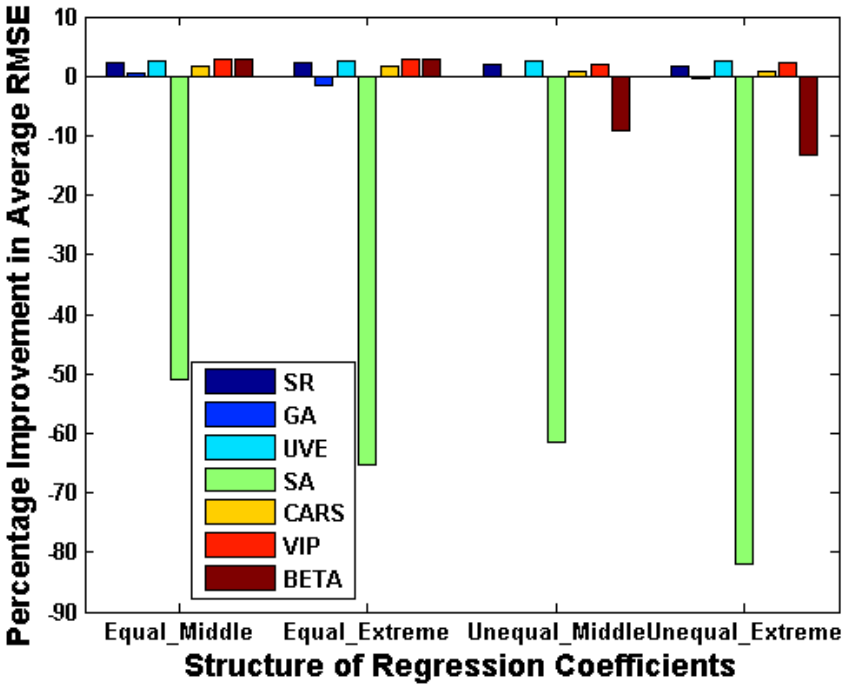

Figure 4.15 Sensitivity of Regression Coefficient Structure in Terms of Average RMSE in Validation Set

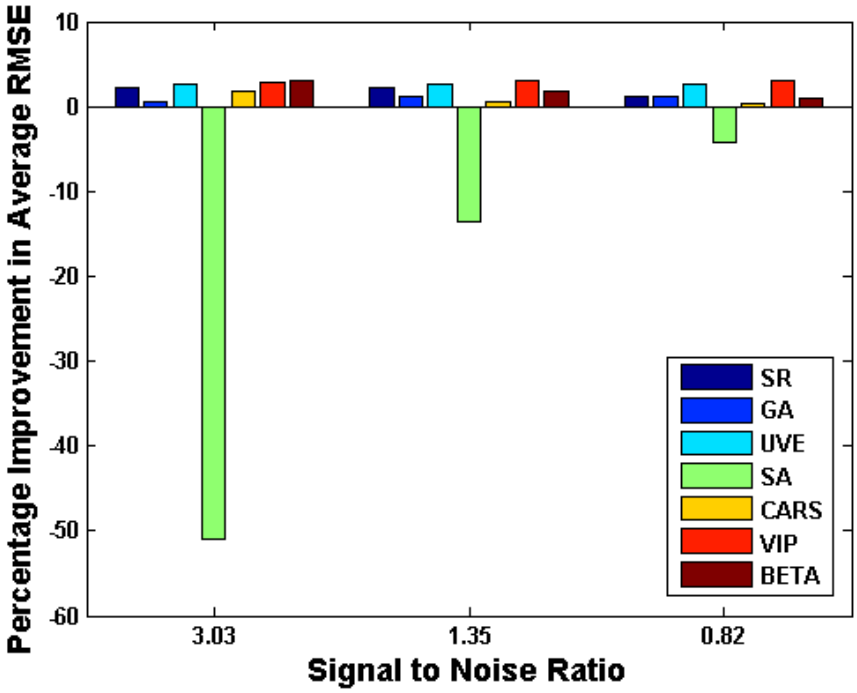

Figure 4.16 Sensitivity of Magnitude of Signal to Noise Ratio in Terms of Average RMSE in Validation Set

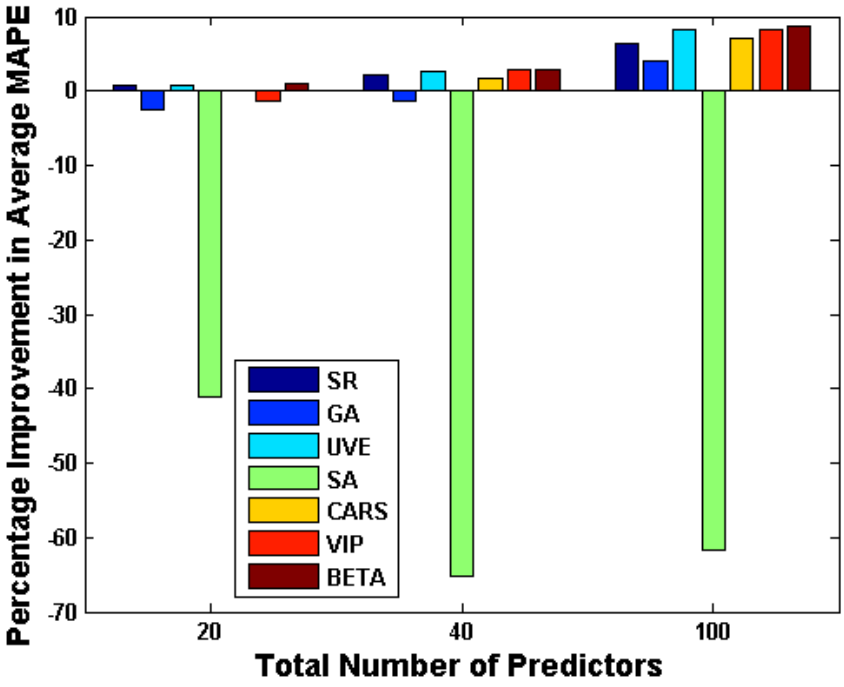

Figure 4.17 Sensitivity of Proportion of Relevant Predictors in Terms of Average MAPE in Validation Set

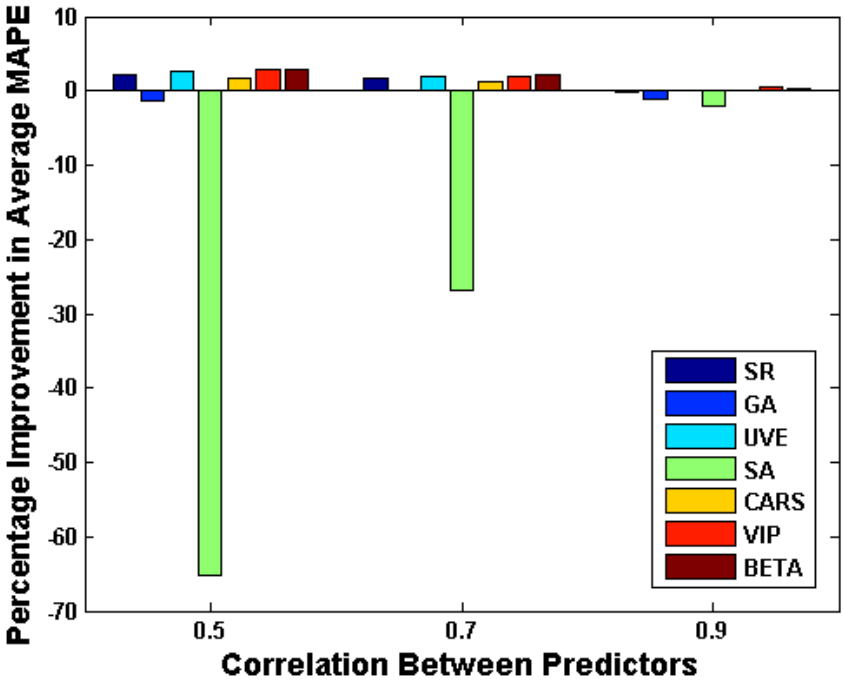

Figure 4.18 Sensitivity of Magnitude of Correlation between Predictors in Terms of Average MAPE in Validation Set

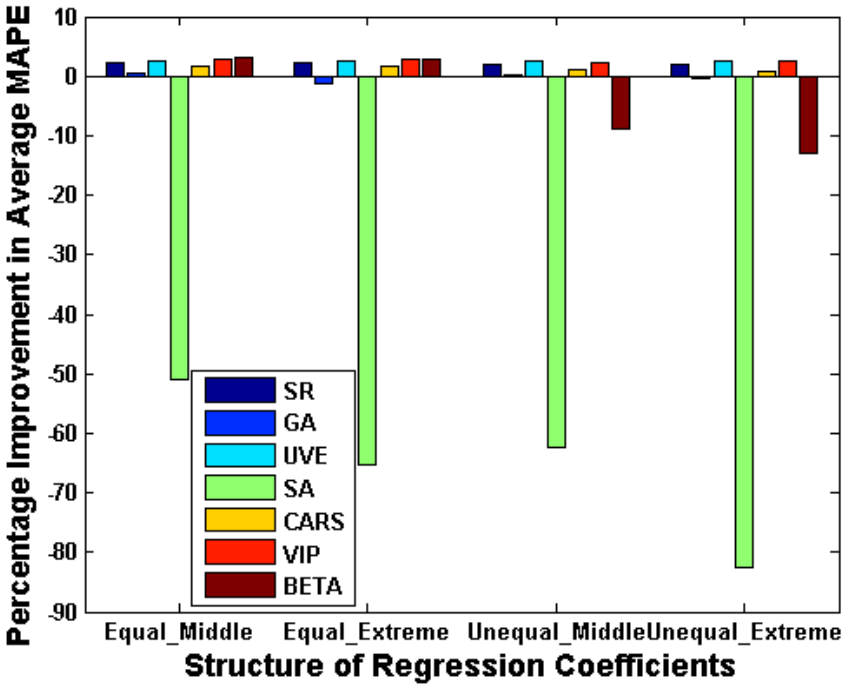

Figure 4.19 Sensitivity of Regression Coefficient Structure in Terms of Average MAPE in Validation Set

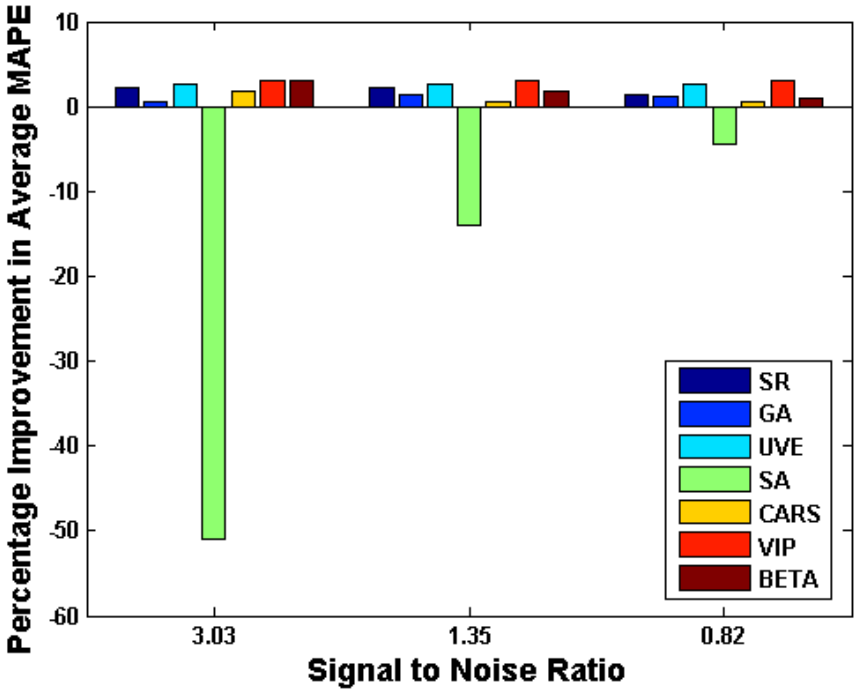

Figure 4.20 Sensitivity of Magnitude of Signal to Noise Ratio in Terms of Average MAPE in Validation Set

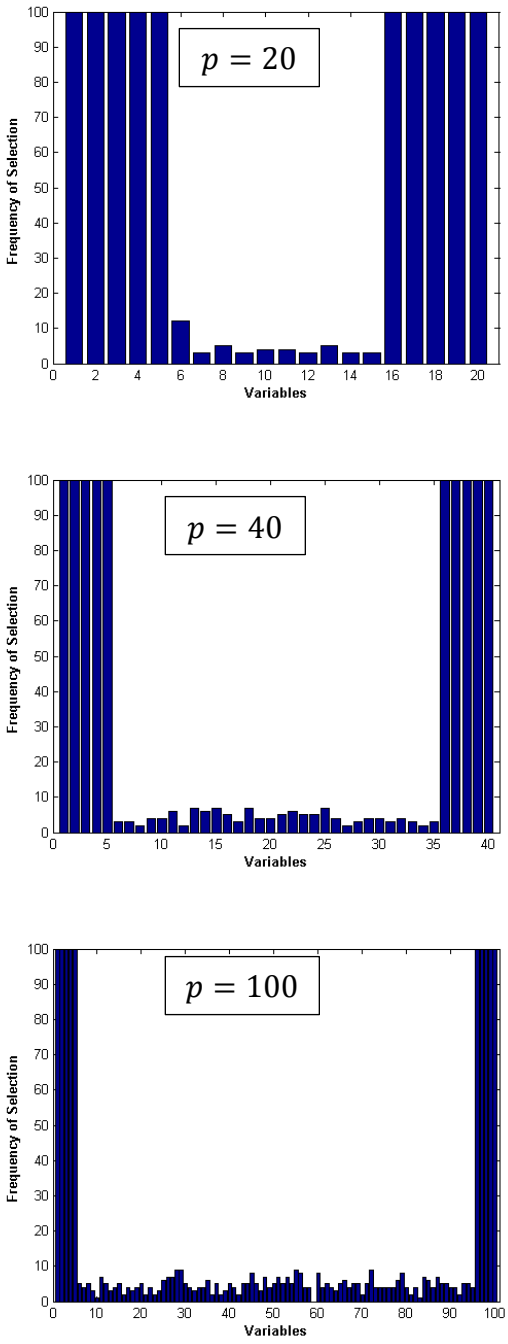

Figure 4.21 Frequency of Variables Selected by SR

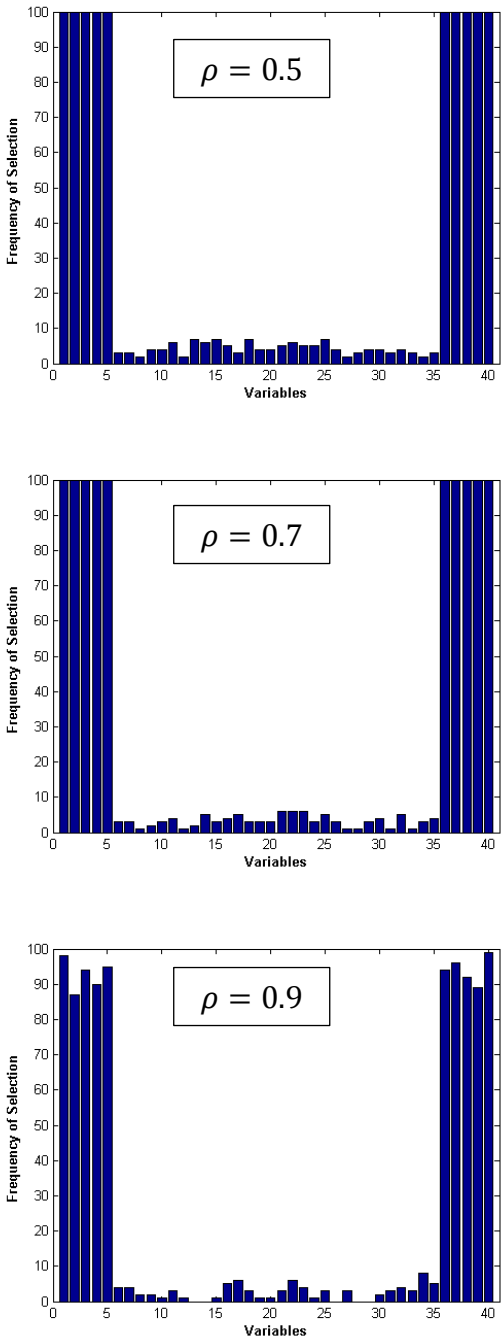

Figure 4.22 Frequency of Variables Selected by SR

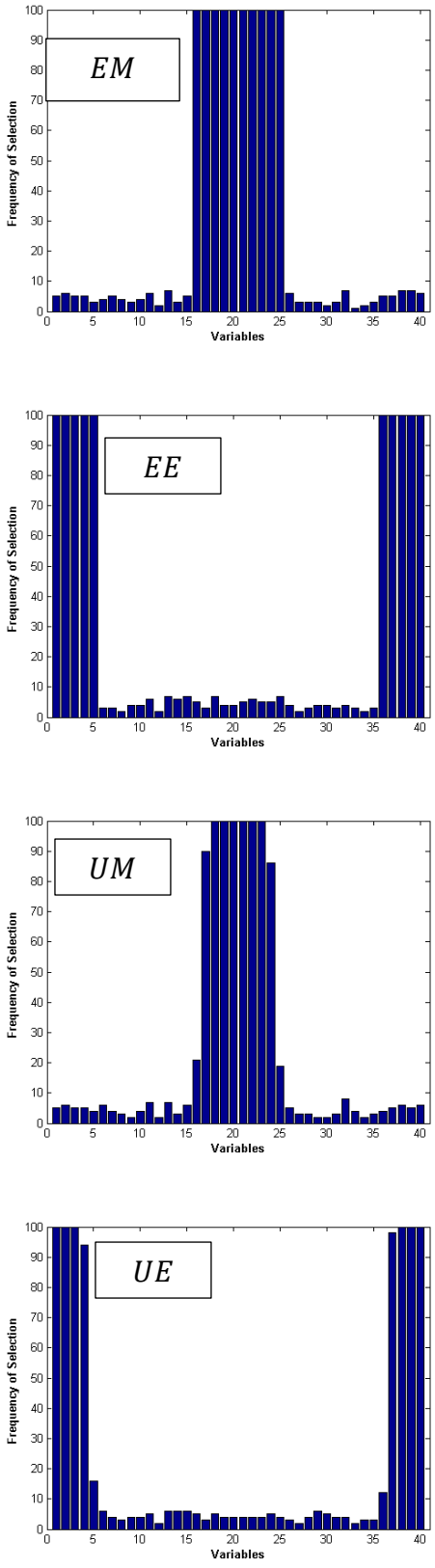

Figure 4.23 Frequency of Variables Selected by SR

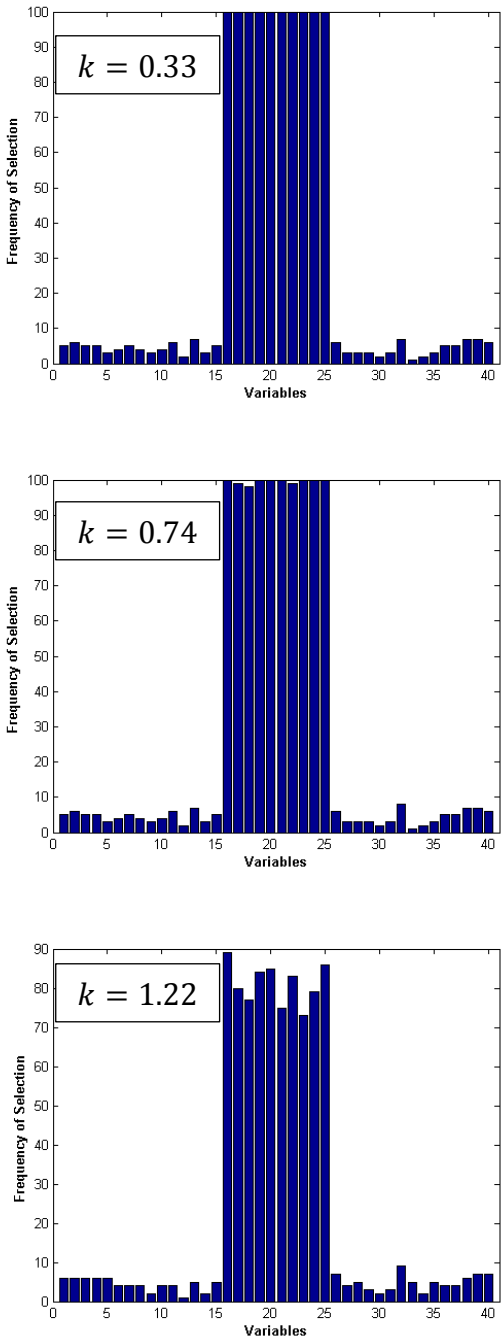

Figure 4.24 Frequency of Variables Selected by SR

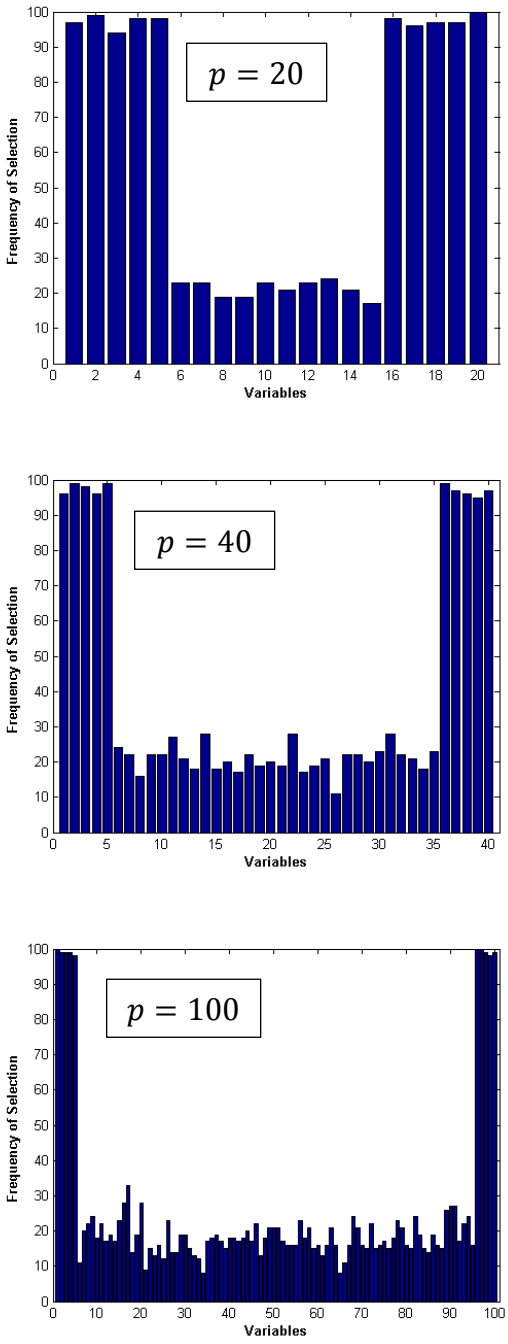

Figure 4.25 Frequency of Variables Selected by GA-PLS

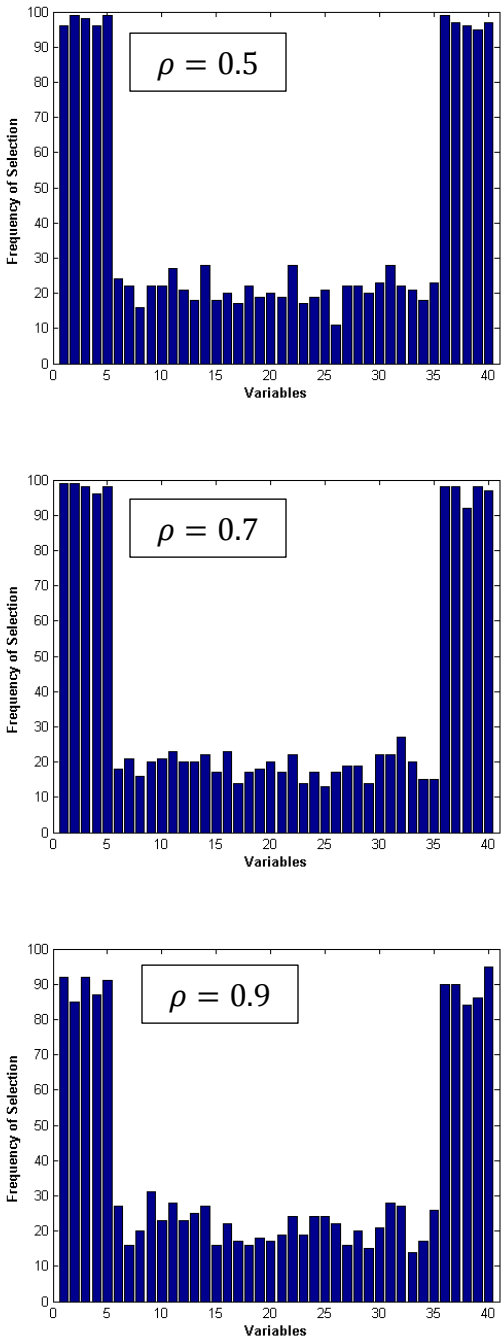

Figure 4.26 Frequency of Variables Selected by GA-PLS

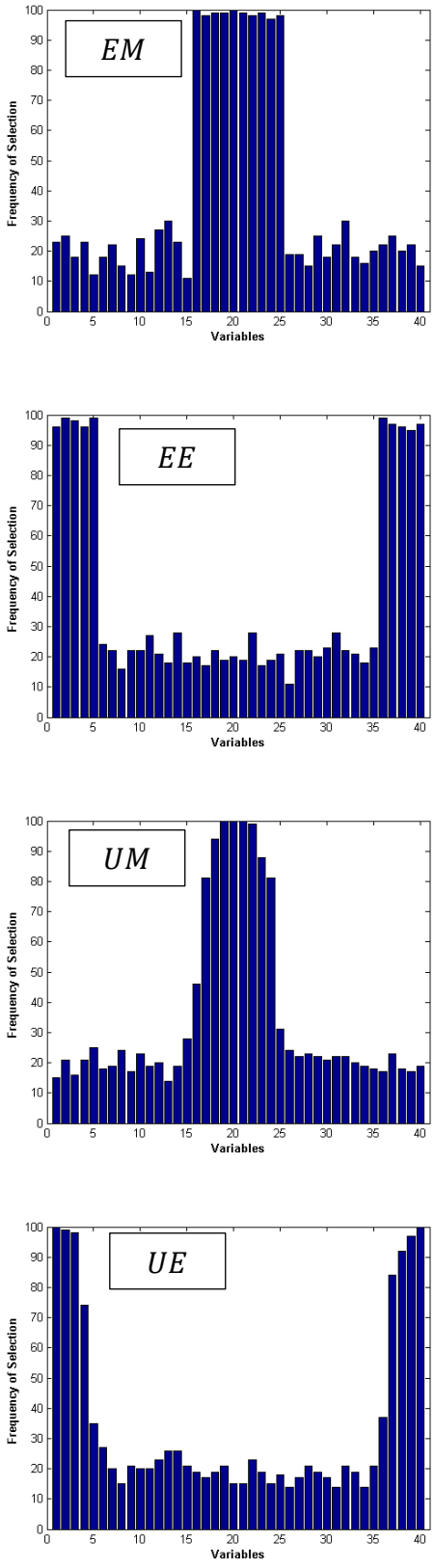

Figure 4.27 Frequency of Variables Selected by GA-PLS

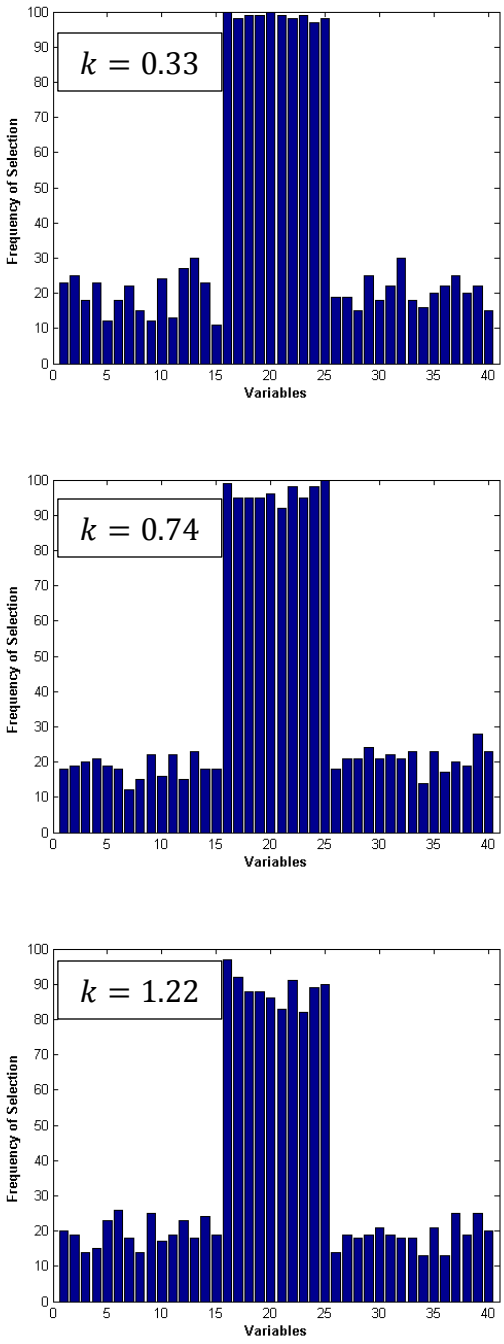

Figure 4.28 Frequency of Variables Selected by GA-PLS

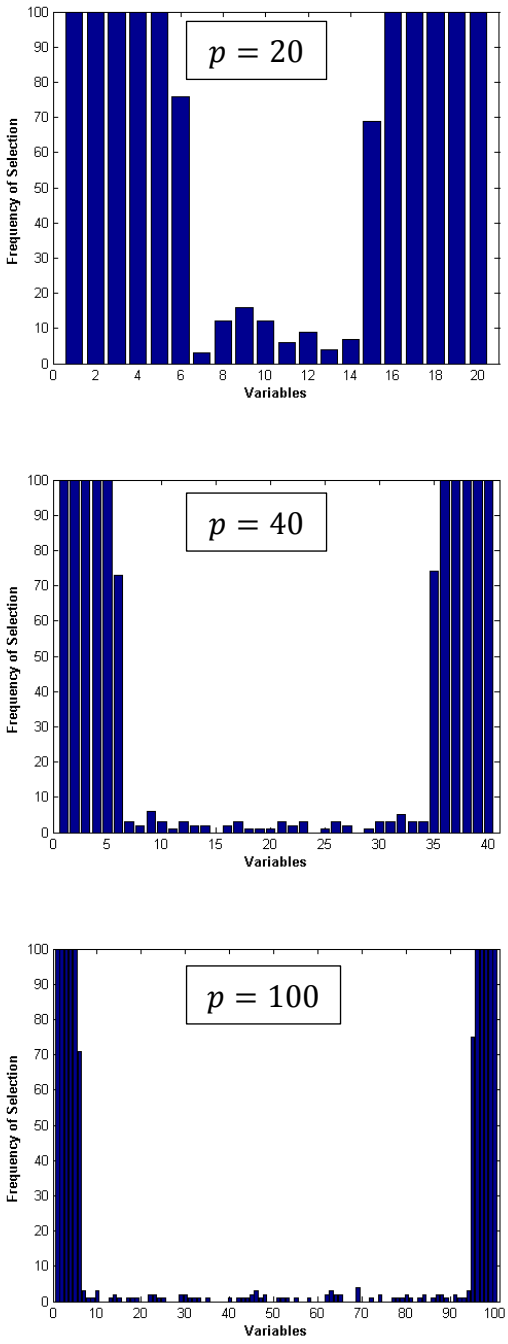

Figure 4.29 Frequency of Variables Selected by UVE-PLS

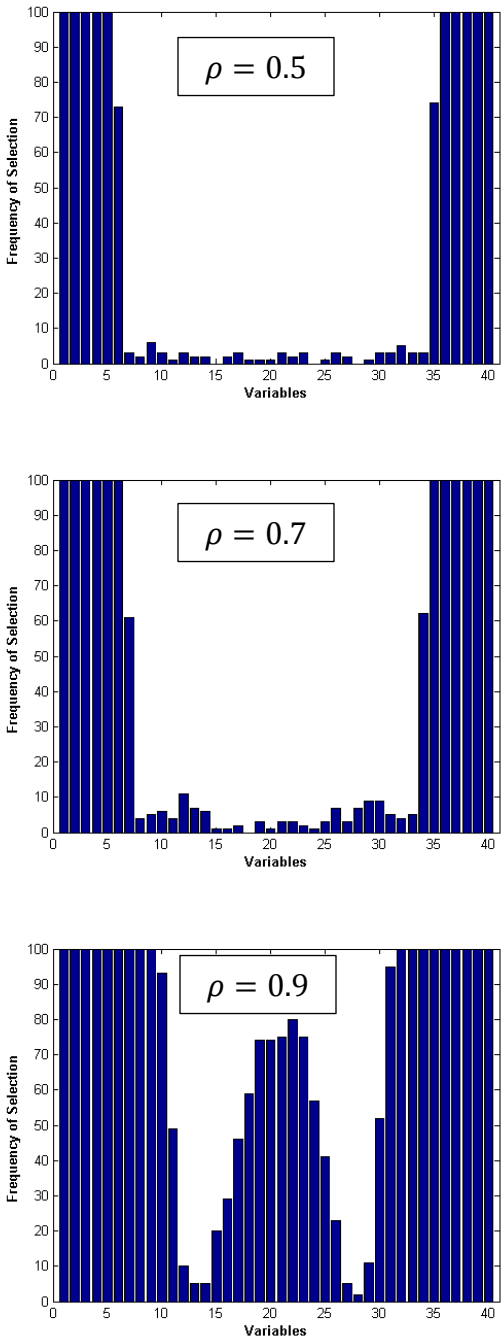

Figure 4.30 Frequency of Variables Selected by UVE-PLS

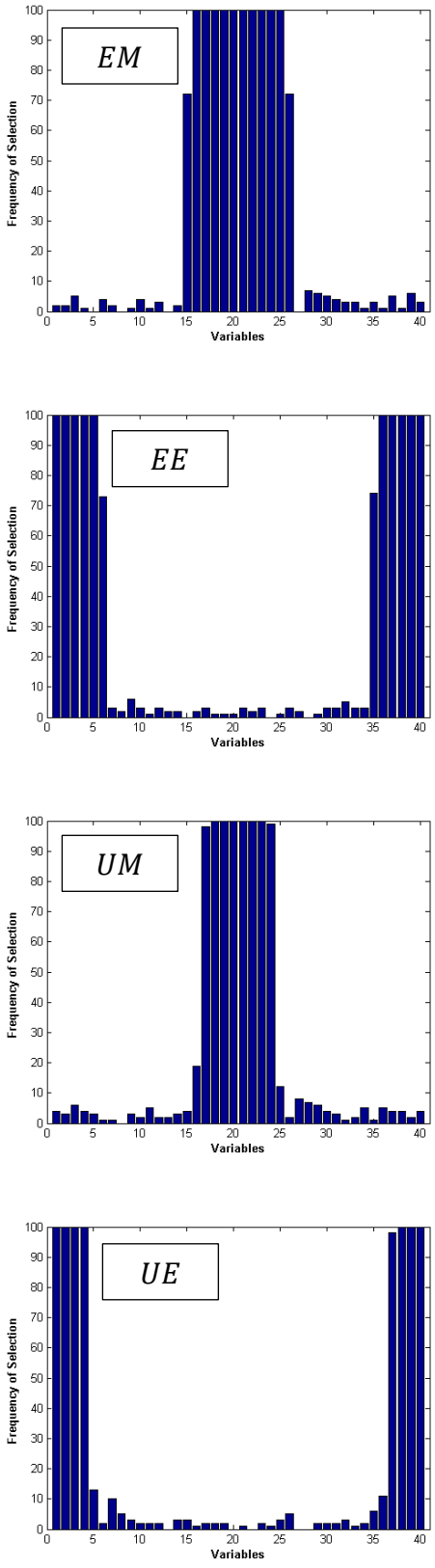

Figure 4.31 Frequency of Variables Selected by UVE-PLS

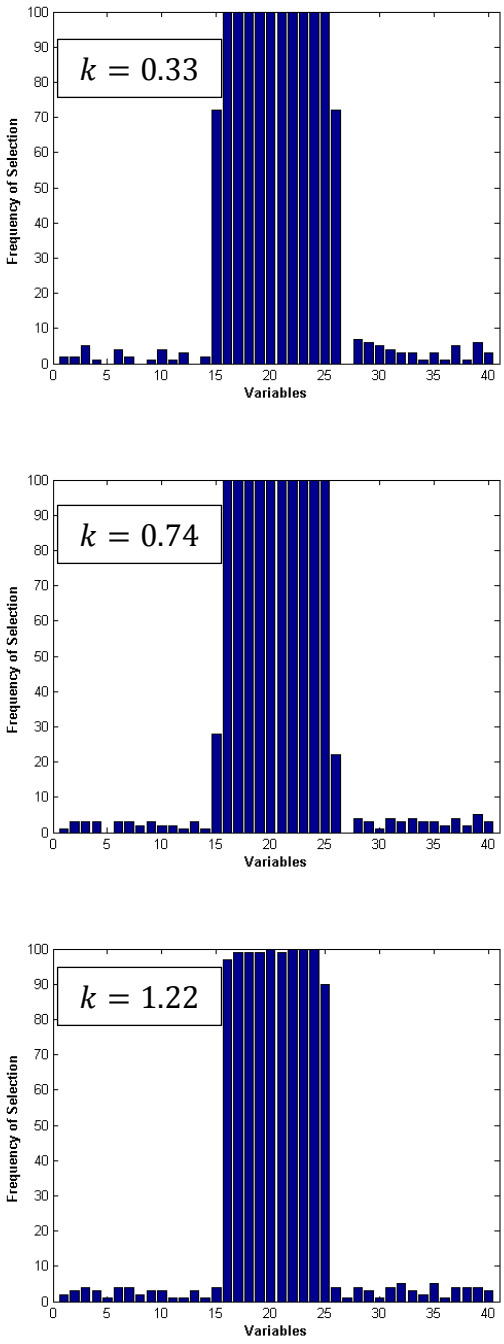

Figure 4.32 Frequency of Variables Selected by UVE-PLS

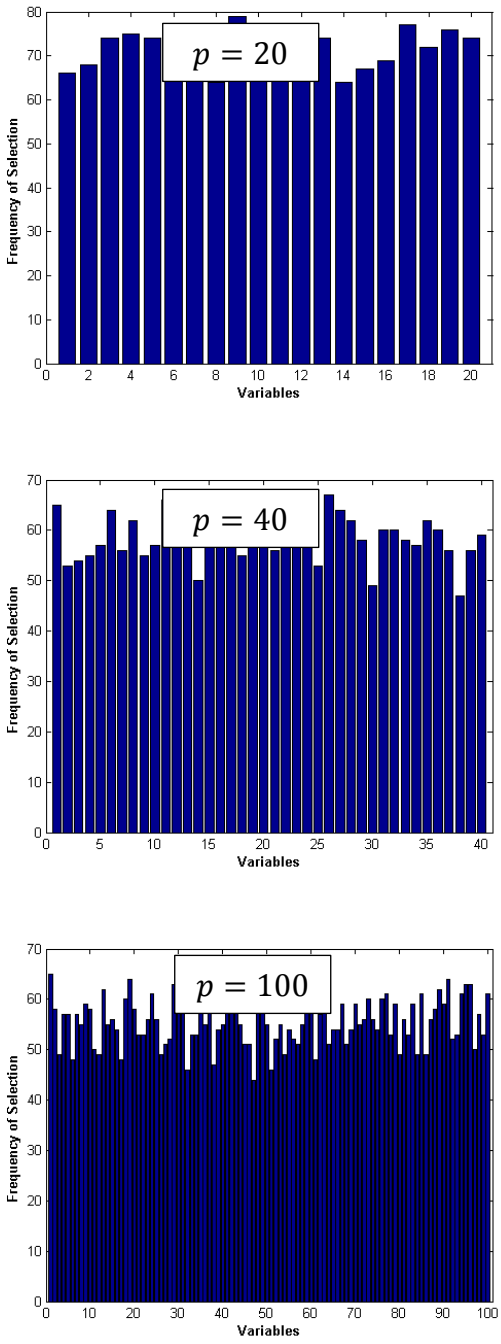

Figure 4.33 Frequency of Variables Selected by PLS-SA

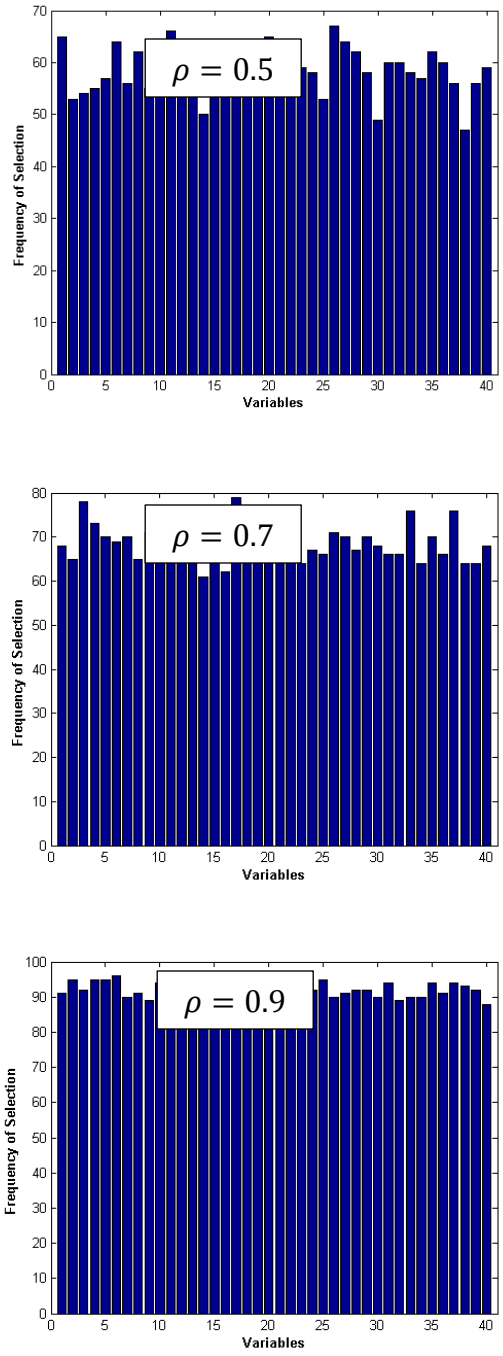

Figure 4.34 Frequency of Variables Selected by PLS-SA

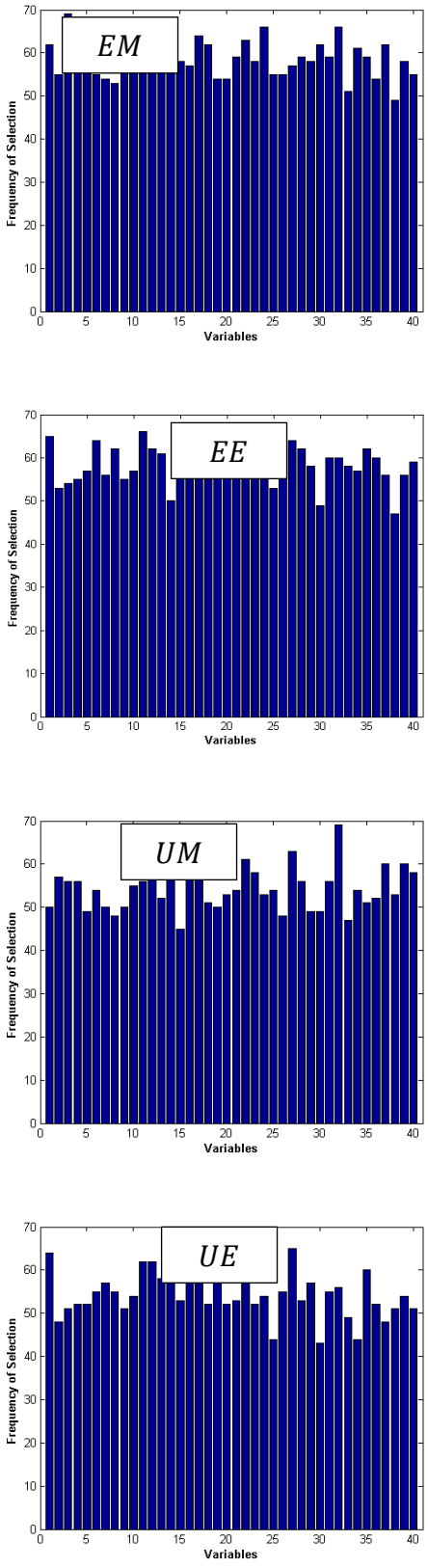

Figure 4.35 Frequency of Variables Selected by PLS-SA

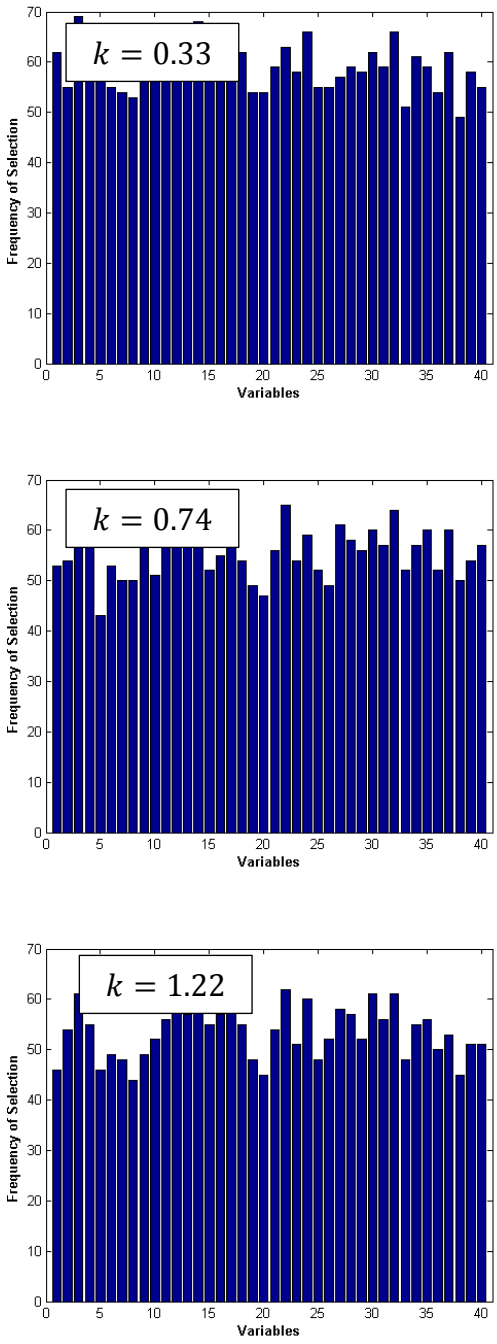

Figure 4.36 Frequency of Variables Selected by PLS-SA

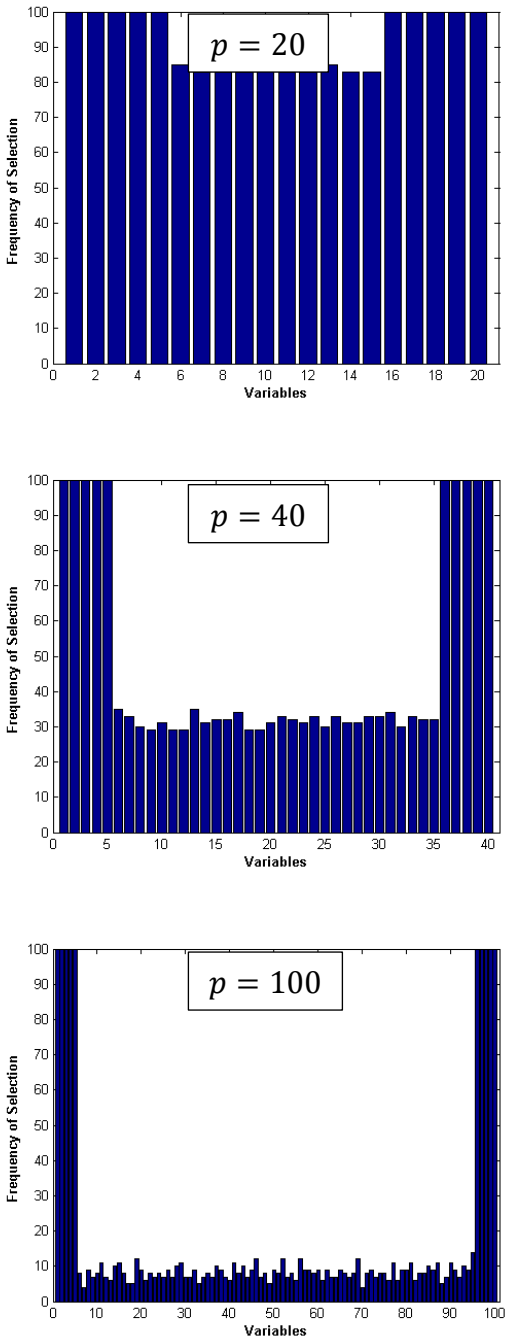

Figure 4.37 Frequency of Variables Selected by CARS-PLS

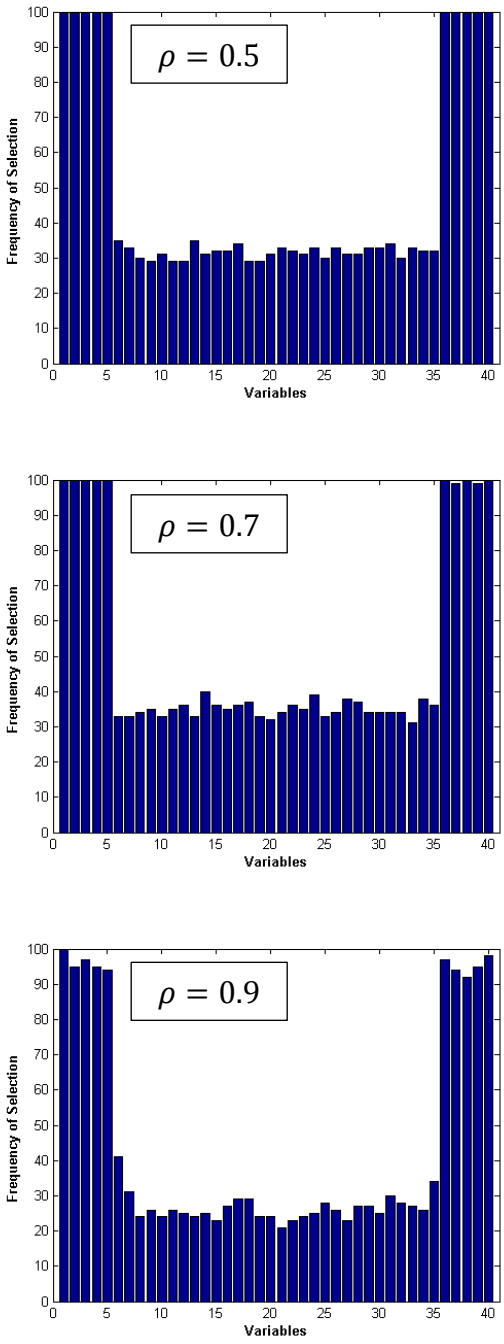

Figure 4.38 Frequency of Variables Selected by CARS-PLS

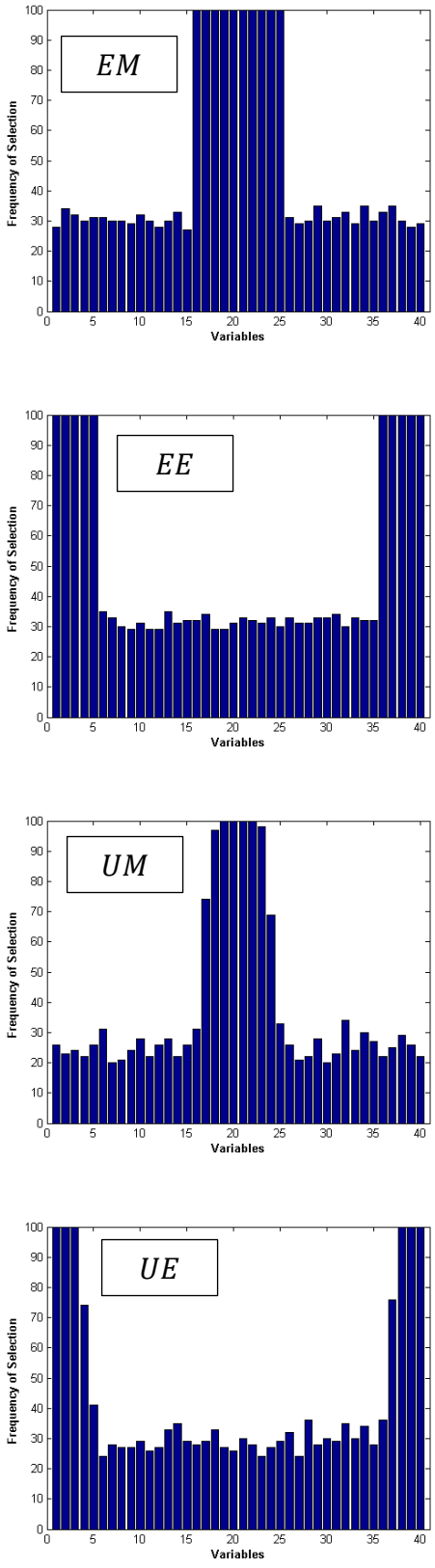

Figure 4.39 Frequency of Variables Selected by CARS-PLS

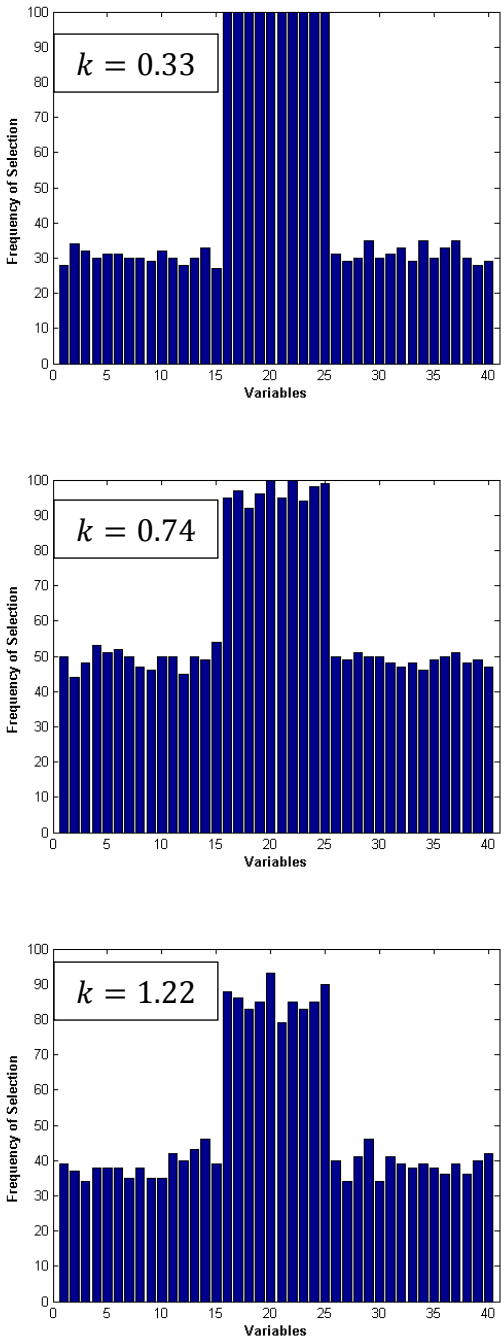

Figure 4.40 Frequency of Variables Selected by CARS-PLS

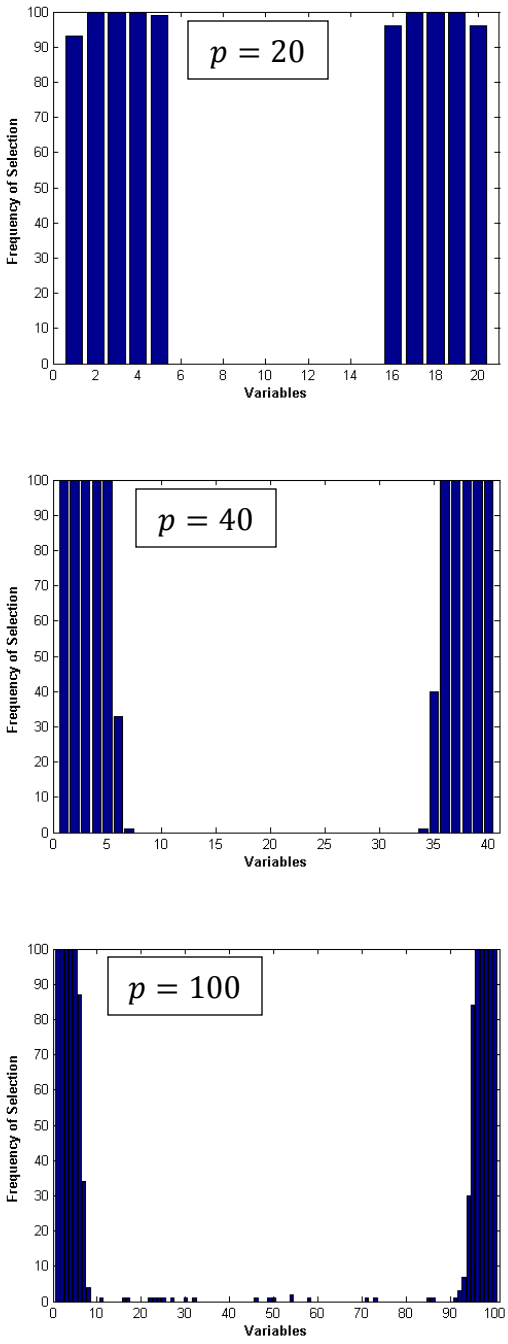

Figure 4.41 Frequency of Variables Selected by PLS-VIP

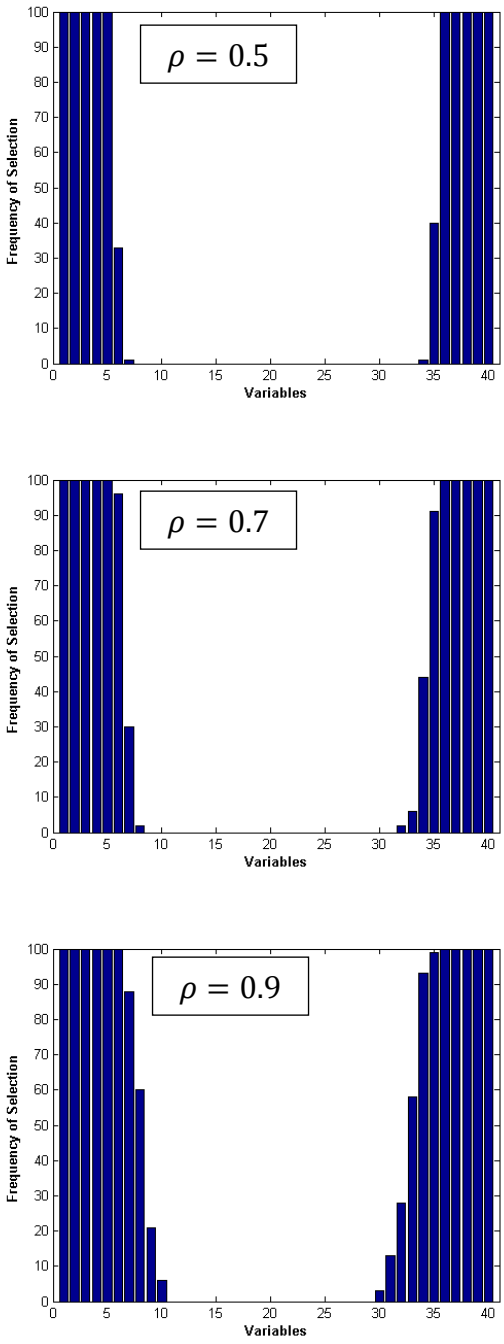

Figure 4.42 Frequency of Variables Selected by PLS-VIP

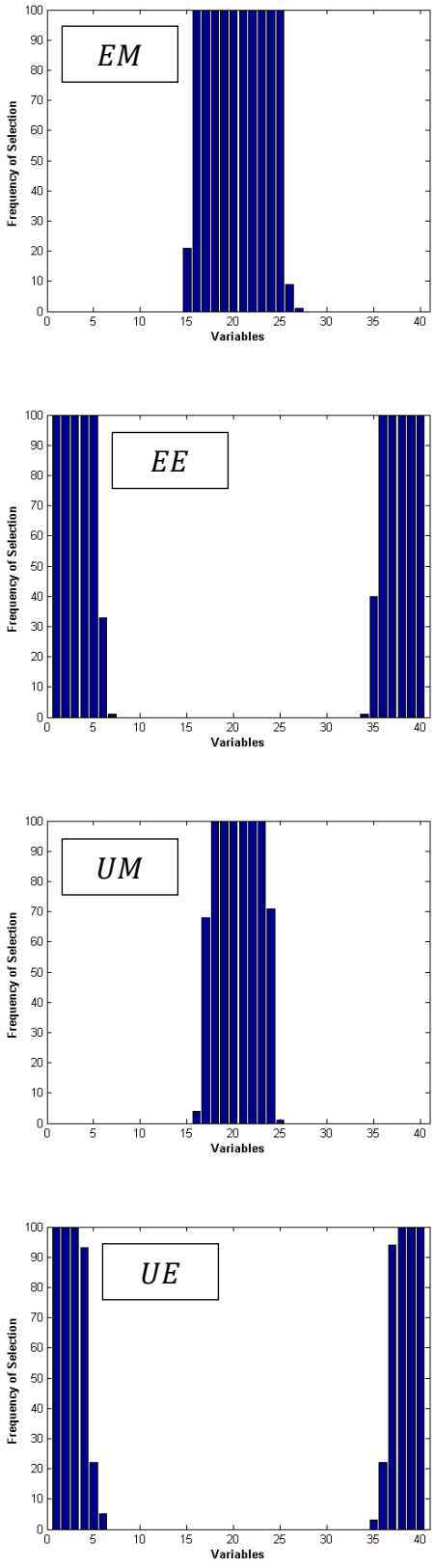

Figure 4.43 Frequency of Variables Selected by PLS-VIP

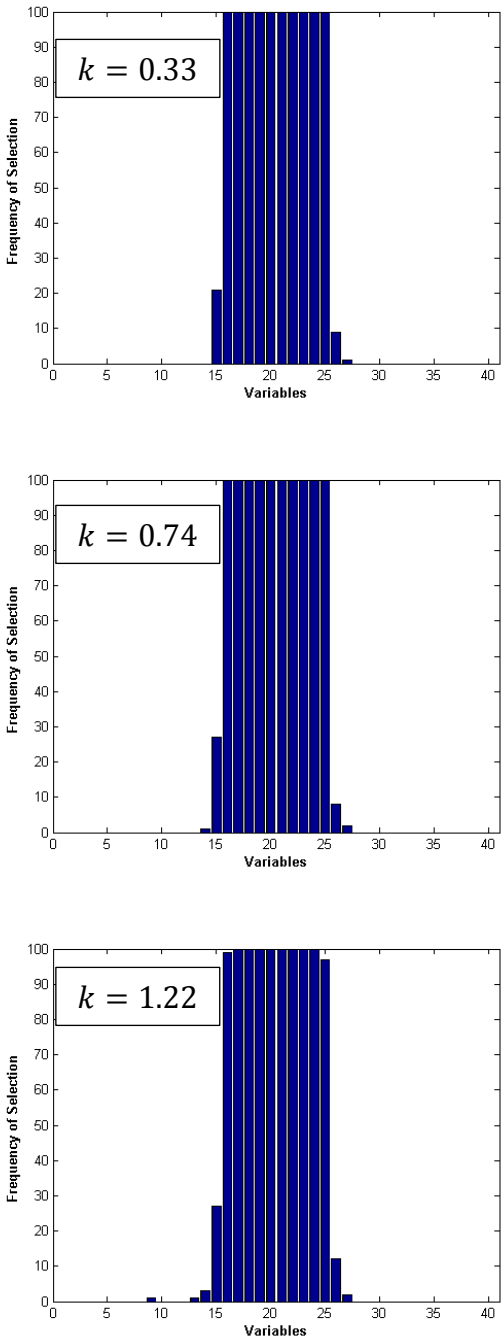

Figure 4.44 Frequency of Variables Selected by PLS-VIP

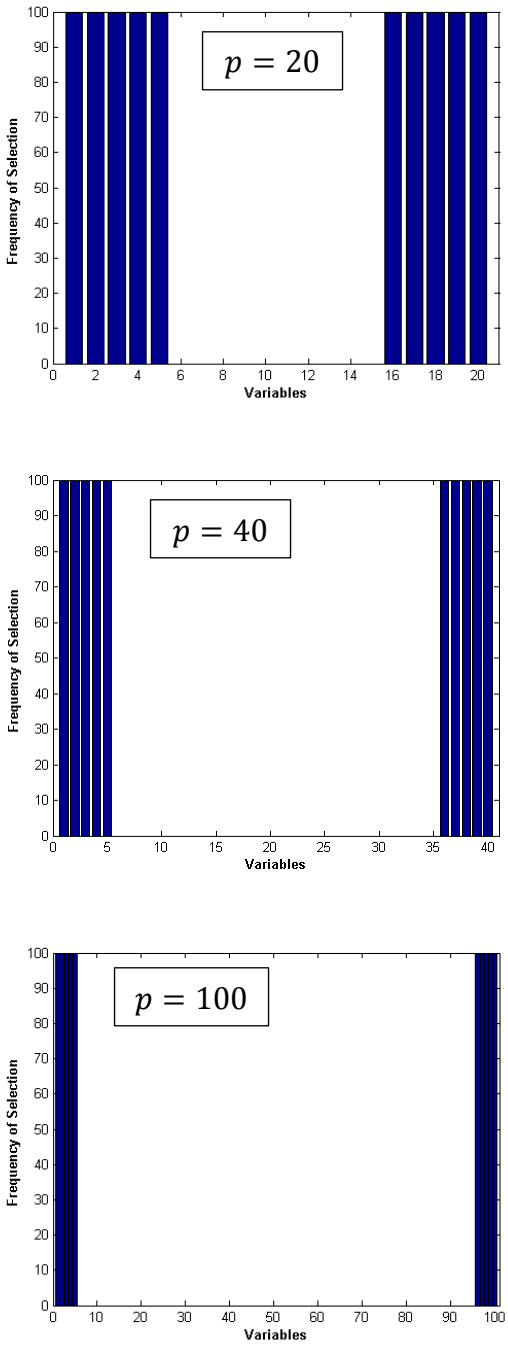

Figure 4.45 Frequency of Variables Selected by PLS-BETA

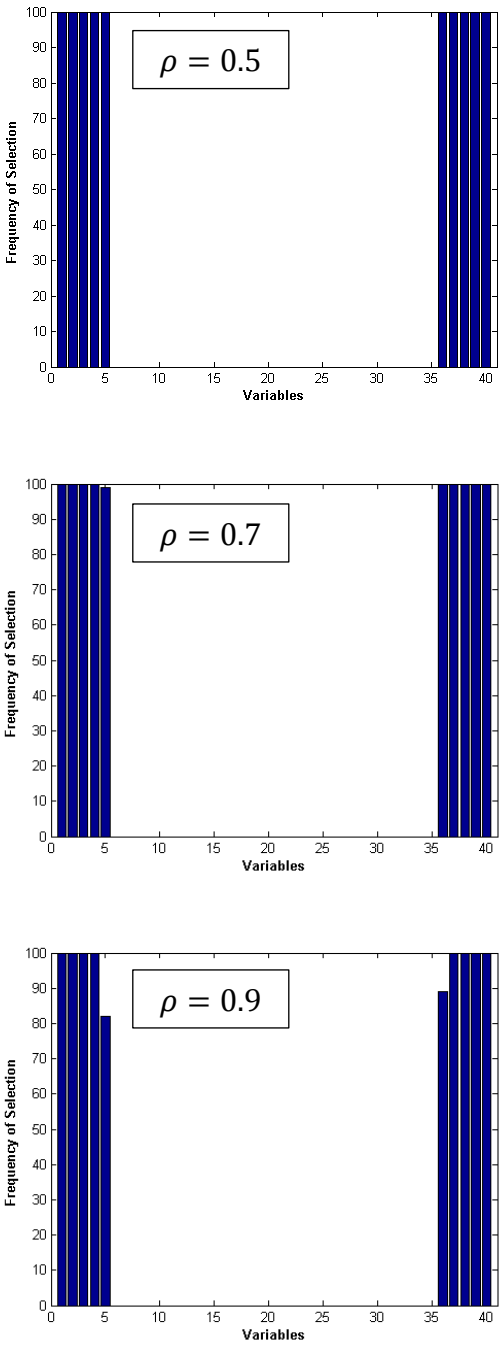

Figure 4.46 Frequency of Variables Selected by PLS-BETA

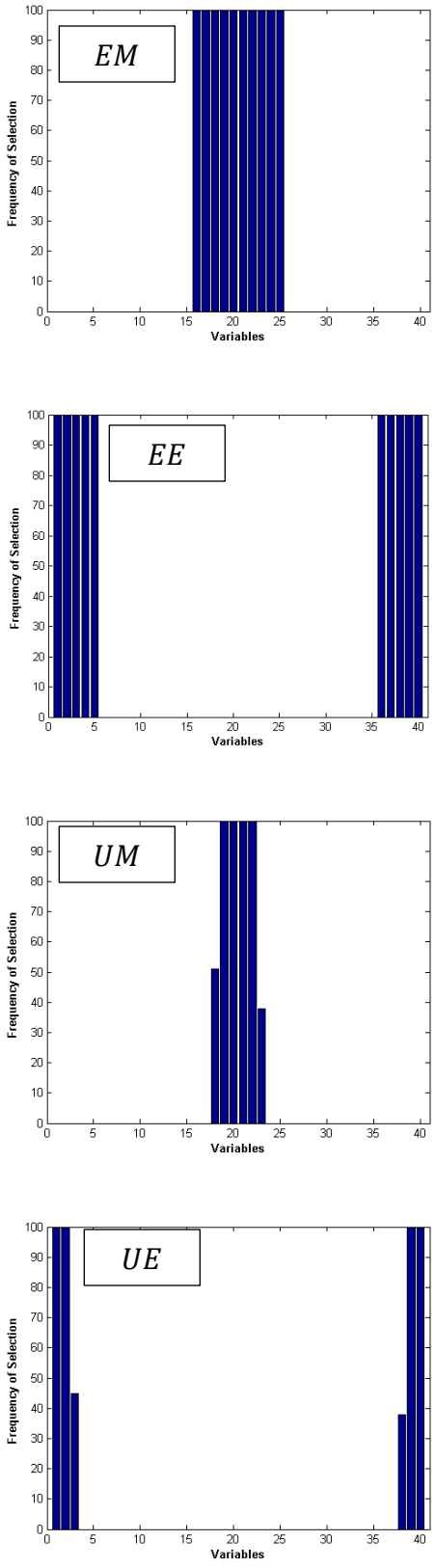

Figure 4.47 Frequency of Variables Selected by PLS-BETA

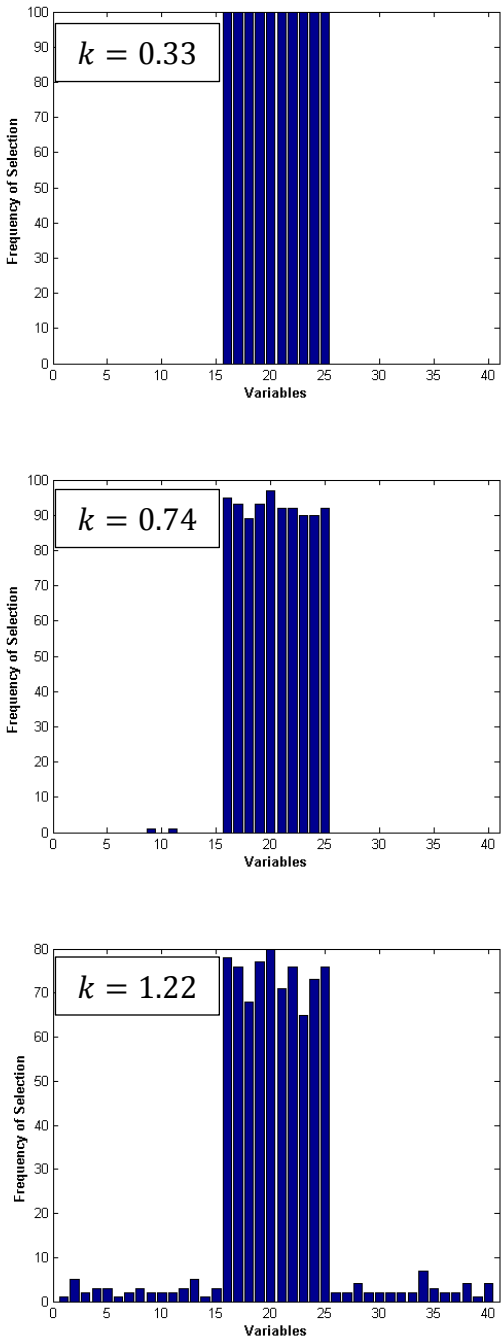

Figure 4.48 Frequency of Variables Selected by PLS-BETA

Another aspect of investigation is to examine the model's consistency, i.e., if a variable is selected once, will this variable be selected in next run? In order to check models' consistency, frequency of variable selection plots are generated and illustrated in Figure 4.21 to Figure 4.48. As shown in Figure 4.21 to Figure 4.24, SR is able to correctly identify most of the relevant predictors with 100% frequency. However, it also selects the irrelevant predictors at times. When the correlation between predictors increases, and signal to noise ratio decreases; SR no longer selects relevant predictors with 100% frequency. None of the relevant predictors are selected by GA-PLS with 100% frequency as shown in Figure 4.25 to Figure 4.28. Its performance worsens when the correlation between predictors and reciprocal of signal to noise ratio increase. Based on Figure 4.29 to Figure 4.32, UVE-PLS performs fairly well in the cases with unequal regression coefficients and low signal to noise ratio. However, UVE-PLS select some irrelevant predictors non-randomly when the correlation between predictors is at its higher level. From Figure 4.33 to Figure 4.36, one can see that PLS-SA selects both relevant and irrelevant predictors with same frequency in all cases. From Figure 4.37, CARS-PLS selects almost all the variables when the proportion of relevant predictors is high. Also, CARS-PLS selects irrelevant predictor with around 30-40% frequency in all other cases, which can be seen in Figure 4.37 to Figure 4.40. PLS-VIP and PLS-BETA are the ones with the most 'clean' frequency plots. In many cases, most of the irrelevant predictors are selected with 0% frequency. Only a few irrelevant predictors are selected with very low frequency in the case with low signal to noise ratio.

4.2.2 Conclusion and Discussion

Based on the results shown in simulated case study, PLS-VIP and PLS-BETA yield the best results among all seven variable selection methods studied. However, the performance of PLS-BETA does decay significantly when the regression coefficients are not equal and cover a wide range. PLS-BETA only selects the ones with large regression coefficients and discards the ones with smaller coefficients even though they are related to the primary variables. The prediction performance also deteriorates accordingly. In the opposite, UVE-PLS performs the best in the case with unequal regression coefficients. Performance of SR is in the middle range among all seven variable selection methods. The consistency of selection is quite good for SR. However, irrelevant predictors are selected by SR at lower frequency. Performance of GA-PLS is one of the ones in the middle range as well. Similar behavior is observed in GA-PLS as in SR. However, in terms of computational effort, SR requires less computation time. CARS-PLS is sensitive to proportion of relevant predictors. It selects all the irrelevant predictors with higher than 80% when the proportion of relevant predictors is high. CARS-PLS also selects irrelevant predictors with around 30-40% frequency in all other cases. Among all the variable selection methods considered in this work, PLS-SA yields the worst performance in most training and validation set. And the computation time of PLS-SA is quite intensively compare to other methods.

These conclusions are made purely based on the results obtained from this simulated case study.

4.3 Industrial Case Study

This industrial dataset was obtained from the request to Dr. Barolo. This process is the production of polyester resin used in the manufacturing of coatings via batch poly-condensation between a diol and a long-chain dicarboxylic acid [49]. The main part of this plant is a 12 m^3 stirred tank reactor, which is used for the production of different resins. Water is also formed in the poly-condensation reaction as a byproduct. A packed distillation column, along with an external water-cooled condenser and a scrubber, are installed to remove the water. In addition, a vacuum pump is equipped to maintain the vacuum in the reactor.

There are several online measurement sensors supplied in the plant. Thirty-four variables are routinely measured online and recorded by a process computer every 30 seconds. The number of samples is in the range between 4500 and 7500 from batch to batch. These are process measurements (temperature, pressures and valve openings, etc.) and controller settings (which are adjusted manually by the operators). A list of these thirty-four variables is shown in Table 4.5. Product quality measurements, acidity number and viscosity, are not measured online and are not available for the entire duration of the batch. The product samples are taken manually by the operators. The sampling is unevenly and infrequently. There are only 15 to 25 measurements available per batch. 33 batches are made available in 16-month period of time. The autoscaled process data of one of the batches is shown in Figure 4.49, and the quality variables are plotted in Figure 4.50. More details about this process can be found in [49], [50].

Table 4.5 List of Process Variables Included in Polyester Resin Dataset

Online Monitored Variable	Column	MA-PLS
Date and time of the day	1	
Mixing rate (%)	2	X
Mixing rate	3	X
Mixing rate SP	4	
Vacuum line temperature (°C)	5	X
Inlet dowtherm temperature (°C)	6	X
Outlet dowtherm temperature (°C)	7	X
Reactor temperature (sensor 1) (°C)	8	X
(dummy)	9	
Column head temperature (°C)	10	X
Valve V25 temperature (°C)	11	
Scrubber top temperature (°C)	12	X
Inlet water temperature (°C)	13	X
Column bottom temperature (°C)	14	X
Scrubber bottom temperature (°C)	15	X
Reactor temperature (sensor 2) (°C)	16	X
Condenser inlet temperature (°C)	17	X
Valve V14 temperature (°C)	18	X
Valve V15 temperature (°C)	19	X
Reactor differential pressure	20	X
(dummy)	21	
Column top temperature PV (°C)	22	X
Column top temperature SP (°C)	23	
V42 way-1 valve opening (%)	24	X
Inlet dowtherm temperature PV (°C)	25	X
Inlet dowtherm temperature SP (°C)	26	
V42 way-2 valve opening (%)	27	X
Reactor temperature PV(°C)	28	X
Reactor temperature SP (°C)	29	
(dummy)	30	
Valve V25 temperature PV (°C)	31	
Valve V25 temperature SP (°C)	32	
Valve V42 valve opening (%)	33	X
Reactor vacuum PV (mbar)	34	X
Reactor vacuum SP (mbar)	35	

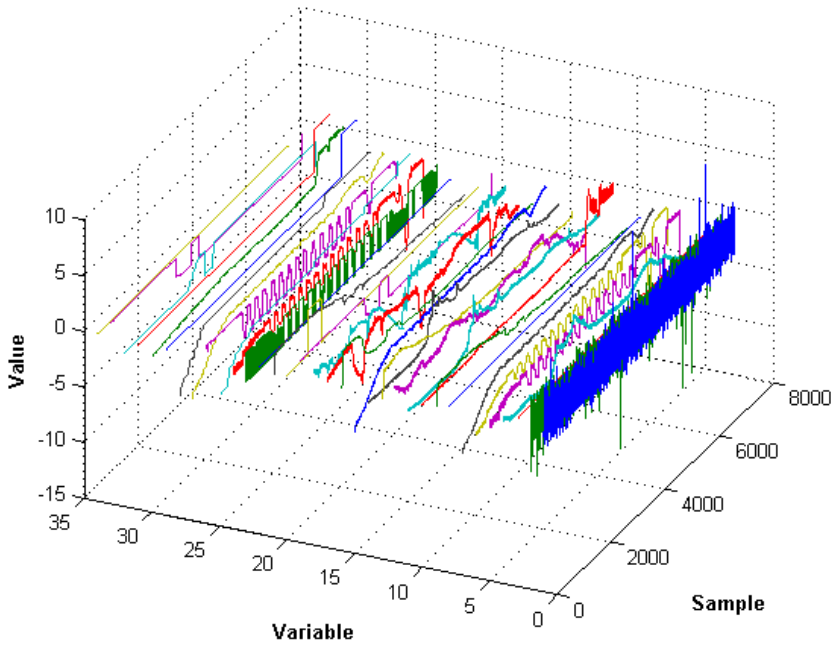

Figure 4.49 Visualization of Autoscaled Process Data from A Reference Batch in Polyester Production

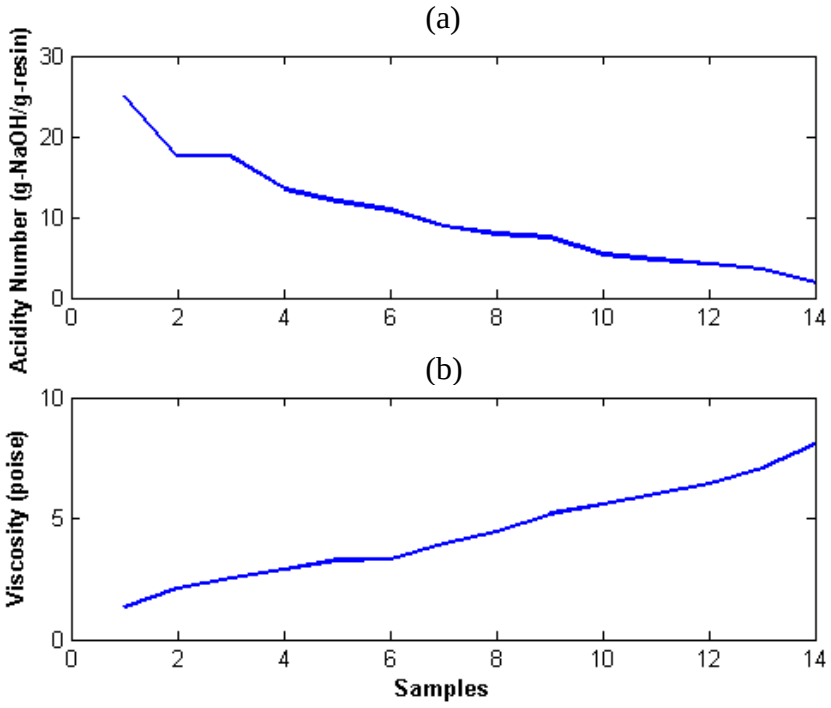

Figure 4.50 Product Quality Variables from A Reference Batch in Polyester Production. (a) is the acidity number in g_{NaOH}/g_{resin}; (b) is the viscosity in poise.

4.3.1 Data Preprocessing

For a batch process, the data are stored in a three-dimension array, $K \times J \times I_k$, as shown in Figure 4.51. Each row corresponds to one of the K batches, while each column contains one of the J variables; $I_k$is the total number of samples taken in k^{th} batch. This is one of the typical characteristics of batch process, where batch duration is not fixed. Thus, a preprocessing step is required to synchronize the batch-to-batch durations. Three preprocessing methods are considered in this work:

1. Only retain the process samples when the quality variables are available. All the process samples without their corresponding quality variables are eliminated.
2. Instead of eliminate all those samples points, the average of them are taken and utilized as soft sensor inputs.
3. Similar to the previous one, but integral over time is taken as the soft sensor inputs.

The approach taken to unfold the three-way array is to preserve the direction of the variables [51]. The resulting matrix has dimension of n by J, where $n = \sum_{k=1}^{K} I_k$. A previous approach proposed by Nomikos and MacGregor [52], [53] is to unfold the three-way matrix so that the batch direction is preserved. This results in matrix with dimension of K by $(\sum_{k=1}^{K} J \times I_k)$. Since variable selection is our purpose, the approach that preserves the direction of variables is adopted.

To provide fair comparison, the batches are permuted 100 times before unfolding to generate different combinations of training and validation set. In every permutation, the first 27 batches are used for training, and the remaining is used for validation. The

visualization of the autoscaled process data using the first preprocessing method and the quality data from one of the permutation runs are shown in Figure 4.52 and Figure 4.53, respectively.

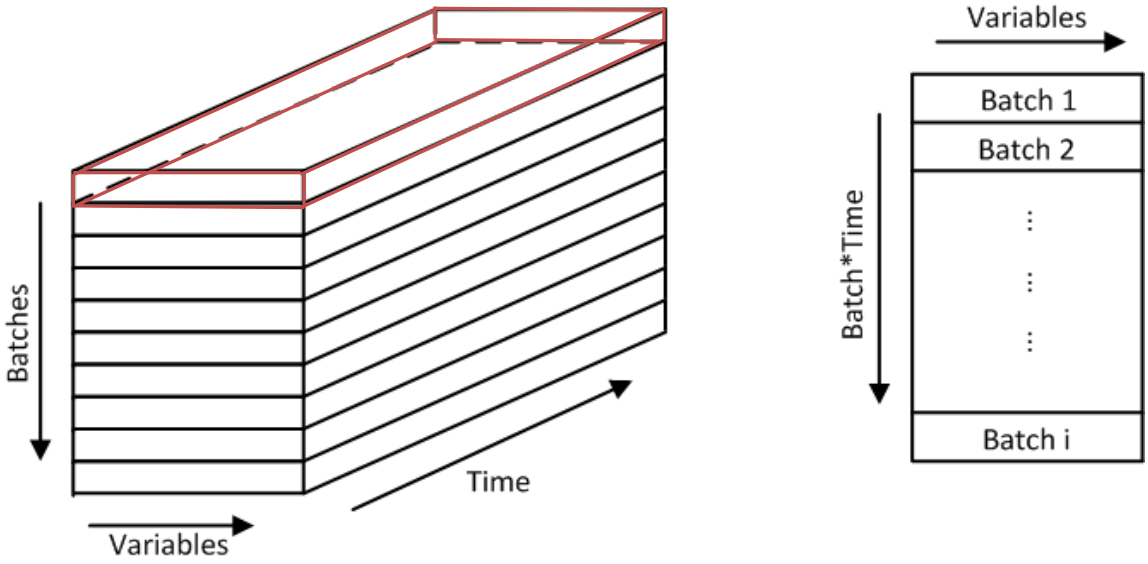

Figure 4.51 Illustration of Unfolding Three-Dimension Array to Preserve the Direction of Variables

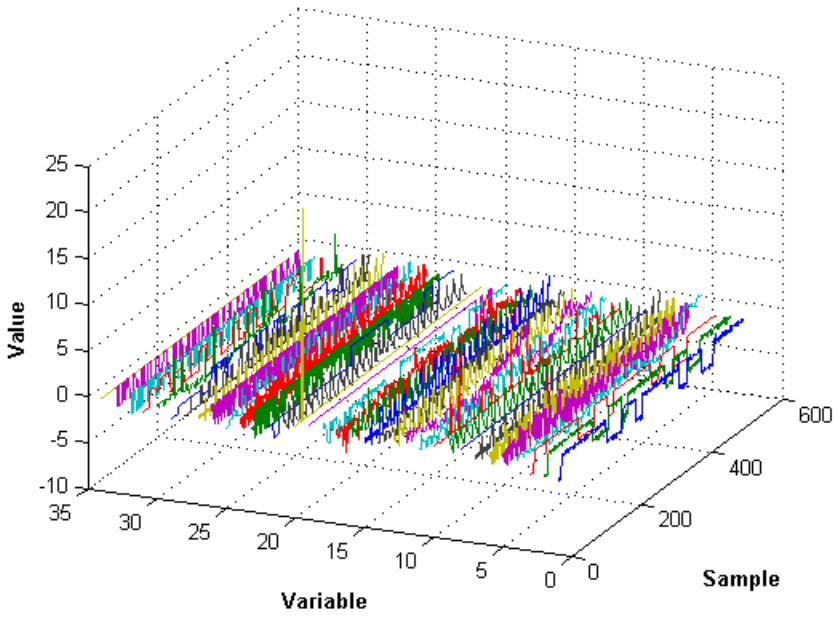

Figure 4.52 Dynamic Parallel Coordinate Plot of Autoscaled Unfolded Process Data of A Permutation Run of Polyester Resin Dataset

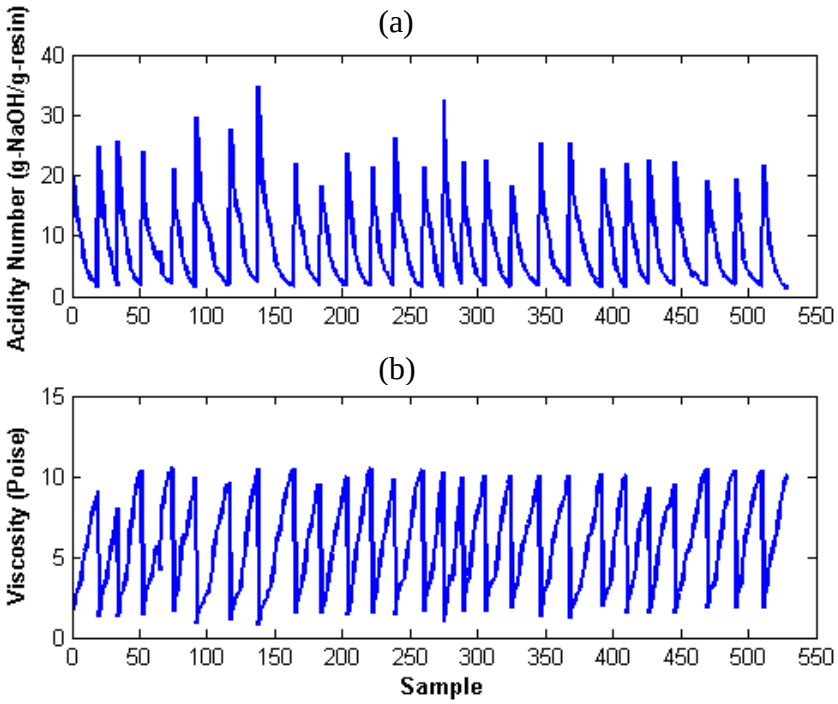

Figure 4.53 Product Quality Variables of A Permutation Run of Polyester Production. (a) is the acidity number in g_{NaOH}/g_{resin}; (b) is the viscosity in poise.

4.3.2 Results

Variable selection methods are applied individually to each quality variables. Only the training set is utilized for variable selection. Models developed by these variable selection schemes are then validated using the validation set. The results of three different preprocessing methods are presented in Table 4.6 and Table 4.8.

Based on the results obtained from the first preprocessing method, all the variable selection methods is able to identify a subset of variables that would improve the model prediction performance, with the exception of PLS-SA. For acidity number model, the one with the best prediction performance is produced by PLS-VIP. Over 100 permutation runs, model with average number of 13 variables are created by PLS-VIP. The prediction performances on the external validation set are improved by 23.2% and 27.1% in RMSE and MAPE, respectively. For the calibration models, the best results are actually given by

SR with an average size of 12 variables. Nevertheless, the results on the validation set are not as great as PLS-VIP. This indicates that the model produced by SR may tend to over-fit the training data. For viscosity model, the best performance model is again produced by PLS-VIP with model size of 14 variables in average. The prediction performances on the external validation set are improved by 28.1% and 23.3% in RMSE and MAPE, respectively. In addition, this is the only model that gives such superior performance; all the other methods only improve the prediction performance up to 6%.

Results of acidity number and viscosity models from the second preprocessing method show that the most superior prediction performance is still given by PLS-VIP. The model size is increased from 13 to 16 variables for acidity number model and 14 to 16 for viscosity model. The prediction performances of acidity number model are boosted by 26% and 13% in RMSE and MAPE, respectively. For viscosity model, the prediction performances are advanced by 36.5 % and 29.2% in RMSE and MAPE, respectively. SR also produces best calibration models for both acidity number and viscosity in this preprocessing method.

Once again, best prediction performance of acidity number model is provided by PLS-VIP in the third preprocessing method, by 10% in RMSE and 8% in MAPE. The average model size is only 9 variables. The results of the viscosity model are different from the previous two methods. The highest prediction performance is actually provided by CARS-PLS with improvement of 13% in both RMSE and MAPE. The model is relatively small with 9 variables. The standard deviation of model size is almost half of the average value, which means CARS-PLS is sensitive to data selection. Furthermore, the prediction errors of the last method are almost doubled compared to the previous ones.

Table 4.6 Comparison of Different Variable Selection for Preprocessing Method 1

	Model	No. of Variables	Training		Validation	
			RMSE	MAPE	RMSE	MAPE
Acidity Number	PLS	34	1.7031	24.4044	2.2175	32.9926
	Stepwise	12+/-2	**1.6166**	**22.1661**	1.8587	26.4972
	GA-PLS	13+/-2	1.6327	23.0392	1.8183	26.6951
	UVE-PLS	18+/-2	1.6795	23.5050	2.0449	29.4195
	PLS-SA	27+/-2	1.7752	25.1340	2.2777	33.8493
	CARS-PLS	8+/-2	1.7402	23.9638	1.8849	27.2733
	PLS-VIP	13+/-1	1.6653	22.5347	**1.7021**	**24.0513**
	PLS-BETA	16+/-1	1.6737	22.7862	1.9947	28.0227
Viscosity	PLS	34	0.6885	11.7796	1.0662	17.5609
	Stepwise	11+/-2	0.6869	12.0458	1.0068	17.0644
	GA-PLS	11+/-3	0.6936	**11.6718**	0.9951	16.5112
	UVE-PLS	18+/-3	0.7475	13.2438	1.0368	17.4080
	PLS-SA	29+/-2	0.7084	12.1746	1.0951	17.9966
	CARSPLS	9+-/4	0.7270	12.1278	1.0234	16.6214
	PLS-VIP	14+/-1	0.7344	13.0173	**0.7661**	**13.4600**
	PLS-BETA	16+/-2	**0.6884**	11.8257	1.0341	17.1061

Table 4.7 Comparison of Different Variable Selection for Preprocessing Method 2

	Model	No. of Variables	Training		Validation	
			RMSE	MAPE	RMSE	MAPE
Acidity Number	PLS	34	1.4498	22.8617	2.0488	29.1595
	Stepwise	11+/-2	**1.4200**	**22.5903**	1.6558	25.9971
	GA-PLS	11+/-2	1.4292	23.1971	1.5874	26.4053
	UVE-PLS	20+/-1	1.4604	23.4277	1.8858	29.4339
	PLS-SA	28+/-2	1.5008	23.7984	2.0197	29.8083
	CARS-PLS	9+/-3	1.4511	23.7552	1.5500	25.6226
	PLS-VIP	16+/-0	1.4605	23.8432	**1.5245**	**25.3108**
	PLS-BETA	17+/-1	1.4406	22.6302	1.7180	26.9167
Viscosity	PLS	34	0.7079	12.6513	1.1777	19.3988
	Stepwise	11+/-2	**0.6902**	**12.0718**	1.0320	17.5915
	GA-PLS	10+/-2	0.6965	12.2571	1.0391	17.5482
	UVE-PLS	21+/-3	0.7196	13.0370	1.1227	18.7715
	PLS-SA	28+/-2	0.7539	13.7273	1.2775	21.1849
	CARSPLS	9+-/5	0.7413	13.4055	1.0976	18.7016
	PLS-VIP	16+/-1	0.7228	13.2627	**0.7476**	**13.7396**
	PLS-BETA	15+/-2	0.6950	12.4778	1.0176	17.6031

Table 4.8 Comparison of Different Variable Selection for Preprocessing Method 3

	Model	No. of Variables	Training		Validation	
			RMSE	MAPE	RMSE	MAPE
Acidity Number	PLS	34	3.0584	53.1167	3.7931	57.7093
	Stepwise	14+/-2	3.0806	53.1855	4.0081	58.4953
	GA-PLS	11+/-2	2.9637	49.3168	4.0584	56.3860
	UVE-PLS	28+/-1	3.1057	52.4875	4.5632	62.0594
	PLS-SA	32+/-1	3.0722	53.4689	3.7646	57.7600
	CARS-PLS	12+/-7	3.0216	**49.0159**	3.7139	53.9567
	PLS-VIP	9+/-1	3.0576	50.6028	**3.4143**	**52.9578**
	PLS-BETA	14+/-1	**2.9966**	51.0542	3.8230	56.2184
Viscosity	PLS	34	1.7437	32.7257	2.2371	43.3408
	Stepwise	10+/-2	1.6001	31.0598	2.4125	45.0485
	GA-PLS	11+/-2	1.5842	30.7862	2.3106	43.2504
	UVE-PLS	27+/-2	1.6804	31.7349	2.3896	44.1047
	PLS-SA	29+/-5	1.7580	32.9440	2.1602	41.7313
	CARSPLS	9+-/5	**1.5539**	**30.3764**	**1.9422**	**37.7724**
	PLS-VIP	14+/-1	1.7392	32.5318	1.9951	39.9705
	PLS-BETA	17+/-1	1.6217	31.0373	2.1733	41.9445

Based on the results summarized in the above Tables and Figure 4.54 to Figure 4.57, among three preprocessing methods, the third method yields the largest prediction error. Also, the improvement by variable selection for the third method is least significant. The results from the first two methods are comparable. In acidity number model, the second preprocessing method yields the best performance, while the first preprocessing method performs the best in viscosity model. The sizes of models produced by the second preprocessing method are slightly larger the first method. However, more significant improvement in prediction performance is observed in the second preprocessing method. The computation time is also equivalent. The results are also illustrated in Figure 4.58 to Figure 4.65. The performance indicators shown are compared with the full model from each preprocessing method, with positive values implying improvement in prediction performance and negative values implying deteriorations in prediction performance.

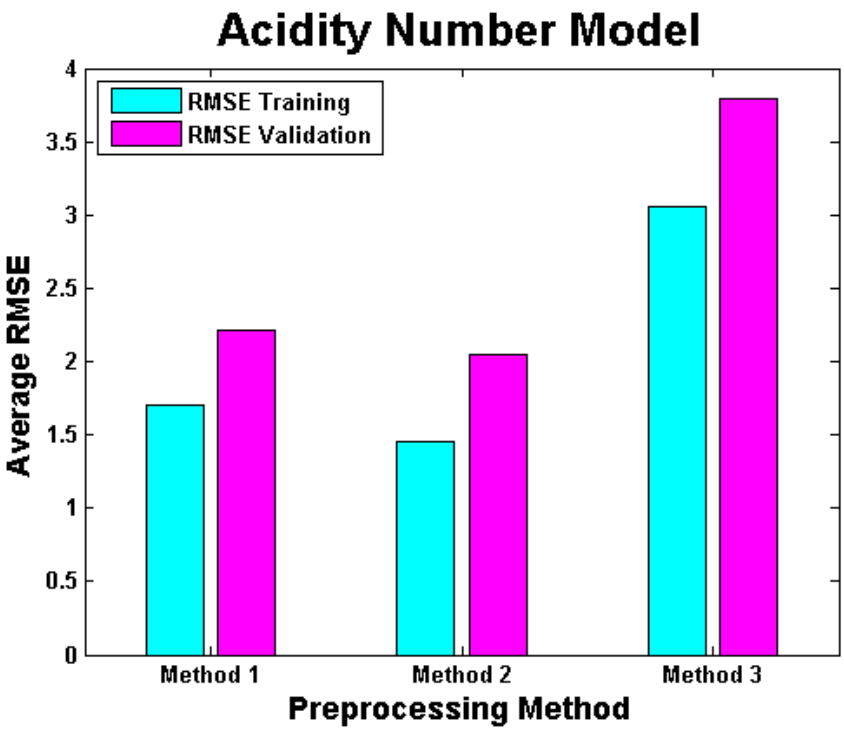

Figure 4.54 Comparison of Acidity Number Full Models from Each Preprocessing Method in Terms of RMSE

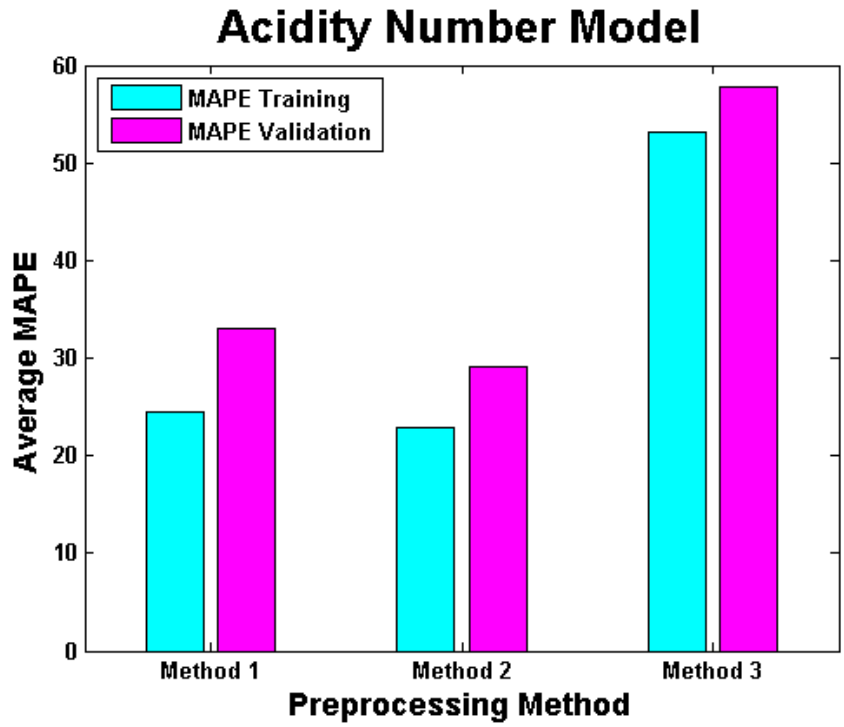

Figure 4.55 Comparison of Acidity Number Full Models from Each Preprocessing Methods in Terms of MAPE

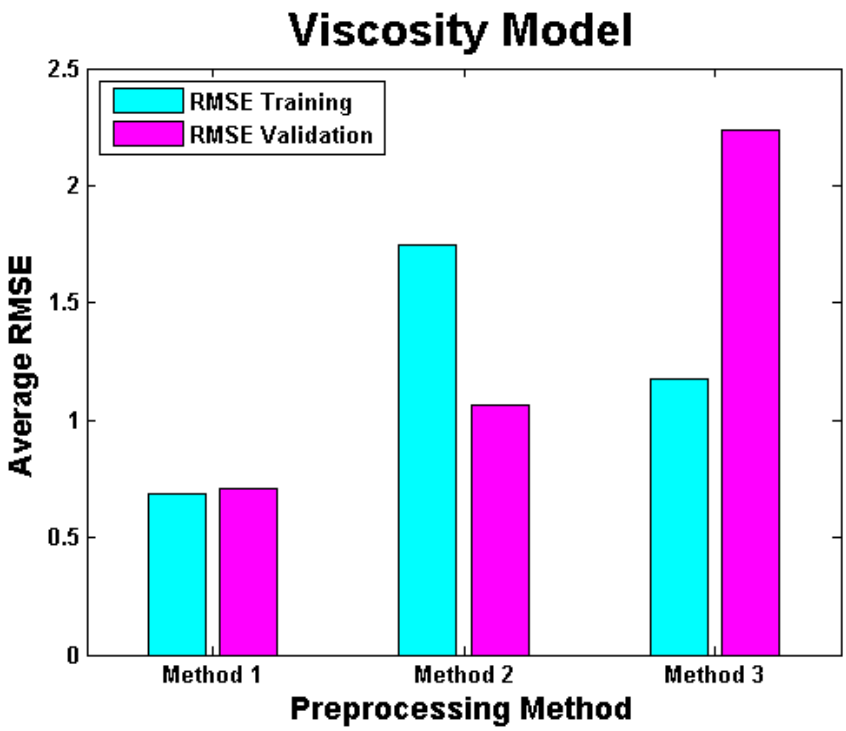

Figure 4.56 Comparison of Viscosity Full Models from Each Preprocessing Methods in Terms on RMSE

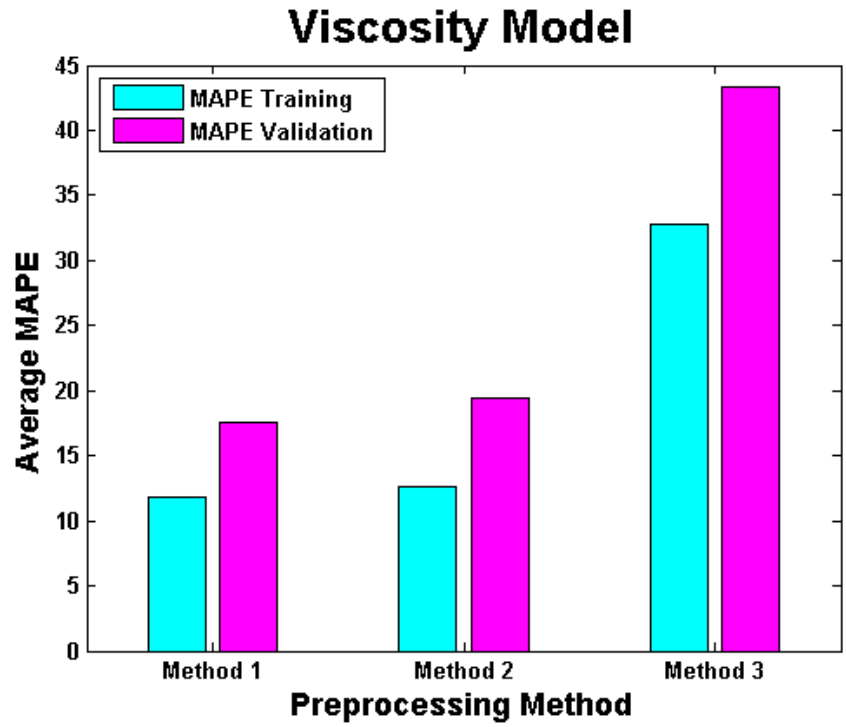

Figure 4.57 Comparison of Viscosity Full Models from Each Preprocessing Methods in Terms of MAPE

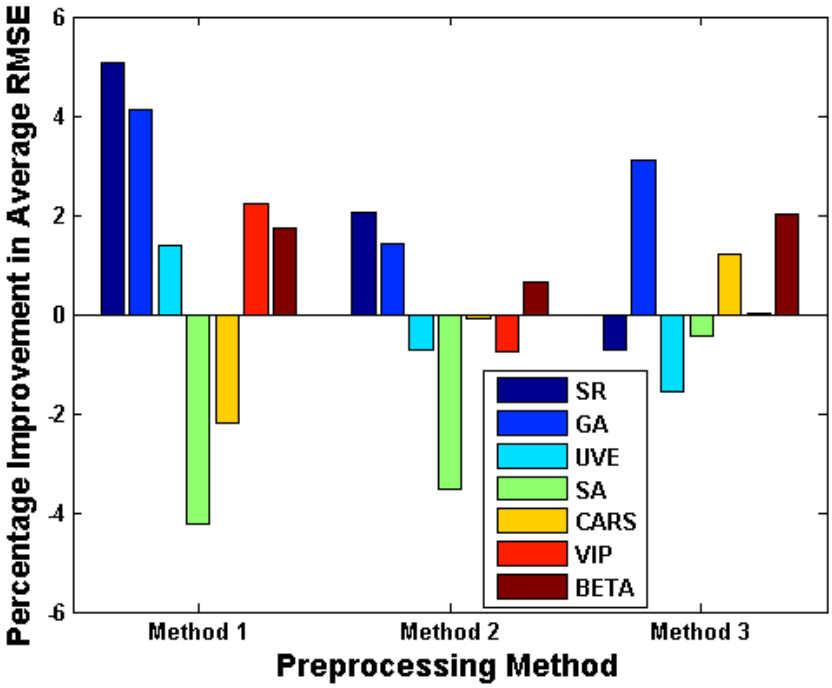

Figure 4.58 Comparison of Different Preprocessing Method of Acidity Number Model in Terms of RMSE in Training Set

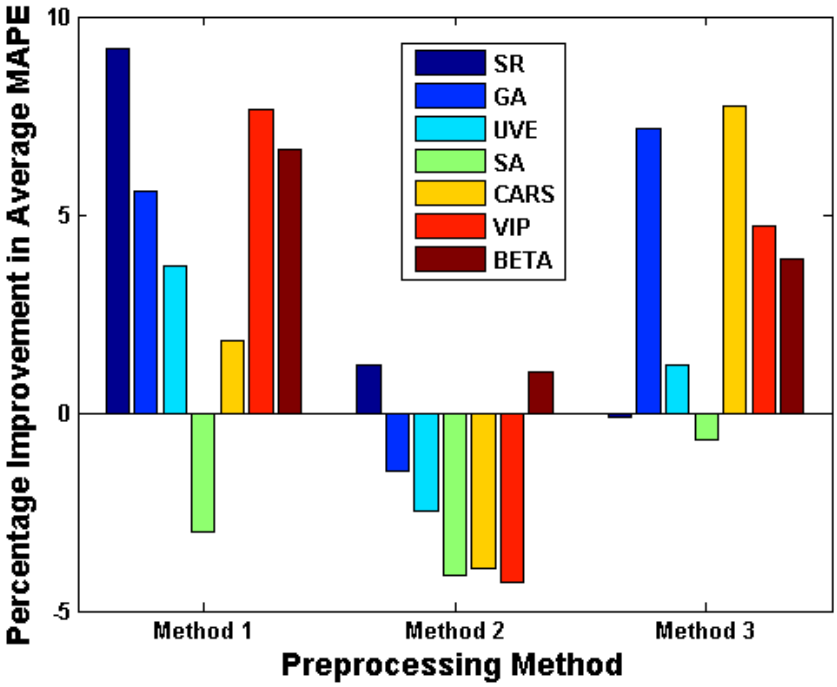

Figure 4.59 Comparison of Different Preprocessing Method of Acidity Number Model in Terms of MAPE in Training Set

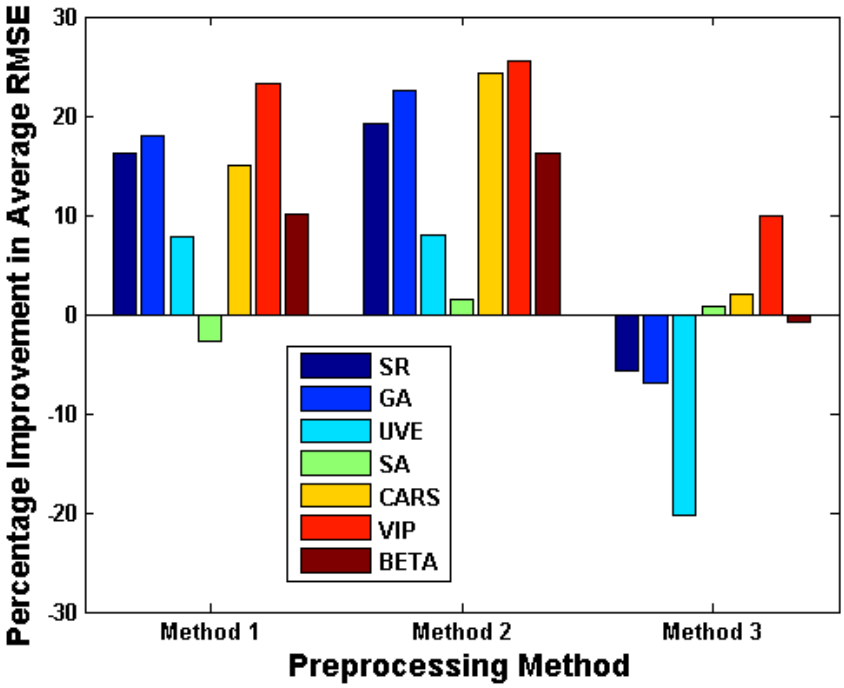

Figure 4.60 Comparison of Different Preprocessing Method of Acidity Number Model in Terms of RMSE in Validation Set

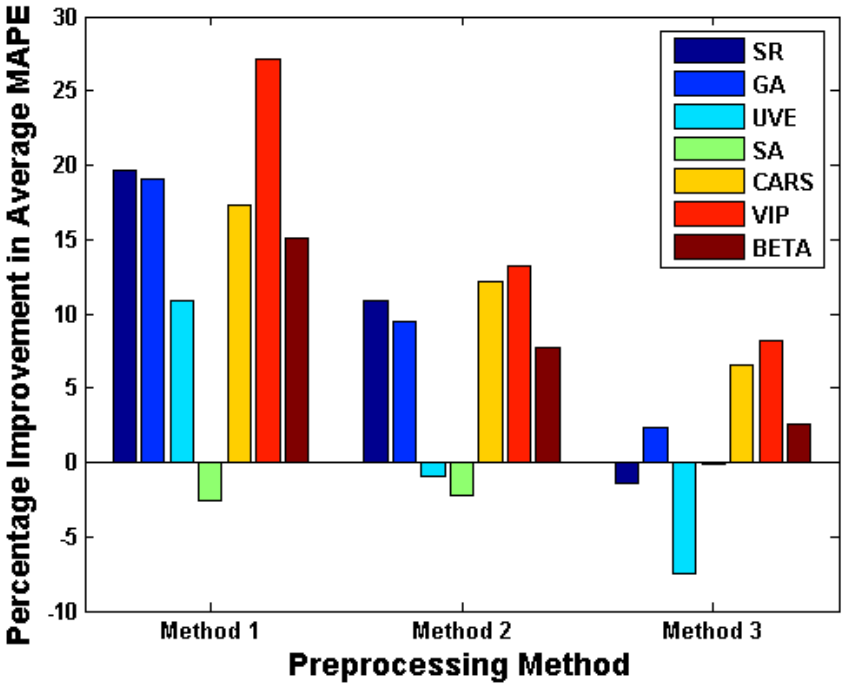

Figure 4.61 Comparison of Different Preprocessing Method of Acidity Number Model in Terms of MAPE in Validation Set

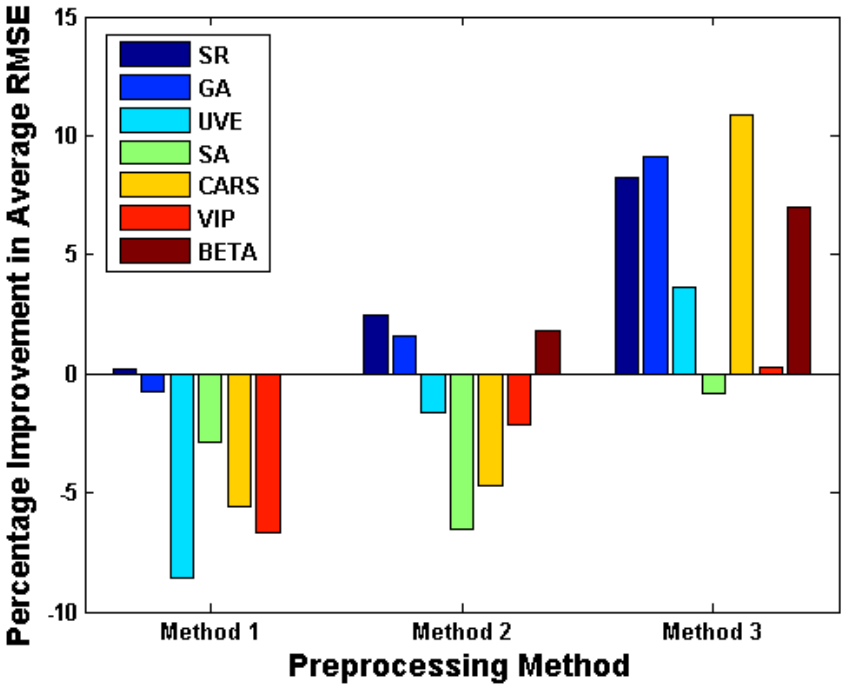

Figure 4.62 Comparison of Different Preprocessing Method of Viscosity Model in Terms of RMSE in Training Set

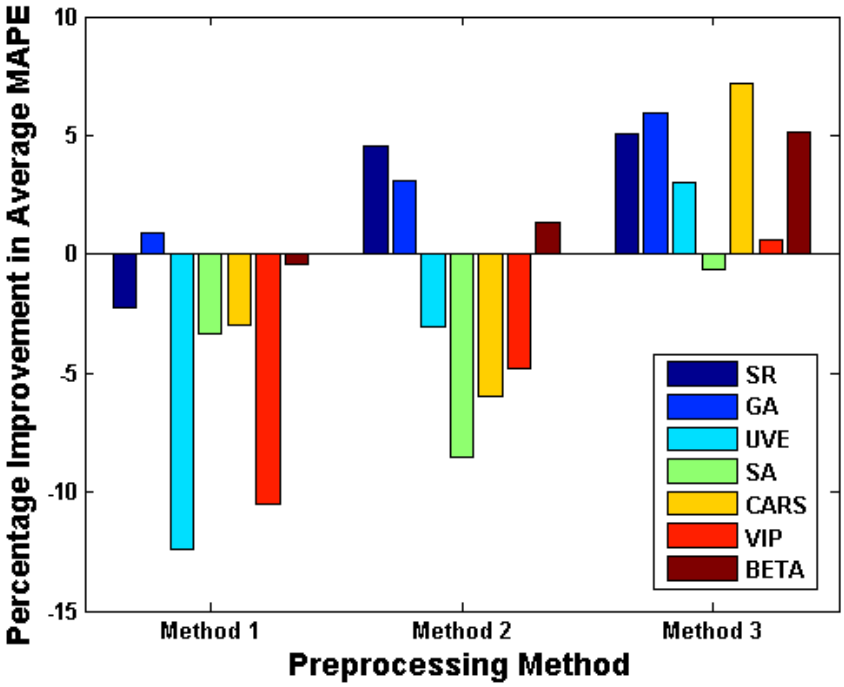

Figure 4.63 Comparison of Different Preprocessing Method of Viscosity Model in Terms of MAPE in Training Set

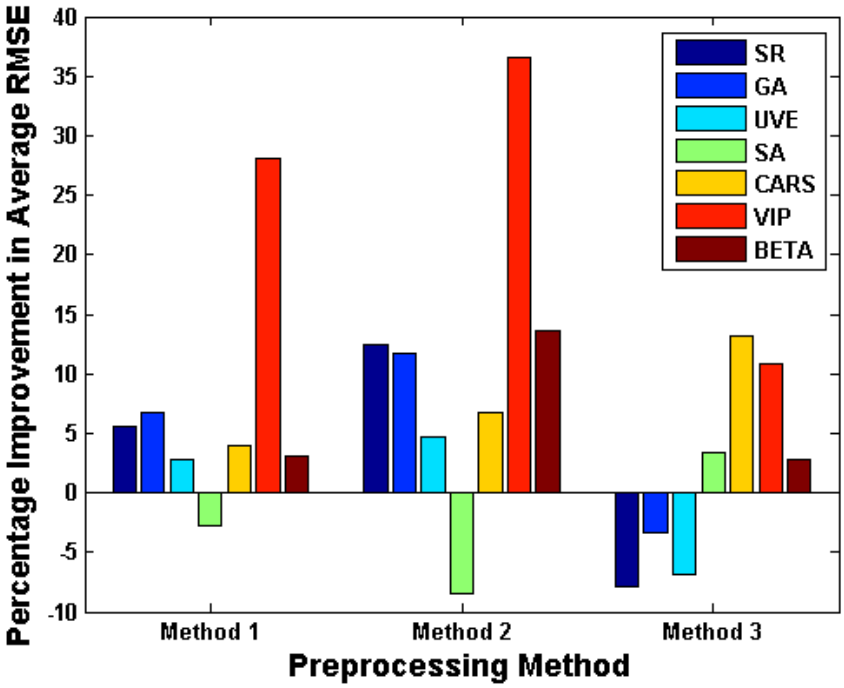

Figure 4.64 Comparison of Different Preprocessing Method of Viscosity Model in Terms of RMSE in Validation Set

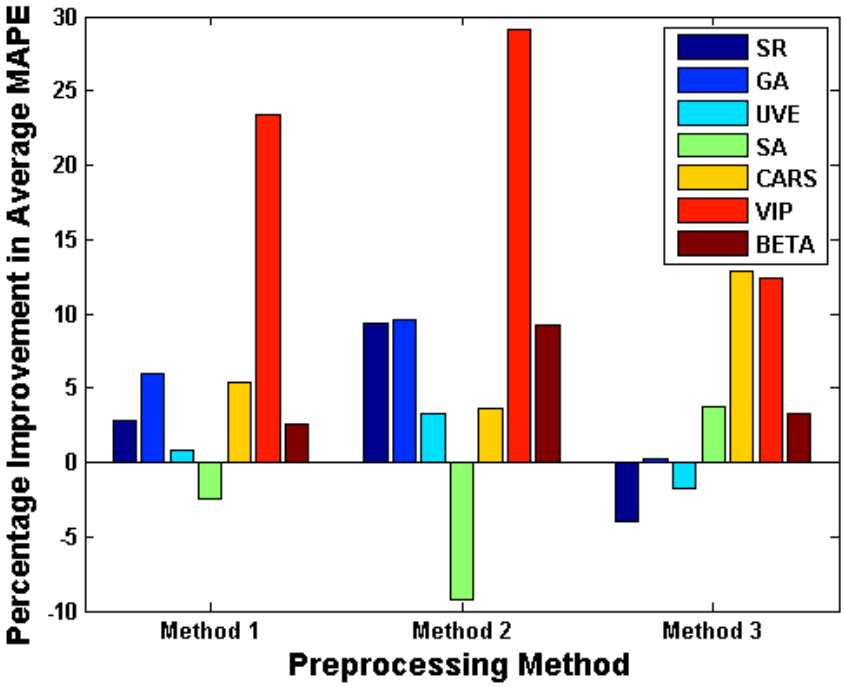

Figure 4.65 Comparison of Different Preprocessing Method of Viscosity Model in Terms of MAPE in Validation Set

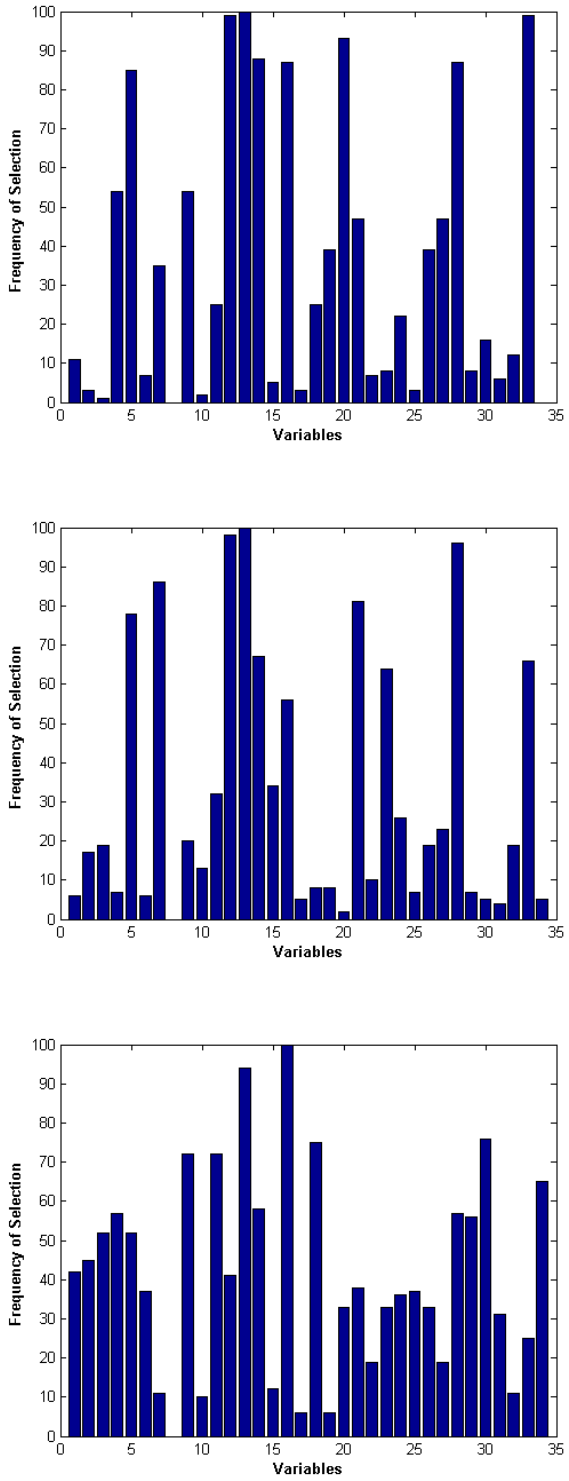

Figure 4.66 Selection Frequency of SR Acidity Number Model

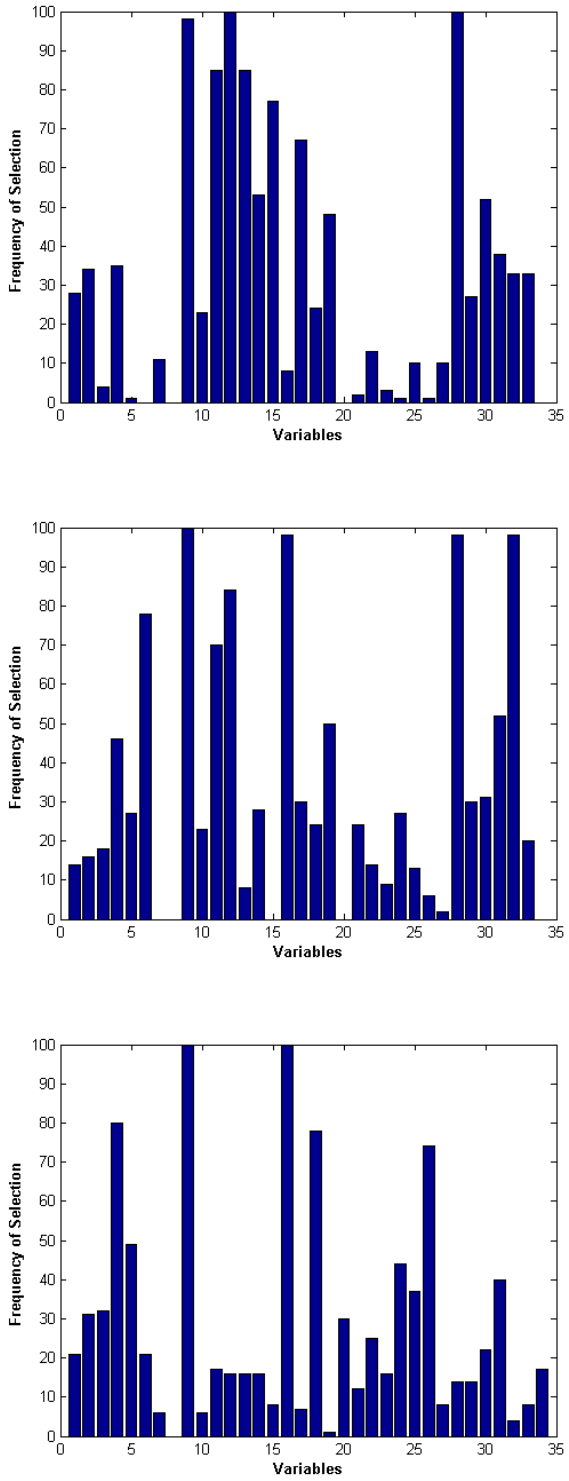

Figure 4.67 Selection Frequency of SR in Viscosity Model

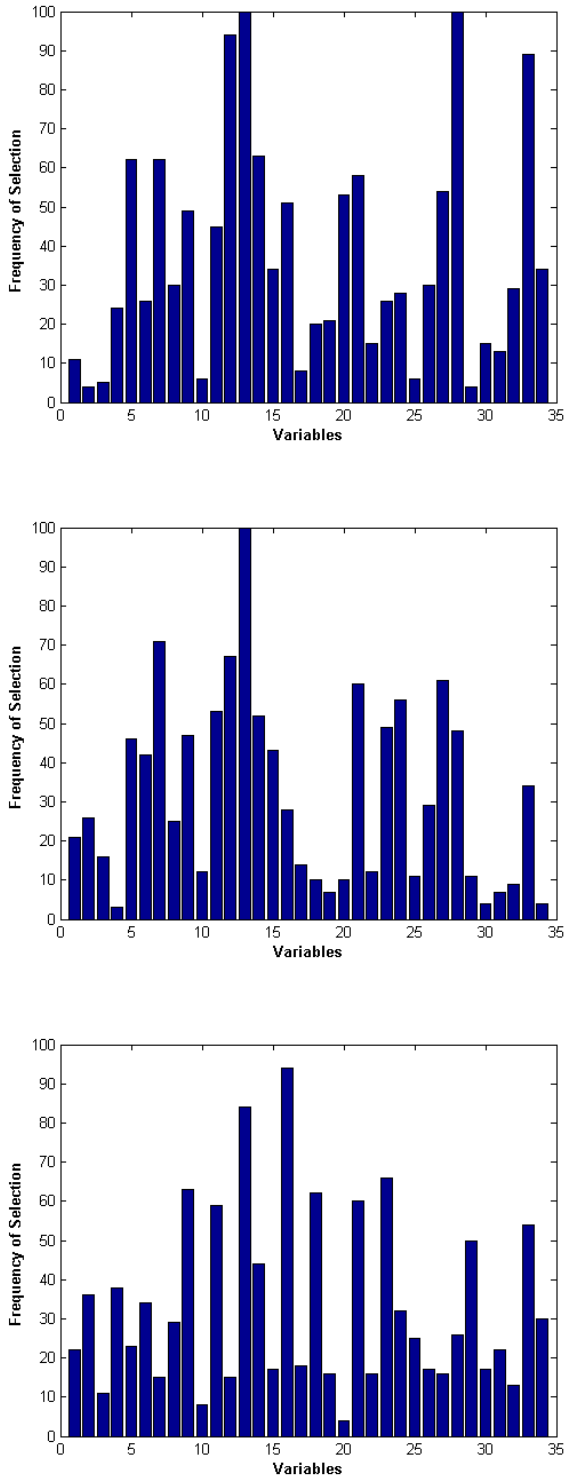

Figure 4.68 Selection Frequency of GA in Acidity Number Model

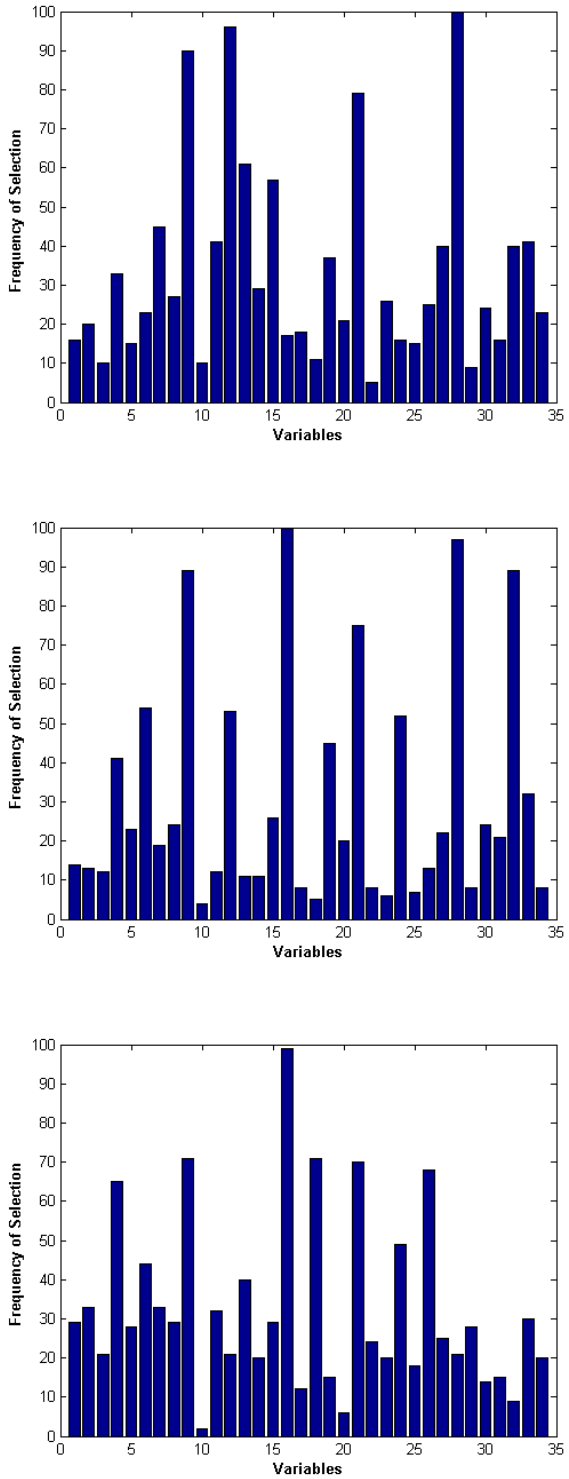

Figure 4.69 Selection Frequency of GA in Viscosity Model

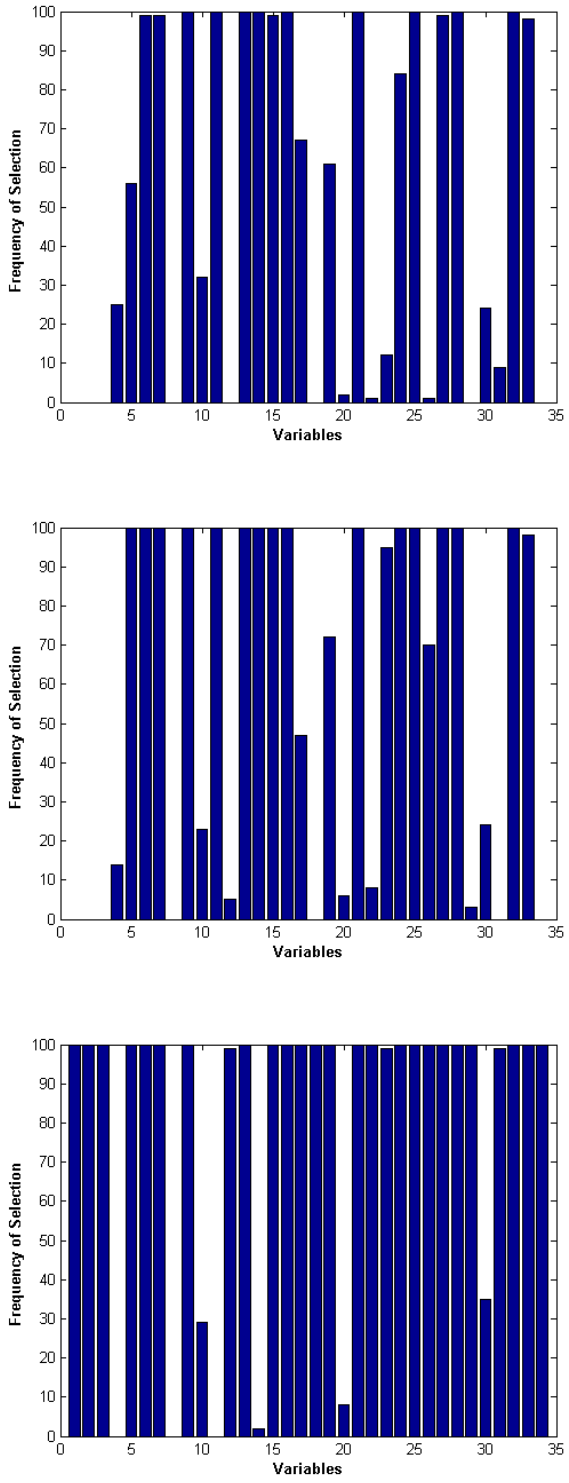

Figure 4.70 Selection Frequency of UVE in Acidity Number Model

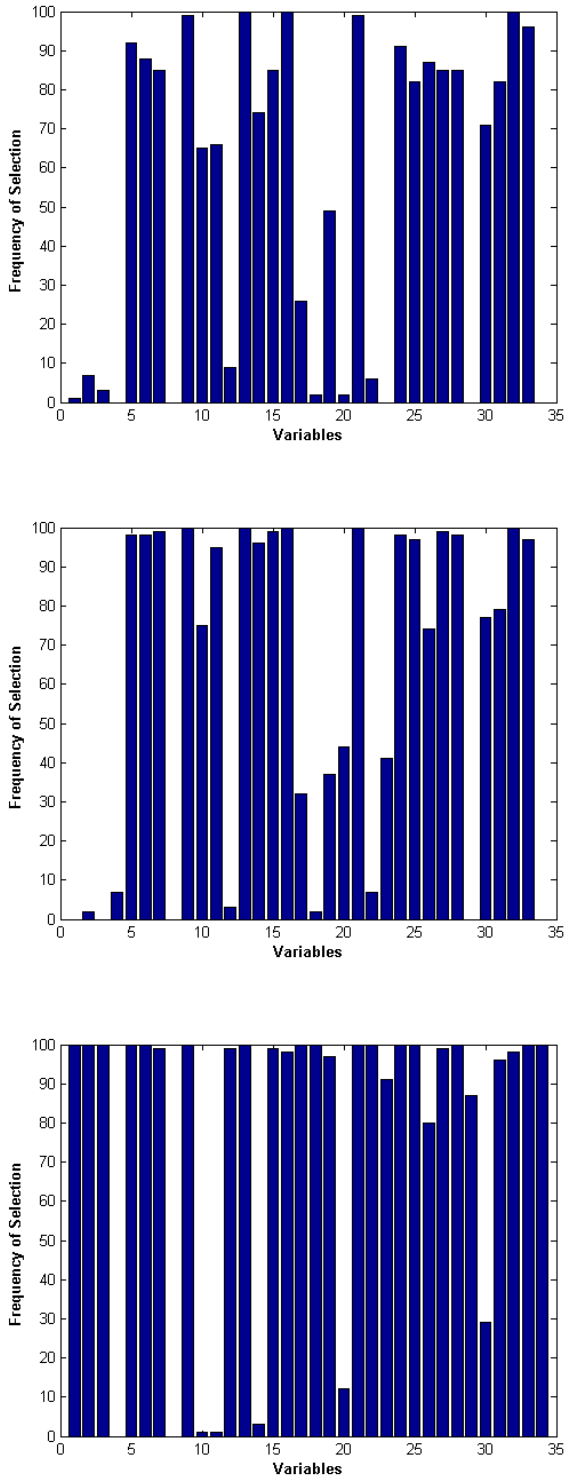

Figure 4.71 Selection Frequency of UVE in Viscosity Model

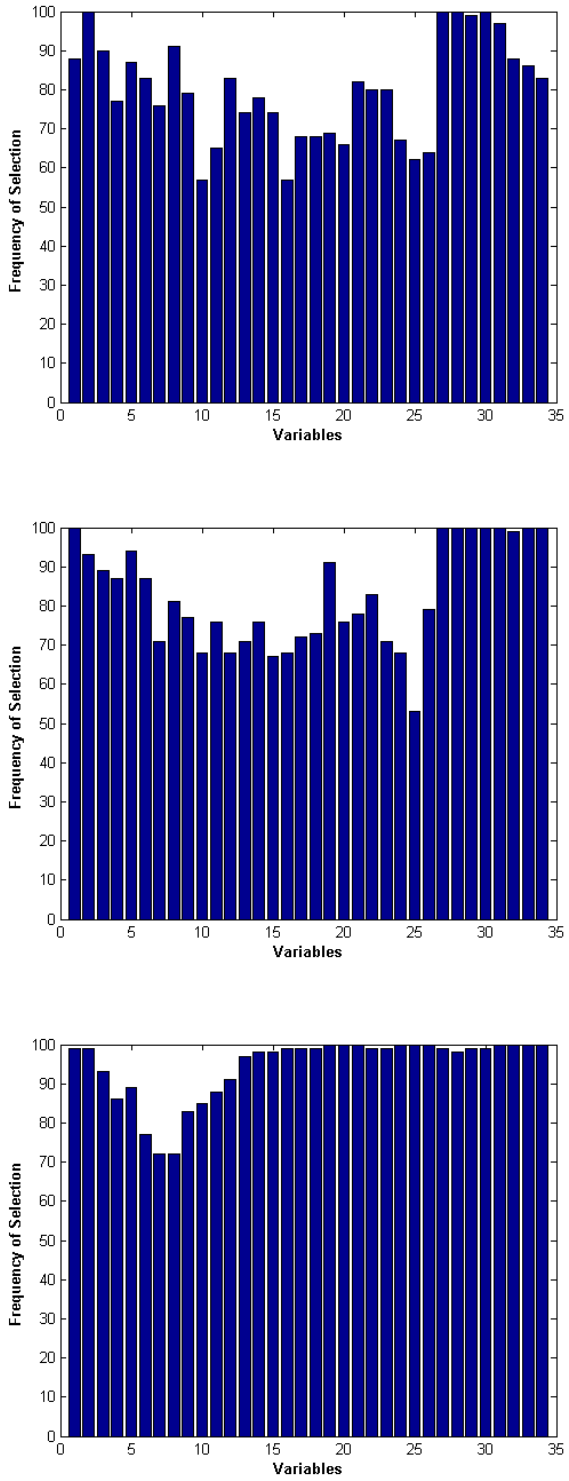

Figure 4.72 Selection Frequency of PLS-SA in Acidity Number Model

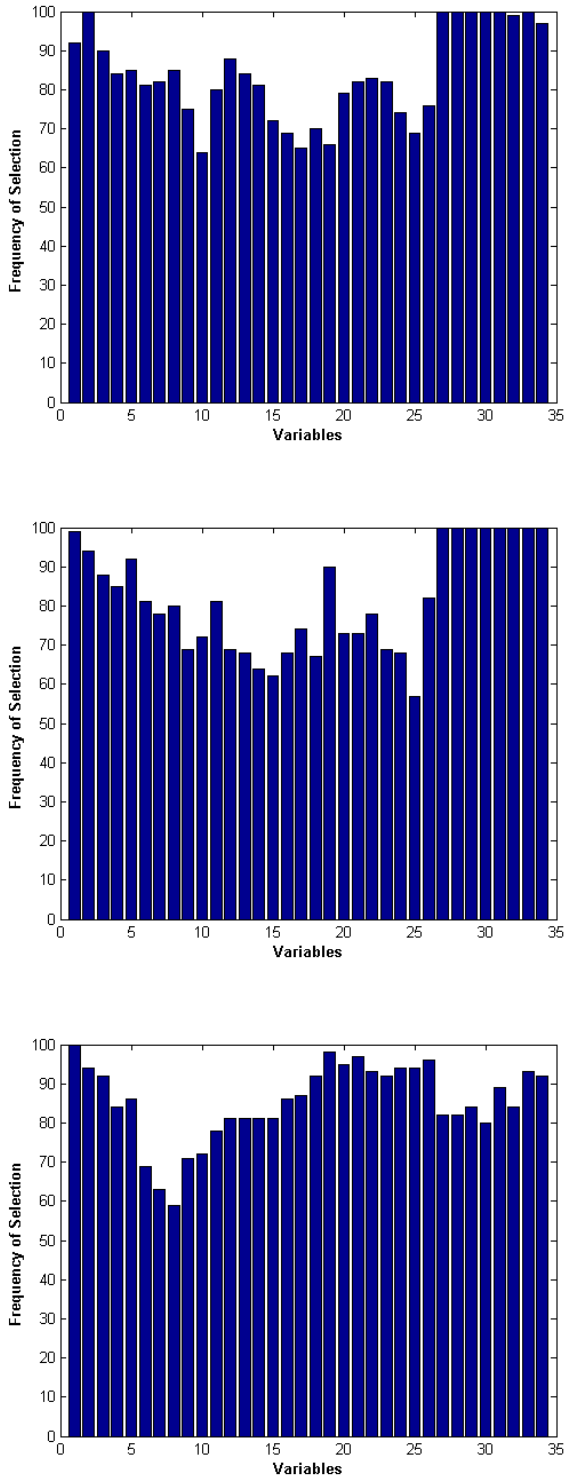

Figure 4.73 Selection Frequency of PLS-SA in Viscosity Model

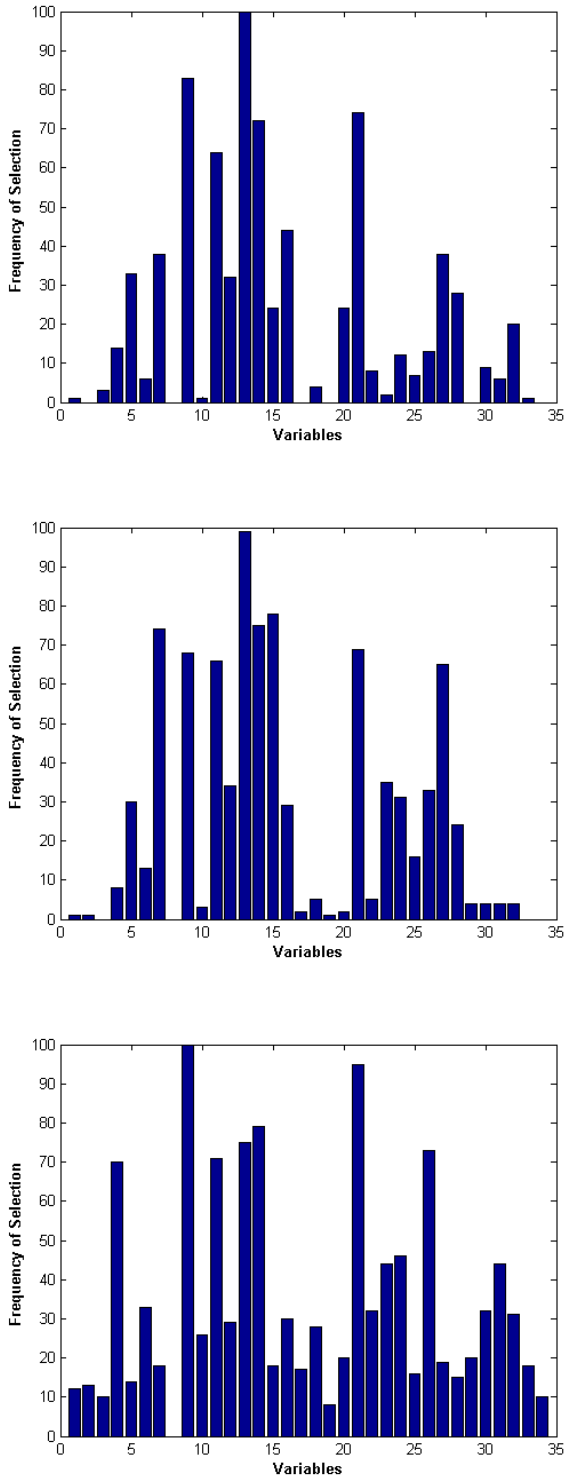

Figure 4.74 Selection Frequency of CARS-PLS in Acidity Number Model

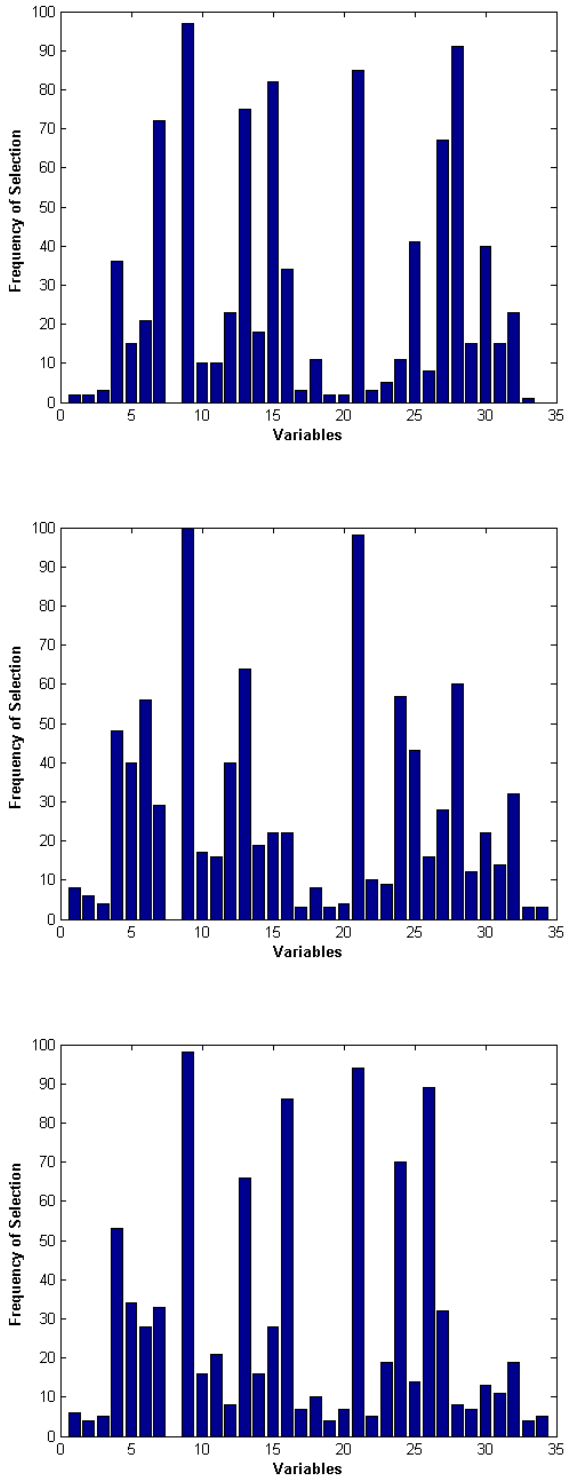

Figure 4.75 Selection Frequency of CARS-PLS in Viscosity Model

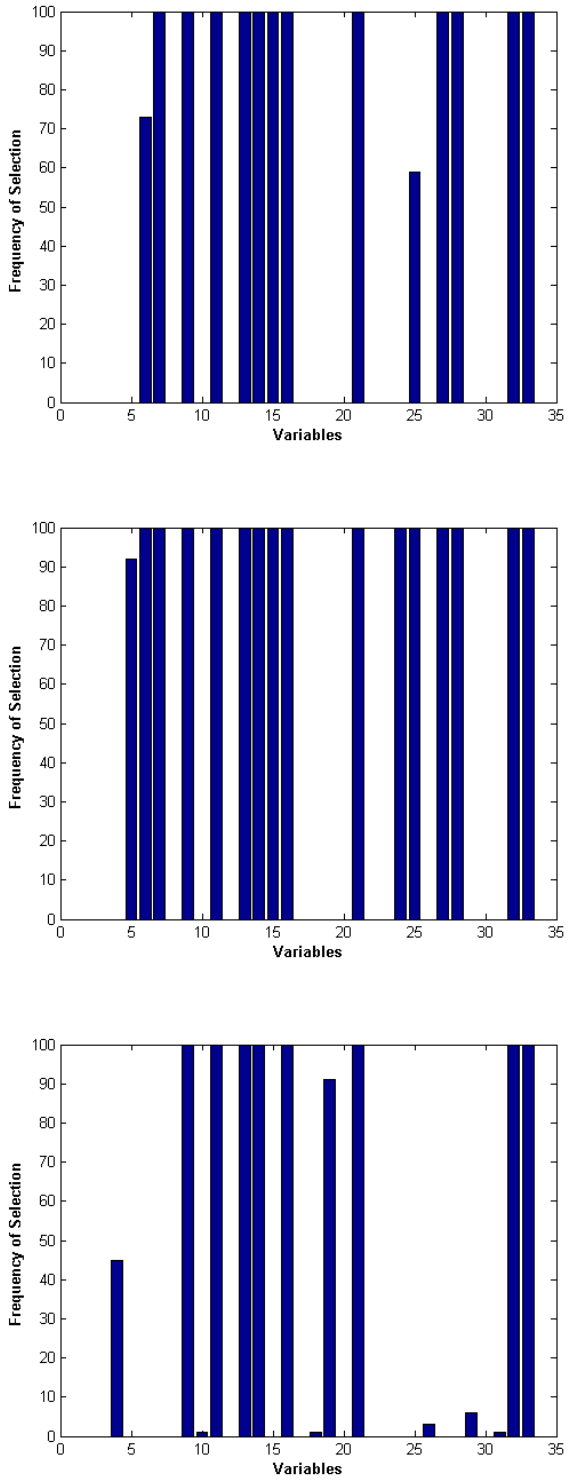

Figure 4.76 Selection Frequency of PLS-VIP in Acidity Number Model

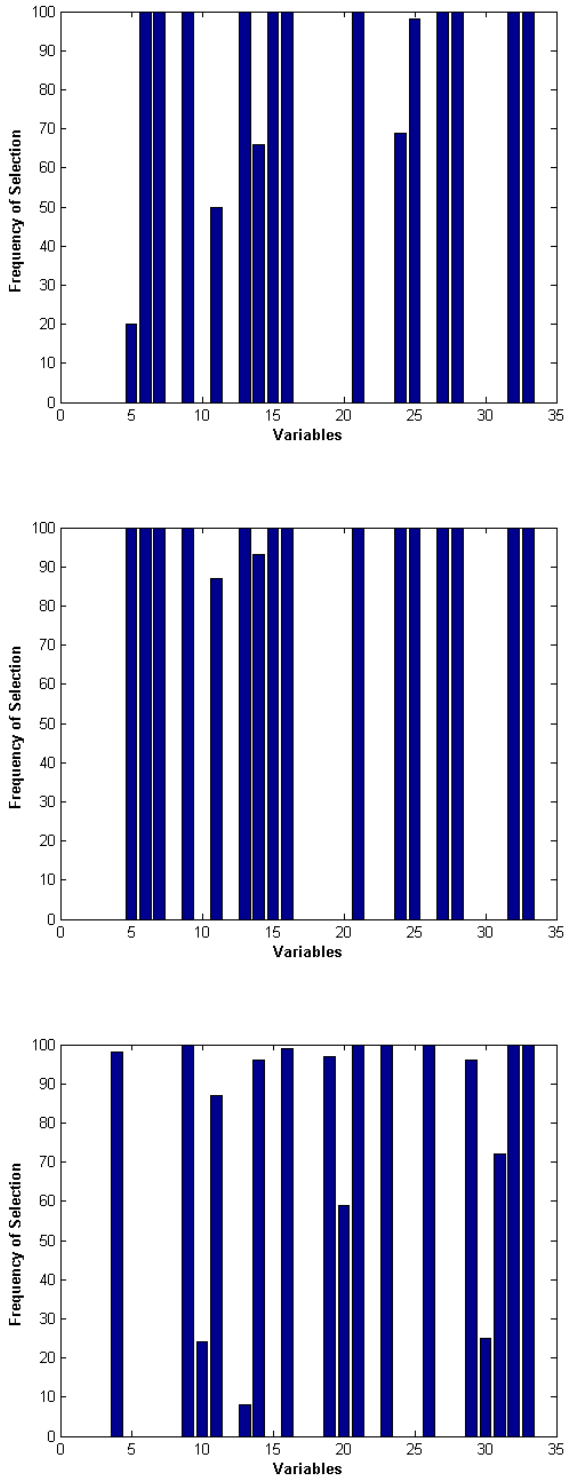

Figure 4.77 Selection Frequency of PLS-VIP in Viscosity Model

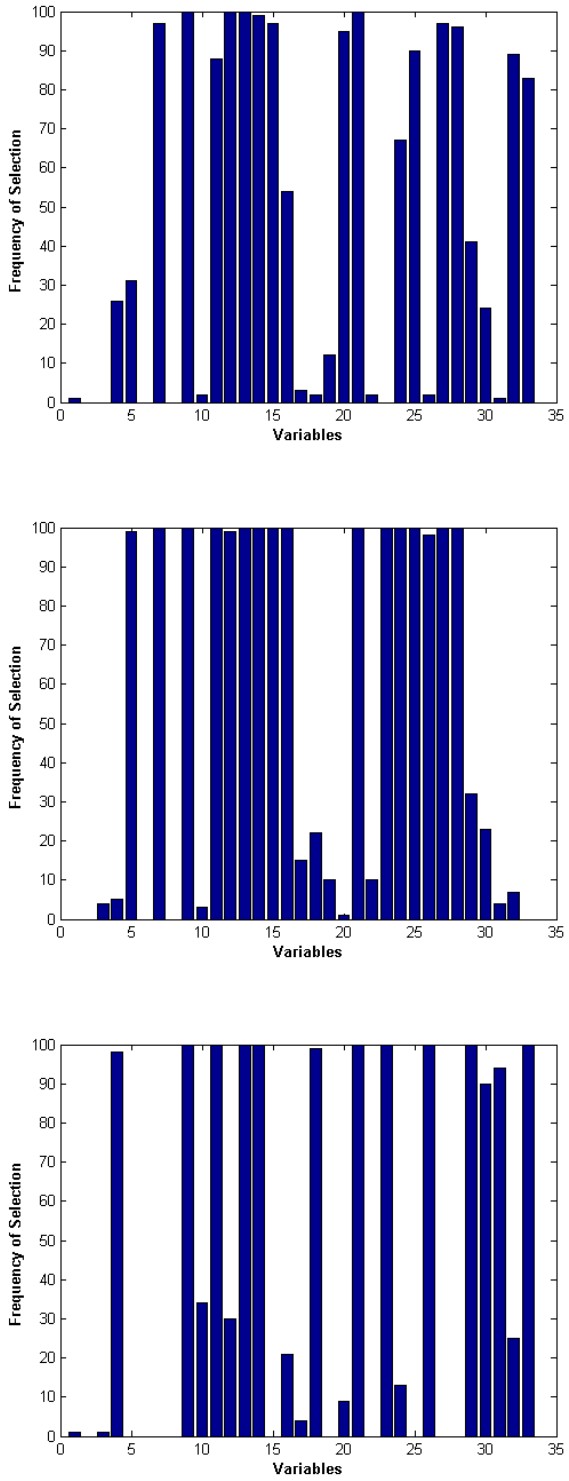

Figure 4.78 Selection Frequency of PLS-BETA Acidity Number Model

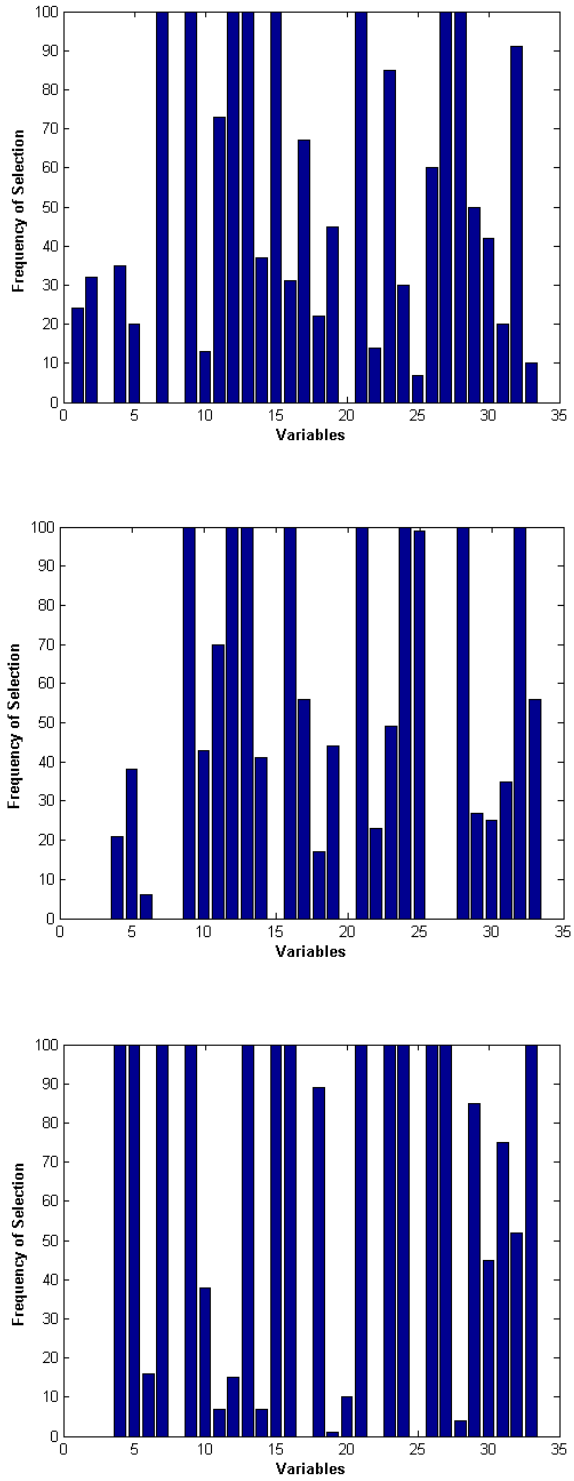

Figure 4.79 Selection Frequency of PLS-BETA in Viscosity Model

Consistency of variable selection is also studied for the industrial case. The frequency plots are presented in Figure 4.66 through Figure 4.79. Frequency of selection is shown in percentage. As results shown, PLS-VIP is the most consistent one among all seven variable selection methods. Variables are either selected with extreme high frequency or not selected at all. Only a few variables are selected in the lower or middle range. The consistency of PLS-BETA is better in the acidity number model than that of the viscosity model. Compare to PLS-VIP, the model size of PLS-BETA are generally larger than PLS-VIP. CARS-PLS produces the smallest models. The consistency of CARS-PLS is unacceptable. Only a few variables are selected with high frequency, and many variables are selected with frequency in the lower range. This agrees with results found in the simulated case study that CARS-PLS is sensitive to data selection. The performance of SR and GA-PLS are the ones in the middle range. UVE-PLS produces models with second largest size. It frequency of selection is relatively consistent. PLS-SA generates the largest models. Also, the prediction performances are actually worsened after variable selection by PLS-SA. This is also the only method that yields worse performance than the original model.

4.3.3 Conclusion and Discussion

According to the analysis of results obtained from three preprocessing methods, the first two preprocessing methods should be adopted. The performances of the first preprocessing methods are very competitive, while the third method is not quite comparable. Especially in prediction performance, the prediction errors are doubled compared to those of the first two preprocessing methods. The improvement after variable selection is not as significant as the other ones.

Based on results obtained from industrial case study, PLS-VIP yields the most superior performance in terms of prediction and selection consistency. In the first two preprocessing methods, PLS-VIP outperforms the other variable selection methods for both acidity number model and viscosity model. Even though CARS-PLS gives best model in the third preprocessing method, due to its inconsistency, CARS-PLS should be applied with care.

Chapter 5. Conclusions and Future Works

5.1 Conclusions

The goal of this project is to implement variable selection algorithms in data-driven soft sensors to improve their predictive power. Seven variable selection methods are investigated: stepwise regression (SR), genetic algorithm (GA) with PLS, uninformative variable elimination (UVE) with PLS, PLS with sensitivity analysis (SA), competitive adaptive reweighted sampling (CARS) with PLS, PLS with variable importance in projection (VIP) and regression coefficients (BETA). The characteristics of these methods are explored by using a simulated case study and an industrial case study.

Based on the analysis results, PLS-VIP gives the most superior performance in both simulated and industrial case. PLS-VIP is very straight-forward, which selects the relevant predictors based on its importance in the PLS projection. It is shown that the submodels produced by PLS-VIP outperform other methods significantly, especially in the industrial case study. Furthermore, the models produced by PLS-VIP are very consistent from one sampling run to the other, which shows its robustness to data selection. On the other hand, CARS-PLS is quite sensitive to data selection. The standard deviations of the models produced by CARS-PLS are much larger than the other ones. The next in line would be PLS-BETA and SR. PLS-BETA performs very well when the contributions of each relevant predictor are in the same range. However, this is not always the case the industrial processes. PLS-BETA tends to only select the variables with dominating contribution, which may over simplify the model and cause overfit. SR also has

issues with overfitting. From the results shown in the industrial case study, SR always gives the greatest calibration models, but prediction performance on the external validation set is not quite ideal. GA-PLS yields similar results to SR. However, the computation time of GA-PLS is much longer than that of SR. GA-PLS also required more tedious preliminary setting of GA parameters. UVE-PLS and PLS-SA generates models with relatively large size. Nonetheless, UVE-PLS is able to identify a subset of variables that would improve the prediction performance, whereas the submodels produced by PLS-SA worsen the prediction performance. The strength and limitations of each method are summarized in Table 5.1.

Table 5.1 Limitations and Strengths of Each Variable Selection Method

Models	Pros	Cons
SR	• Produce high performance training model • Relatively consistent selection	• Developed model tend to overfit the model
GA-PLS	• Performance can be improved by tuning the parameters	• Require a lot of user input to optimize performance • Selects irrelevant predictors • Computation load
UVE-PLS	• Relatively consistent selection • Improvement observed in prediction performance	• Large model size
PLS-SA	• Selection consistency	• Low prediction performance • Heavy computation load
CARS-PLS	• Use as preliminary variable reduction in wavelength selection	• Selection inconsistency
PLS-VIP	• Selection consistency • High prediction performance • Least computation load	• May select some irrelevant predictors around the relevant ones
PLS-BETA	• Selection consistency • High prediction performance • Low computation load	• User input requires for the cutoff value of BETA

5.2 Future Works

Variable reduction can be carried out prior to variable selection based on two rules: elimination of variables with zero-variance and elimination of highly correlated variables. The variable selection methods can also be improved by considering the application of modeling power approach. Modeling power approach balances the predictive and descriptive abilities of model.

Application of wavelength selection is also of interest. The next step of research is to implement these seven variable selection methods on a benchmark dataset, NIR spectral of diesel fuel.

The optimal goal of our study is to implement variable selection method in the framework of Statistics Pattern Analysis (SPA). Due to the characteristics of SPA, it is very likely that the number of regressors would be greater than the number of samples. Variable selection method could be implemented to eliminate the uninformative variables prior to SPA. In addition, variable selection can also be employed to select useful statistics in statistics pattern generation.

Bibliography

[1] C. M. Andersen and R. Bro, "Variable selection in regression — a tutorial," *Journal of Chemometrics*, vol. 24, no. 11–12, pp. 728-737, 2010.

[2] J. Reunanen, "Overfitting in making comparisons between variable selection methods," *The Journal of Machine Learning Research*, vol. 3, pp. 1371-1382, 2003.

[3] M.-D. Ma et al., "Development of adaptive soft sensor based on statistical identification of key variables," *Control Engineering Practice*, vol. 17, no. 9, pp. 1026-1034, Sep. 2009.

[4] I.-G. Chong and C.-H. Jun, "Performance of some variable selection methods when multicollinearity is present," *Chemometrics and Intelligent Laboratory Systems*, vol. 78, no. 1–2, pp. 103-112, Jul. 2005.

[5] V. Centner, D. L. Massart, O. E. de Noord, S. de Jong, B. M. Vandeginste, and C. Sterna, "Elimination of uninformative variables for multivariate calibration.," *Analytical chemistry*, vol. 68, no. 21, pp. 3851-3858, Nov. 1996.

[6] F. A. Arciniegas, M. Embrechts, and I. E. A. Rueda, "Variable Selection with Partial Least Squares Sensitivity Analysis: An Application to Currency Crises' Real Effects," *SSRN eLibrary*, Jun. 2006.

[7] L. H. Chiang and R. J. Pell, "Genetic algorithms combined with discriminant analysis for key variable identification," *Analytical Sciences*, vol. 14, pp. 143-155, 2004.

[8] M. Forina, S. Lanteri, M. Casale, and M. C. Cerrato Oliveros, "Stepwise orthogonalization of predictors in classification and regression techniques: An 'old' technique revisited," *Chemometrics and Intelligent Laboratory Systems*, vol. 87, no. 2, pp. 252-261, Jun. 2007.

[9] J.-P. Gauchi and P. Chagnon, "Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 171-193, Oct. 2001.

[10] D. Broadhurst, R. Goodacre, A. Jones, J. J. Rowland, and D. B. Kell, "Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry," *Analytica Chimica Acta*, vol. 348, no. 1–3, pp. 71-86, Aug. 1997.

[11] M. J. Arcosa, M. C. Ortizav, B. Villahoz, and L. A. Sarabiab, "Genetic-algorithm-based wavelength selection in multicomponent spectrometric determinations by PLS : application on indomethacin and acemethacin mixture," 1997.

[12] H. Kaneko and K. Funatsu, "A New Process Variable and Dynamics Selection Method Based on a Genetic Algorithm-Based Wavelength Selection Method," vol. 58, no. 6, 2012.

[13] G. Jones, P. Willett, and R. Glen, "Genetic algorithms for chemical structure handling and molecular recognition," in *In Genetic Algorithms in Molecular Modeling*, 1996, pp. 211-242.

[14] W. Cai, Y. Li, and X. Shao, "A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 90, no. 2, pp. 188-194, Feb. 2008.

[15] X. Shao, F. Wang, D. Chen, and Q. Su, "A method for near-infrared spectral calibration of complex plant samples with wavelet transform and elimination of uninformative variables," *Analytical and Bioanalytical Chemistry*, vol. 378, no. 5, pp. 1382-1387, 2004.

[16] J. Koshoubu, T. Iwata, and S. Minami, "Application of the Modified UVE-PLS Method for a Mid-Infrared Absorption Spectral Data Set of Water-Ethanol Mixtures," *Applied Spectroscopy*, vol. 54, no. 1, pp. 148-152, Jan. 2000.

[17] J. Koshoubu, T. Iwata, and S. Minami, "Elimination of the uninformative calibration sample subset in the modified UVE(Uninformative Variable Elimination)-PLS (Partial Least Squares) method.," *Analytical sciences : the international journal of the Japan Society for Analytical Chemistry*, vol. 17, no. 2, pp. 319-22, Feb. 2001.

[18] S. Ye, D. Wang, and S. Min, "Successive projections algorithm combined with uninformative variable elimination for spectral variable selection," *Chemometrics and Intelligent Laboratory Systems*, vol. 91, no. 2, pp. 194-199, Apr. 2008.

[19] E. Zamprogna, M. Barolo, and D. E. Seborg, "Optimal selection of soft sensor inputs for batch distillation columns using principal component analysis," *Journal of Process Control*, vol. 15, no. 1, pp. 39-52, Feb. 2005.

[20] Q. Li and C. Shao, "Soft sensing modelling based on optimal selection of secondary variables and its application," *International Journal of Automation and Computing*, vol. 6, no. 4, pp. 379-384, Oct. 2009.

[21] H. Li, Y. Liang, Q. Xu, and D. Cao, “Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration.,” *Analytica chimica acta*, vol. 648, no. 1, pp. 77-84, Aug. 2009.

[22] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, and D.-S. Cao, “Model population analysis for variable selection,” *Journal of Chemometrics*, vol. 24, no. 7–8, pp. 418-423, Jul. 2010.

[23] H.-dong Li, Y.-zeng Liang, and Q.-song Xu, “Model Population Analysis for Statistical Model Comparison,” no. 1, pp. 3-21.

[24] T. Mehmood, H. Martens, S. Sæbø, J. Warringer, and L. Snipen, “A Partial Least Squares based algorithm for parsimonious variable selection.,” *Algorithms for molecular biology : AMB*, vol. 6, no. 1, p. 27, Jan. 2011.

[25] F. Lindgren, B. Hansen, and W. Karcher, “MODEL VALIDATION BY PERMUTATION TESTS :,” *Journal of Chemometrics*, vol. 10, pp. 521-532, 1996.

[26] R. Gosselin, D. Rodrigue, and C. Duchesne, “A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications,” *Chemometrics and Intelligent Laboratory Systems*, vol. 100, no. 1, pp. 12-21, Jan. 2010.

[27] C. M. Andersen and R. Bro, “Variable selection in regression—a tutorial,” *Journal of Chemometrics*, vol. 24, no. 11-12, pp. 728-737, Nov. 2010.

[28] P. P. Roy and K. Roy, “On Some Aspects of Variable Selection for Partial Least Squares Regression Models,” *QSAR & Combinatorial Science*, vol. 27, no. 3, pp. 302–313, 2008.

[29] D. Wang and R. Srinivasan, “Data-Driven Soft Sensor Approach for Quality Prediction in a Refining Process,” *IEEE Transactions on Industrial Informatics*, vol. 6, no. 1, pp. 11-17, Feb. 2010.

[30] P. Kadlec, B. Gabrys, and S. Strandt, “Data-driven Soft Sensors in the process industry,” *Computers & Chemical Engineering*, vol. 33, no. 4, pp. 795-814, Apr. 2009.

[31] P. Kadlec, R. Grbić, and B. Gabrys, “Review of adaptation mechanisms for data-driven soft sensors,” *Computers & Chemical Engineering*, vol. 35, no. 1, pp. 1-24, Jan. 2011.

[32] P. Kadlec and B. Gabrys, “Adaptive Local Learning Soft Sensor for Inferential Control Support,” 2008, pp. 243-248.

[33] I. T. Jolliffee, *Principal Component Analysis*. Springer, 2002.

[34] S. Wold, M. Sjostrom, and L. Eriksson, "PLS-regression : a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, pp. 109-130, 2001.

[35] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning : Data Mining , Inference and Prediction Probability Theory : The Logic of Science The Fundamentals of Risk Measurement Mathematicians , pure and applied , think there is something weirdly different about," vol. 27, no. 2, pp. 83-85, 2005.

[36] C. Lin and C. Lee, *Neural fuzzy systems: A neuro-fuzzy synergism to intelligent systems*. Upper Saddle River: Prentice-Hall Inc., 1996.

[37] V. N. Vapnik, *Statistical learning theory*. Wiley New York:, 1998.

[38] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica Chimica Acta*, vol. 185, pp. 1-17, 1986.

[39] S. Wold, H. Martens, and H. Russwurm Jr, *Food Research and Data Analysis*. London: Applied Science Publishers, 1983.

[40] S. Wold and B. Kowalski, *Chemometrics: Mathematics and Statistics in Chemistry*. Dordrecht: Reidel, 1984.

[41] R. Leardi and A. Lupiáñez González, "Genetic algorithms applied to feature selection in PLS regression: how and when to use them," *Chemometrics and Intelligent Laboratory Systems*, vol. 41, no. 2, pp. 195-207, Jul. 1998.

[42] D. Broadhursta, J. J. Rowlandb, and D. B. Kelp, "Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression , with applications to pyrolysis mass spectrometry," *Analytica Chimica Acta*, vol. 348, pp. 71-86, 1997.

[43] L. Davis, *Genetic algorithms and simulated annealing*. 1987.

[44] D. E. Goldberg, *Genetic Algorithms in Search, Optimization & Machine Learning*. Addison-Wesley, 1988.

[45] D. E. Goldberg and J. H. Holland, "Genetic Algorithms and Machine Learning," in *Machine Learning*, vol. 3, Kluwer Academic Publishers, 1988, pp. 95-99.

[46] R. Leardi, R. Boggia, and M. Terrile, "Genetic algorithms as a strategy for feature selection," *Journal of Chemometrics*, vol. 6, no. 5, pp. 267-281, Sep. 1992.

[47] R. H. Kewley, M. J. Embrechts, and C. Breneman, "Data strip mining for the virtual design of pharmaceuticals with neural networks.," *IEEE transactions on neu-*

ral networks / a publication of the IEEE Neural Networks Council, vol. 11, no. 3, pp. 668-79, Jan. 2000.

[48] S. Wold, E. Johansson, and M. Cocchi, "PLS_Partial Least Squares Projections to Latent Structures.pdf," in *3D QSAR in Drug Design Theory Methods and Application*, ESCOM, 1993, pp. 523-550.

[49] P. Facco, F. Doplicher, F. Bezzo, and M. Barolo, "Moving average PLS soft sensor for online product quality estimation in an industrial batch polymerization process," *Journal of Process Control*, vol. 19, no. 3, pp. 520-529, Mar. 2009.

[50] P. Facco, F. Bezzo, and M. Barolo, "Nearest-Neighbor Method for the Automatic Maintenance of Multivariate Statistical Soft Sensors in Batch Processing," *Industrial & Engineering Chemistry Research*, vol. 49, no. 5, pp. 2336-2347, Mar. 2010.

[51] S. Wold, N. Kettaneh, H. Fridén, and A. Holmberg, "Modelling and diagnostics of batch processes and analogous kinetic experiments," *Chemometrics and Intelligent Laboratory Systems*, vol. 44, no. 1–2, pp. 331-340, Dec. 1998.

[52] P. Nomikos and J. F. MacGregor, "Multivariate SPC Charts for Monitoring Batch Processes," *Technometrics*, vol. 37, no. 1, pp. 41-59, Feb. 1995.

[53] P. Nomikos and J. F. MacGregor, "Multi-way partial least squares in monitoring batch processes," *Chemometrics and Intelligent Laboratory Systems*, vol. 30, no. 1, pp. 97-108, Nov. 1995.