Variable Selection for Industrial Process Modeling and Monitoring

by

Zixiu Wang

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
December 12, 2015

Keywords: Variable Selection, Soft Sensor, Data-Driven Models, Process Industry, Model Sampling, Fault Detection and Diagnosis

Approved by

Jin Wang, Chair, B. Redd Associate Professor of Chemical Engineering
Qinghua He, Associate Professor of Chemical Engineering, Tuskegee University
W. Robert Ashurst, Associate Professor of Chemical Engineering
Steve R. Duke, Alumni Associate Professor of Chemical Engineering
Tong Li, Senior Research Scientist, Air Liquide

Abstract

In recent years, rapid developments in technology facilitated the collection of vast amount of data from different industrial processes. Data-driven soft sensors have been widely used in both academic research and industrial applications for predicting hard-to-measure variables or replacing physical sensors to reduce cost. It has been shown that the performance of these data-driven soft sensors could be greatly improved by selecting only the vital variables that strongly affect the primary variables, rather than using 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 model complexity and computational load, provide better insight into the nature of the process, and lower the cost of measurements. Given the importance of variable selection, a systematic evaluation of variable selection performance becomes essential. However, the existing performance indicators all have limitations.

In this work, a comprehensive evaluation of different variable selection methods for PLS-based soft sensor development is presented, and a new metric is proposed to assess the performance of different variable selection methods. The new performance indicator incorporates information entropy to measure how consistently variable selection performs over multiple Monte Carlos runs. When the ground truth of the data is not available, only consistency index can be accessed to evaluate the variable selection performance, along with the prediction capability. The following seven variable selection methods are compared: stepwise regression (SR), partial least squares (PLS) with regression coefficients (PLS-BETA), PLS with variable importance in projection (PLS-VIP), uninformative variable elimination with PLS (UVE-PLS), genetic algorithm with PLS (GA-PLS), competitive adaptive reweighted sampling with PLS (CARS-PLS), and least absolute shrinkage and selection operator (Lasso). 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 demonstrated by static case studies (a simulated case and an industrial polyester production) and dynamic case studies (a digester simulator and an industrial Kamyr digester). 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 industrial data, by considering four factors: proportion of relevant predictors, magnitude of correlation between predictors, magnitude of signal to noise ratio, and structure of regression coefficients. In addition, the algorithms are applied to an industrial case study, the production of polyester resin, to test their performance. In both simulated and industrial polyester case studies, Monte Carlos (MC) simulation is adopted to generate different combinations of training, tuning, and testing datasets. Independent tuning datasets are used to optimize each method and to analyze the sensitivities of each method to its tuning parameters. Then independent test datasets are used to compare the prediction performances of PLS models built from different subsets of regressors retained by these variable selection methods. Based on the results, PLS-VIP is the most consistent method, based on both selection and prediction performances. Along with data preprocessing and correlation removal, around 30% of improvement is obtained on the polyester case study. Moreover, the effect of process dynamics on variable selection is examined with applications of a digester simulator and industrial Kamyr digester case studies. Due to the dynamic nature of the process, the selection performances are not as consistent. Therefore, a new variable selection technique is needed. The performances of different variable selection methods are compared and their advantages and disadvantages are discussed with the aim 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. My gratitude also extends to Dr. W. Robert Ashurst, Dr. Steve R. Duke, and Dr. Li for their willingness to serve on my committee and their extraordinary supports, as well as my University Reader, Dr. Dmitry Glotov. I would also like to thank Dr. Mario R. Eden 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, Andrew Damiani, Kyle Stone, and Dr. Gongxian Xu. Also special thanks to Hector Galicia for his endless guidance, assistance and support. He has been very instrumental in successful completion of this work.

Friends have always been important to me, and thankfully during my journey great friendships were always there. I would like to thank my friends: Arianna Tieppo, Pengfei Zhao, Achintya Sujan, Xinquan Cheng, and Yulin Jin. They have brought laughter, happiness, and comfort to not only this journey but my life.

My love and gratitude go to my parents, Haibin Wang and Yanyun 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, Yudong (Jeffrey) Wang, for his support and understanding. Their love and encouragement have always been a constant source of comfort and support to me when times were difficult.

Last but not least, I dedicate this dissertation to Jesus Christ and church family. Praise, thanks and glory go to my savior Jesus Christ, whom I have met during my study in Auburn. Thanks to my church families, Susan Pan, Hanqin Tian, Kai Chang, Ke Liu, and Ting Zhang, for their supports throughout this process and words of wisdom.

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
CBP	Correlation between predictors
CSTR	Continuous stirred tank reactor
DCS	Distributed control system
DPLS	Dynamic partial least squares
EDF	Exponential decreasing function
G	Geometric mean of sensitivity and specificity
GA	Genetic algorithm
GAVDS	Genetic algorithm-based process variables and dynamics selection
I_k	Number of samples in k^{th} batch
J	Number of variables in batch
K	Number of batches, number of MC runs
k	Reciprocal of signal to noise ratio
KRR	Kernel ridge regression
Lasso	Least absolute shrinkage and selection operator

MAPE	Mean absolute percentage error
MLR	Multiple linear regression
MC	Monte Carlos
MS_E	Mean square error
N	Total number of simulation runs
nb	Normalized regression coefficient
NIR	Near infrared
n_u	Number of samples after unfolding the three-array batches
p	Number of predictors
PCA	Principal component analysis
PCR	Principal component regression
PLS	Partial least squares
PLSLDA	Partial least squares linear discriminant analysis
PR	Proportion of relevance
RIVAL	Removing irrelevant variables amidst Lasso iteration
RMSEP	Root mean square error of prediction
RV	Random variable added in UVE-PLS
SNR	Signal to noise ratio
SPA	Statistics pattern analysis or successive projection algorithm
SR	Stepwise regression
SRC	Structure of regression coefficients
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
β	Coefficients of predictors
Γ	Variance-covariance matrix
ϵ	Normal distributed random noise
λ	Positive regularization parameter in RIVAL
ρ	Magnitude of correlation between predictors
σ	Standard deviation of error, ϵ
Ω	Variable subset
ω	Non-negative weighting vector

Chapter 1. Introduction

Due to advancement of technology, tremendous amount of process measurements are collected and stored every day. These data have been used to build data-driven soft sensors [1]–[4]. Soft sensors are mathematical models that relate primary variables with the secondary variables. By correlating the secondary variables with the primary variables, one application of soft sensor is to provide information on those hard-to-measure, but important variables, such as product quality [3]. Another application of soft sensor is to provide prediction on infrequently measured process variables so that prompt control actions can be taken [4]. It has been shown by many studies that the performance of these data-driven schemes can be tremendously improved by selecting only the vital variables that strongly affect the primary variables, rather than all the available process variables [5], [6], even though it has not been studied what factors would determine the level of improvement in soft sensor performance with variable selection. By identifying the relevant variables, variable selection can improve the prediction performance of soft sensor, reduce the model complexity and computational load, obtain better insight into the nature of the process, and lower the cost of measurements [5], [6]. Variable selection has been one of the most important practical concerns in data-driven approaches.

In the past few decades, many different variable selection approaches have been reported for various applications with different soft sensor modeling methods. In general, variable selection techniques can be categorized into three groups: filter, wrapper, and embedded approaches [5], [7]–[11]. Filter approaches can simply be viewed as variable ranking, and it is independent from the learning machines. Compared to the simplicity of filter methods, wrapper and embedded methods are more complex and closely related to each other. Wrapper methods wrap around an appropriate learning machine, which is employed as the evaluation criterion, such as prediction or classification error. Wrapper methods are proven to outperform filter methods [5], [7]–[11]. The embedded methods are similar to wrapper methods, except that the variable selection is performed simultaneously with the training process. Since the main focus of this work is to improve the performances of partial least squares (PLS) based models, only wrapper approaches are investigated. The following seven variable selection methods are explored and compared in this work: PLS based on variable importance in projection (PLS-VIP) [12], [13], PLS with regression coefficients (PLS-BETA) [12], genetic algorithm combined with PLS (GA-PLS) [14]–[16], uninformative variable elimination combined with PLS (UVE-PLS) [17]–[19], stepwise regression (SR) [20], [21], competitive adaptive reweighted sampling method with PLS (CARS-PLS) [22], and least absolute shrinkage and selection operator (Lasso) (removing irrelevant variables amidst Lasso iterations (RIVAL), the improved version of Lasso, will be used in the simulation case study) [23]–[26]. Even though PLS-VIP and PLS-BETA are more of ranking techniques, both methods are tuned to optimize the prediction performance. Therefore, they can be considered as wrapper methods.

Stepwise regression has been applied to the selection of predictors for both classification and multivariate calibrations [21], especially in near-infrared (NIR) spectra. Gauchi and Chagnon proposed a stepwise variable selection method based on maximum Q^2, prediction ability criterion, and applied to manufacturing processes in oil, chemical and food industries [27].

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 [15]. 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 [28]. A modified genetic algorithm-based wavelength selection method has been proposed by Hiromasa Kaneko and Kimito Funatsu to select process variables and dynamics simultaneously [29]. This method is called 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 applications of GA in chemical structure handling and molecular recognition [30].

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. [19]. 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 [19], [31]. In the work of Koshoubu et al., the authors have extended UVE to eliminate uninformative samples (USE) that do not contribute much in the calibration model [32], [33]. They proposed an algorithm in which 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 in [34]. UVE is implemented to remove uninformative variables before application of SPA to improve the efficiency of variable selection by SPA.

Least absolute shrinkage and selection operator (Lasso) has been applied in many areas, such as for genomic selection [35], nonlinear system identification [36], Chemometrics data analysis [6], and sparse modeling [37], etc.

Competitive adaptive reweighted sampling (CARS) method has been proposed by Li et al. [22]. CARS is model independent, i.e., CARS can be combined with any regression or classification models. In [38], [39], CARS has been applied in combination with partial least squares linear discriminant analysis (PLSLDA) to effectively identify two classes of samples in colorectal cancer data.

Variable importance in the projection (VIP) and regression coefficients (BETA) have been broadly adopted 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 [40]. 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 [41]. 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. [13]. 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 employed in food science. Andersen and Bro applied PLS-VIP and PLS-BETA to NIR spectra of beer sample and obtained useful insight of the process, by identifying the important variables [6]. A variable selection algorithm based on the standardized regression coefficients are proposed in [42]. The developed models are optimized by the leave-one-out Q^2 values and validated by an external testing set. It is worth noting that there are many more variable selection methods in the literature.

In this work, we use static case studies (one simulated and one industrial polyester case study) and dynamic case studies (one digester simulator and one industrial Kamyr digester) to evaluate the properties of the variable selection methods. A new metric is proposed to assess the performance of different variable selection methods when the ground truth of the data is unknown. 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 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 by considering four factors: proportion of relevant predictors (PR), magnitude of correlation between predictors (CBP), magnitude of signal to noise ratio (SNR), and structure of regression coefficients (CBP) [12]. In addition, the algorithms are applied to an industrial polyester soft sensor case study. In both cases, independent test sets are used to provide fair comparison and analysis of different algorithms. The soft sensor prediction performance of models developed by these variable selection methods are compared using PLS. Furthermore, a digester simulator [43] and an industrial Kamyr [3] digester case will be utilized to further inspect the effect of process dynamics on variable selection. The overall 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 is 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), Least absolute shrinkage and selection operator (Lasso), 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). Chapter 4 introduces the performance indicators used in this work to evaluate the different variable selection methods. In Chapter 5, application of all seven variable selection methods on simulated case study and industrial polyester case study are investigated. Detailed descriptions of simulation data generation and industrial polyester data are provided. The industrial case study is focused on the process data of polyester resin production plant. A brief specification of the plant is included, followed by discussion of characteristics of batch process. The results and comparison of variable selections on both simulated and industrial case studies are discussed. Applications on digester case studies are presented in Chapter 6. Brief descriptions of the case studies are included. Variable selection is with dynamic models for this process. Chapter 7 concludes this work with major discussion and contributions. Furthermore, suggestions on future works are provided.

Chapter 2. Soft Sensor Development

Soft sensors, mathematical models that correlates primary variables with secondary variables, 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 [1], [44]. 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 [1], [2], [44], [45]. Data-driven soft sensors can also be viewed as mathematical models that correlate the secondary measurements to the primary measurements. The most popular soft sensor techniques include principal component analysis (PCA) [46] and partial least squares (PLS) [47], artificial neural networks (ANN) [48], neuro-fuzzy (NF) systems [49] and support vector machines (SVM) [50]. In our work, only the linear models are considered.

2.1 Multiple linear regression (MLR)

The goal of multiple linear regressions (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 collected over time, β_j is the sensitivity or coefficients of secondary variable j, 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:

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

2.2 Principal component analysis (PCA)

Principal component analysis (PCA) is linear technique that transforms the original data matrix, $\boldsymbol{X}$, into a smaller set of uncorrelated variables, $\boldsymbol{T}$, that would capture most of the information in the original space. This linear transformation can be expressed as in Equation (2.3), where $\boldsymbol{T}$ is the score matrix, $\boldsymbol{P}$ is the loading matrix, and $\boldsymbol{W}^*$ is the weighted loading matrix scaled by weighting matrix $\mathbf{W}$. $\boldsymbol{E}$ is the general term of model residual. The decomposition is done in such a way that the covariance between the original variables is maximized.

$$\begin{gathered}\boldsymbol{X} = \boldsymbol{TP}' + \boldsymbol{E} \\ \boldsymbol{T} = \boldsymbol{XW}^* \\ \boldsymbol{W}^* = \boldsymbol{W}(\boldsymbol{P}'\boldsymbol{W})^{-1}\end{gathered} \tag{2.3}$$

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. This gives the expression for PCR as shown in Equation (2.4).

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

However, the disadvantage of PCR is that the components may not be good at explaining the primary variables.

2.4 Partial least squares regression (PLS)

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 [51]. In PLS, the decomposition of matrix $\boldsymbol{X}$ and $\boldsymbol{Y}$ are done in such a way that the covariance is maximized. The algorithm of PLS were developed by Wold et al. [52]. The decomposition of data matrix $\boldsymbol{X}$ is done by Equation (2.3). And the decomposition of $\boldsymbol{Y}$ can also be done in a similar way by Equation (2.5), where $\boldsymbol{U}$ and $\boldsymbol{Q}$ is the score and loading matrices of $\boldsymbol{Y}$, respectively, and $\boldsymbol{F}$ is the residual.

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}' + \boldsymbol{F} \tag{2.5}$$

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

$$\boldsymbol{U} = \boldsymbol{T}\boldsymbol{B} \tag{2.6}$$

Consequently, matrix $\boldsymbol{Y}$ can be estimated as in Equation (2.7), in which $\boldsymbol{F}$ is to be minimized. The detail algorithm of PLS can be found in [51]–[53].

$$\widehat{\boldsymbol{Y}} = \boldsymbol{T}\boldsymbol{B}\boldsymbol{Q}' + \boldsymbol{F} \tag{2.7}$$

PLS is not only to establish the maximum variance of the secondary variables, but also to maximize the variability of the primary variables, explained by the correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$. When the original variables are highly correlated, redundant, noisy, and of high dimensionality, the orthogonal scores can be obtained through decompositions of $\boldsymbol{X}$ and $\boldsymbol{Y}$, which would contain sufficient information on $\boldsymbol{X}$ and predictive information. In other words, it removes the correlation, noise, etc., between the original variables by projection and dimension reduction. PLS models are more stable than the models built upon the original variables, since the regression is done on the scores instead of the original variables. In this work, nonlinear iterative partial least squares (NIPALS) is used to implement PLS.

Chapter 3. Variable Selection Theory and Algorithm

Due to prompt development of various technologies, thousands of process measurements are collected and stored by process computers every day. Researchers have been utilizing these data to build soft sensors, which are also known as data-driven soft sensors [1]–[4]. By correlating the secondary variables with the primary variables, sensors can provide information on those hard-to-measure or immeasurable, but important variables, such as product quality [3]. Furthermore, soft sensors can provide prediction on infrequently measured process variables so that prompt control actions can be taken to prevent process failure [4]. It has been proved by many studies that the performance of these data-driven schemes can be tremendously improved by selecting only the vital variables that strongly affect the primary variables, rather than all the available process variables [5], [6], even though it has not been studied what factors would determine the level of improvement in soft sensor performance with variable selection. 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 model complexity and computational load, provide better insight into the nature of the process, and lower the cost of measurements [5], [6].

In the past few decades, many different variable selection approaches have been reported for various applications with different soft sensor modeling methods. In general, variable selection techniques can be categorized into three groups: filter, wrapper, and embedded approaches [5], [7]–[11]. Filter approaches are independent from the learning machines, in which variables are only ranked based only chosen criterion. Selection is solely done according to the ranking of the variables. Compared to the simplicity of filter methods, wrapper and embedded methods are more complex and more closely related to each other. Wrapper methods, differ from filter methods, wrap around an appropriate learning machine, which is employed as the evaluation criterion, such as prediction or classification error. Wrapper methods are proven to outperform filter methods [5], [7]–[11] given the optimal goal to improve the performance of soft sensors. The embedded methods are similar to wrapper methods, except that the variable selection is performed simultaneously with the training process. In other words, the central learning algorithm is updated together with variable selection, which leads to more complex computation. Since the main focus of this work is to improve the performances of partial least squares (PLS) based models, only wrapper approaches are investigated. The following seven variable selection methods are explored and compared in this work: PLS based on variable importance in projection (PLS-VIP) [12], [13], PLS with regression coefficients (PLS-BETA) [12], genetic algorithm combined with PLS (GA-PLS) [14]–[16], uninformative variable elimination combined with PLS (UVE-PLS) [17]–[19], stepwise regression (SR) [20], [21], competitive adaptive reweighted sampling method with PLS (CARS-PLS) [22], and least absolute shrinkage and selection operator (Lasso) (removing irrelevant variables amidst Lasso iterations (RIVAL), the improved version of Lasso, will be used in the simulation case study) [23]–[26]. Even though PLS-VIP and PLS-BETA are more of ranking techniques, both methods are tuned to optimize the prediction performance of the soft sensors. Therefore, they can be considered as wrapper methods. In this chapter, we will review the seven variable selection methods with their strengths and limitations summarized in Table 3.1.

3.1 Partial least squares with variable important in projection (PLS-VIP)

Variable importance in the projection (VIP) score estimates the importance of each variable in the projection used in a PLS model. It was first introduced in [54]. The VIP score for the j^{th} variable, in PLS model with A principal components, can be calculated using Equation (3.1).

$$VIP_j = \sqrt{p\sum_{a=1}^{A}\left(SS(q_a t_a)\left(\frac{w_{ja}}{\|w_a\|}\right)^2\right) / \sum_{a=1}^{A} SS(q_a t_a)} \tag{3.1}$$

where $SS(q_a t_a) = q_a^2 t_a' t_a$. t_a is the a^{th} column vector of score matrix $\boldsymbol{T}$. q_a is the a^{th} element of regression coefficient vector q (column vector of matrix $\boldsymbol{BQ'}$) of $\boldsymbol{T}$. w_a is the a^{th} column vector of weighting matrix $\boldsymbol{W}$, which gives the weighted variability of j^{th} variable in the retained dimensions. p is the number of variables in regression matrix $\boldsymbol{X}$. VIP score calculates the contribution of each variable according to variance explained by each PLS component [13]. The expression $w_{ja}/\|w_a\|$ represents the importance of j^{th} variable in the a^{th} PLS component. The $SS(q_a t_a)$ is the variance of y explained by the a^{th} PLS component. And the summation of $SS(q_a t_a)$, the denominator term, is the total variance explained by the PLS model with A components.

A variable selection method based on VIP scores estimated by PLS regression model is known as PLS-VIP. The ‘greater than one rule’ is conventionally used as criterion for variable selection. According to this rule, only variables with VIP values greater than one are considered significant. Overall PLS-VIP procedure can be described as follows:

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 (3.1).
3. Select variables with VIP scores greater than the cutoff value.
4. Rebuild PLS model with only the retained variables.
5. Evaluate the model performance using various performance indexes.

3.2 Partial least squares with regression coefficients (PLS-BETA)

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. PLS-BETA directly 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.

3.3 Uninformative variable elimination with PLS (UVE-PLS)

A method for eliminating uninformative variables by comparing with artificial variables was proposed by V. Center et al. [17]. Models are built using both experimental and artificial variables. The significance of each variable is assessed by comparing its reliability index, a function of the regression coefficients, with reliability indices of artificial random variables.

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

1. For a given set of experimental variables with n number of samples, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, generate an artificial random variable matrix, $\boldsymbol{R}$, with very

Table 3.1: Strengths and limitations of each variable selection method

Models	Strengths	Limitations
PLS-VIP	• Selection consistency, insensitive to training data selection • High prediction performance, results reflect process knowledge • Simple implementation, cheap computation • Only one parameter to tune; general guideline available	• Affected by correlation • Somewhat sensitive to tuning parameter
PLS-BETA	• Insensitive to training data selection • Simple implementation, cheap computation, and only one parameter to tune	• Sensitive to tuning parameter • Directly affected by the magnitude of contribution
Lasso/RIVAL	• Explicitly penalize the size of the model, proved set consistency for RIVAL	• Highly sensitive to tuning parameter • Expensive computation • Require large dataset to reach convergence
UVE-PLS	• Straightforward algorithm • Insensitive to tuning parameter	• Performance strongly affected by magnitude of correlation
SR	• Easy interpretation between the results and tuning parameters	• May be trapped in local optima • Degrade noticeably with increased collinearity
CARS-PLS	• Can easily control the percentage of variables to be retained; could be used as a pre-selection for high dimension data	• Sensitive to training data selection
GA-PLS	• Could escape from local optima due to randomized search • Handles problems with multiple objectives	• Global optima not guaranteed • Requires a lot of user inputs to optimize the tuning parameters • Computation expensive due to evaluation of fitness function, which depends on the population size

small magnitude and same dimension as the experimental variables. This results in a matrix with dimension of n by $2p$, $\boldsymbol{XR} = [\boldsymbol{X} \quad \boldsymbol{R}]$.

2. Build PLS model for $\boldsymbol{XR}$ based on leave-one-out procedure. This will yield a regression coefficient matrix, $\boldsymbol{B} \in \mathbb{R}^{n\times 2p}$.
3. Calculate the reliability index of each variable j using Equation (3.2),

$$c_j = \frac{m(b_j)}{s(b_j)} \tag{3.2}$$

$$m(b_j) = \frac{\sum_{i=1}^{n} b_{ij}}{n} \tag{3.3}$$

$$s(b_j) = \left(\frac{\sum_{i=1}^{n}\left(b_{ij} - m(b_j)\right)^2}{n-1}\right)^{1/2} \tag{3.4}$$

Where c_j is the reliability index of variable j, $m(b_j)$ and $s(b_j)$ are the mean and standard deviation of regression coefficient of variable j obtained from leave-one-out procedure, b_{ij}.

4. Determine the maximum absolute reliability index of the artificial variables, $|\max(\mathrm{c}_{\mathrm{artif}})|$. The cutoff threshold is defined as $cut_{uve} = k \times |\max(\mathrm{c}_{\mathrm{artif}})|$, where k controls the role of reliability of artificial variables. The experimental variables with absolute reliability index less than that are eliminated, i.e., $|c_j| < cut_{uve}$.
5. A new PLS model is built using only the remaining variables.

3.4 Genetic algorithm with PLS (GA-PLS)

Genetic algorithm has been used widely in solving complex problems of optimization and search problems [55]. 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 [15].

The original algorithm can be found in [56]–[58]. Generally speaking, there are five steps in GA: coding of variables, initiation of population, evaluation of the responses, reproductions, and mutations [14]. 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 (10%-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 criteria.
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. Only the remained 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 a certain percentage of the total variables. This is repeated multiple times depending on the population size. A

PLS model is built for each population/chromosome. Populations are then sorted in descending order by its cross validation metrics. In this work, we use root mean square error (RMSE) as 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 the lack of prediction improvement.

3.5 Competitive adaptive reweighted sampling with PLS (CARS-PLS)

Hongdong Li et al. have proposed a novel strategy based on the 'survival of the fittest' principle, called competitive adaptive reweighted sampling (CARS) [22], [38]. This method utilizes the absolute values of the regression coefficients to evaluate the importance of variables. In an iterative manner, K subsets of variables are selected by CARS from K Monte Carlo (MC) sampling runs. At the end, cross validation is employed to evaluate each subset. The general procedure can be described as follows:

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 (3.5), where p is the total number of variables.

$$nb_j = \frac{|b_j|}{\sum_{j=1}^{p}|b_j|} \tag{3.5}$$

2. In CARS, an exponentially decreasing function (EDF) is introduced as in Equation (3.6). EDF is utilized to eliminate variables with relatively small absolute regression coefficients by force. The ratio of variables to be re-

tained in the i^{th} sampling run, r_i, is calculated by EDF shown in Equation (3.6) to (3.8),

$$r_i = de^{-hi} \tag{3.6}$$

$$d = \left(\frac{p}{2}\right)^{\frac{1}{K-1}} \tag{3.7}$$

$$h = \frac{\ln(p-1)}{K-1} \tag{3.8}$$

where constants d and h are determined so that $r_1 = 1$ and $r_K = 2/p$.

3. Following EDF-based reduction, adaptive reweighted sampling (ARS) is implemented in each subset of variables to further eliminate variables in a competitive way. In other words, variables with larger regression coefficients will be selected with higher frequency.

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 3.1. Hence, EDF becomes a very efficient algorithm for removing the variables with little information.

ARS in Step 3 mimics the ‘survival of the fittest’ principle. The idea of ARS is illustrated in Figure 3.2. Three scenarios are considered: equal weight, little weight difference, and large weight difference. As a result, the variables with larger weights are selected with higher frequency.

3.6 Least absolute shrinkage and selection operator (Lasso)

The objective of least absolute shrinkage and selection operator (Lasso) is to minimize the residual sum of squares subject to the sum of the absolute values of the coefficients being less than a constant, t [12], [23], [25], which directly controls the number of variables being selected. Lasso is similar to Ridge regression [12], the regression coefficients are shrunk by placing a penalty on their size. It can be mathematically expressed in the following two ways.

$$J(\beta) = \arg\min_{\beta} \|y_n - \boldsymbol{X}_n \beta\|^2 \tag{3.9}$$

$$subject\ to \sum_{j=1}^{p} |\beta_j| \leq t$$

$$J(\beta) = \arg\min_{\beta} \|y_n - \boldsymbol{X}_n \beta\|^2 + \lambda(n) \sum_{j=1}^{p} |\beta_j| \tag{3.10}$$

where β is the coefficients of the predictors. Equation (3.10) is the Lagrangian relaxation of Equation (3.9). This can be solved by the standard quadratic programming with linear

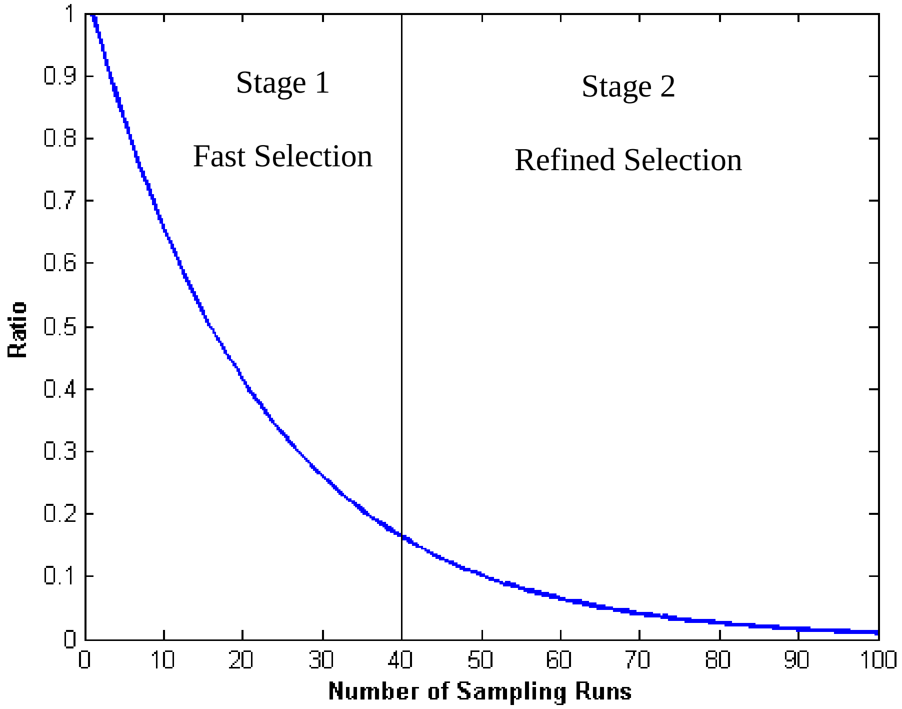
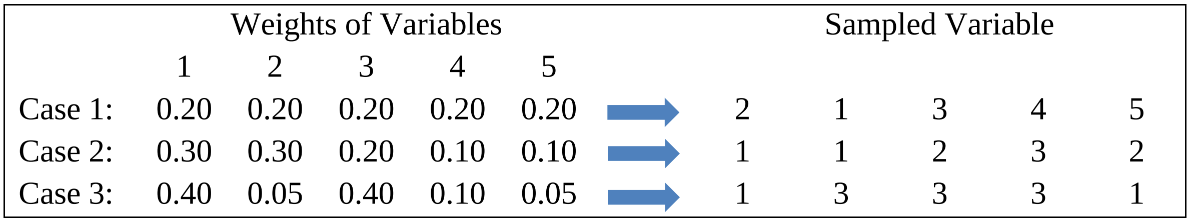


Figure 3.1: Graphical illustration of the exponentially decreasing function

inequality constraints. The use of least angle regression (LARS) algorithm can reduce the computation burden.



	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 3.2: 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.

An improved version Lasso, removing irrelevant variables amidst Lasso iterations (RIVAL), was proposed by Kump et al. [24]–[26]. In RIVAL, Lasso is modified by incorporating a priori information that all the regression coefficients are greater than zero into this minimization problem, resulting in the positive Lasso with a penalty term.

$$J(\beta) = \underset{\beta \geq 0}{\arg\min} \|y_n - \boldsymbol{X}_n \beta\|^2 + \lambda(n) \sum_{j=1}^{p} \omega_j \beta_j \tag{3.11}$$

where $\lambda(n)$ is the positive regularization parameter that depends on the number of data points n, and ω is a non-negative weighting vector.

In positive Lasso, the set consistency, i.e., all the variables are identified correctly, and parameter consistency, i.e., the magnitudes of all the regression coefficients are identified correctly, can only be met when the number of data points approaches infinity. The new algorithm, RIVAL, ensures the set consistency for a large but fixed/finite number of data. The idea of RIVAL is to shrink the coefficients toward zero as $\lambda(n)$ increases. Therefore, the selection of $\lambda(n)$ is critical to the performance of positive Lasso. In this work, RIVAL is only applied to the simulation case study due to its positivity constraints.

3.7 Stepwise regression (SR)

Stepwise regression has been widely used for variable selection in linear regression [20]. Stepwise regression is a combination of forward selection and backward elimination methods [27]. 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, determined by partial F-test or t-test, in a regression [21].

The standard stepwise regression procedure is 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 $\Omega_1 = \{x_1, \quad x_2, \quad \ldots, \quad x_k\}$ is a subset of variables included in linear regression model. The unselected variables are examined by calculating their partial F-statistic using Equations (3.12) and (3.13), where SSR is the sum of squared residuals due to regression, and MSE is the mean square error. The variable with maximum F-statistic among all the unselected ones is added to the model, provided that $F_j > F_{in}$.

$$F_j = \frac{SSR\left(x_j \middle| x_1, x_2, \ldots, x_k\right)}{MSE\left(x_j, x_1, x_2, \ldots, x_k\right)} \tag{3.12}$$

$$SSR(x_j|x_1,\dots,x_k) = SSR(x_j,x_1,\dots,x_k) - SSR(x_1,\dots,x_k) \qquad (3.13)$$

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 removed. 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.

Chapter 4. Performance Indicators of Variable Selection Methods

In order to evaluate the performance of different variable selection methods, several performance indices have been proposed in the literature. The most common ones are the average mean absolute percentage error (MAPE), coefficient of determination (R^2), and geometric mean of selection sensitivity and specificity (G). Among them, only G directly measures the accuracy of variable selection results while MAPE and R^2 indirectly measure the effects of variable selection through the prediction performance of a soft sensor, such as PLS.

4.1 Mean absolute percentage error (MAPE)

To evaluate model prediction performance, MAPE is commonly used, which is defined as follows.

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|\% \tag{4.1}$$

where y_i is the true measurement and $\hat{y}_i$ is the prediction.

4.2 Coefficient of determination ($\boldsymbol{R^2}$)

R^2 measures how well the data fits the model, as calculated in Equation (5.9

$$R^2 = 1 - \frac{SSR}{SST} \tag{4.2}$$

where $SSR = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ is the sum of squared residual, and $SST = \sum_{i=1}^{N}(y_i - \bar{y})^2$ is the total sum of squares, with $\bar{y}$ as the average of y.

4.3 Geometric mean of sensitivity and specificity (G)

When the information on the true relevant variable is available, confusion matrix, as shown in Table 4.1, can be used to evaluate the variable selection performance. From the confusion matrix, accuracy, sensitivity, and specificity of variable selection can be calculated as follows.

$$Accuracy = (a + d)/(a + b + c + d) \tag{4.3}$$

$$Sensitivity = d/(c + d) \tag{4.4}$$

$$Specificity = a/(a + b) \tag{4.5}$$

In this work, the geometric mean of sensitivity and specificity, G, i.e., Equation (4.6), is used as an overall variable selection performance indicator.

$$G = (Sensitivity\ \times Specificity)^{1/2} \tag{4.6}$$

The value of G ranges from 0 to 1. $G = 1$ indicates that all the predictors are classified correctly. This index is in general only applicable in simulation case studies, where the ground truth of variable relevancy is known.

Table 4.1: Confusion matrix

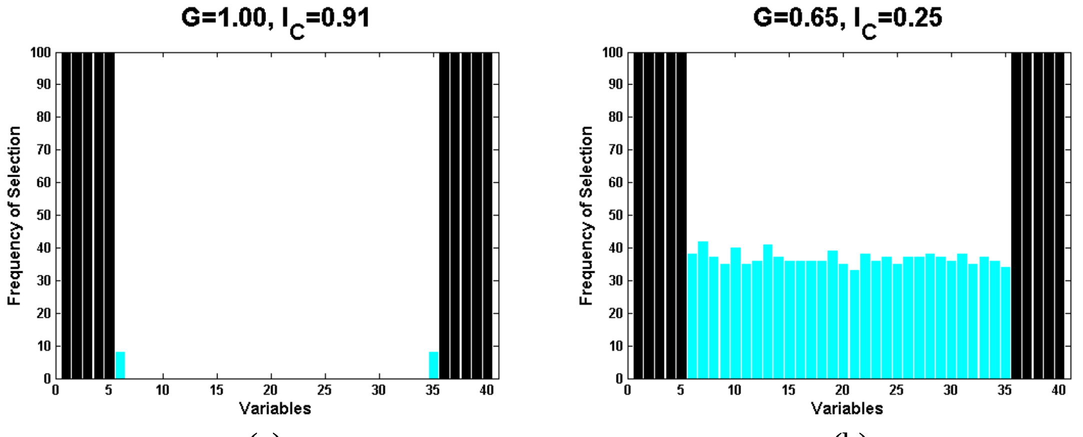
		Predicted classes	
		Irrelevant predictor (IR)	Relevant predictor (R)
True classes	Irrelevant predictor (IR)	a: the number of irrelevant predictors classified correctly	b: the number of irrelevant predictors classified incorrectly
	Relevant predictor (R)	c: the number of relevant predictors classified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

4.4 The proposed indicator: consistency index (I_C)

When the information on the true relevant variables is not available, which is the case for most industrial applications, G cannot be obtained. For such cases, how consistent the variable selection results from different MC runs (i.e., different training data) can provide important information on the robustness and consistency of the variable selection method. However, there is no such metric exists in the literature. Therefore, in this work we propose an entropy based consistency index, named I_C, as defined below.

$$I_C = 1 - \frac{-e \sum_{j=1}^{p} prob(x_j) \ln\left(prob(x_j)\right)}{m} \tag{4.7}$$

where m is the total number of variables being selected among all MC runs (100 in this work); e is the Euler's number; and $prob(x_j)$ is the probability of variable x_j being selected, which can be approximated by the percentage of x_j being selected among all MC runs. Note that $-\sum_{j=1}^{p} prob(x_j) \ln\left(prob(x_j)\right)$ is exactly the information entropy associated with the probability for x_j being selected, and multiplying e scales the maximum entropy to 1. As a result, I_C ranges between 0 and 1. A simple example is used here to illustrate the computation of I_C based on a selection frequency plot, which is a bar chart showing the frequency of each variable being selected among all MC runs. In this example, only the 10 variables at the extreme locations (i.e., variables 1-5 and 36-40) are the true relevant variables among all 40 variables. The selection frequency plots of two variable selection methods, PLS-VIP and CARS-PLS are shown in Figure 4.1 (a) and (b) respectively, where the frequencies of the true relevant variables are shown as the dark-colored bars, while those of the irrelevant variables are shown as the light-colored bars. By approximating the probabilities in Equation (5.9) with the selection frequencies, we can compute the I_C values for both cases. As a comparison, G values are also computed. In Figure 4.1 (a), PLS-VIP selects all the relevant variables correctly all the time, with very few errors of selecting irrelevant variables, which is correctly assessed by the high values in both G and I_C. In the case of CARS-PLS shown in Figure 4.1 (b), although CARS-PLS also selects the 10 relevant variables all the time; it also selects all the 30 irrelevant ones about 35% of the time. This means that, on average, there are more than twenty variables selected each time with only about half of them (i.e., 10) truly relevant. Given this poor performance, CARS-PLS still achieves moderate G value of 0.65. On the other hand, the I_C value of CARS-PLS is 0.25, which seems to be a better reflection of its poor selection performance. It is worth noting that I_C is applicable not only for simulated cases where relevant variables are known, such as the one illustrated above, but also for the industrial cases where the exact relevant variables are often unknown. It is also worth noting that I_C only evaluates the consistency of variables being selected when different training data are used. If the knowledge on the true relevant variable is available, that knowledge is not utilized by I_C.



(a) (b)

Figure 4.1: Comparison of I_C and G for (a) PLS-VIP and (b) CARS-PLS. The dark-colored bars represent true relevant variables, while the light-colored bars representing true irrelevant variables

Chapter 5. Variable Selection Methods with Their Applications to Simulated and Industrial Polyester Datasets

5.1 Introduction

The performances of different variable selection methods are compared using a simulation case study and an industrial polyester case study. In addition, the sensitivities of variable selection methods to its tuning parameters are examined. In both case studies, 100 MC simulations are performed to generate different sets of training, tuning, and test data. Three steps are followed to carry out the comparison. First, models are built from training set with suggested range of tuning parameters of each variable selection method. A summary of the tuning parameters and their search range is listed in Table 5.1. Next, each model is optimized using independent tuning datasets by minimizing the average MAPE. Finally, the optimized model is applied to the testing set and used for performance comparison.

5.2 Simulated case study

The simulation case study introduced in [12] is used in this work. The dataset is generated to mimic typical characteristics of industrial data by considering four factors: proportion of relevant predictors (PR), magnitude of correlations between predictors (CBP), magnitude of signal to noise ratio (SNR), and structure of regression coefficients (SRC). The dataset is generated following a linear model as defined in Equation (5.1).

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \tag{5.1}$$

where ε_i is a normal distributed random error sequence with zero mean and specified standard deviation, σ.

Table 5.1: Summary of tuning parameters and search range

Methods	Tuning Parameters	Range
PLS-VIP	VIP Score	0.01:0.01:3
PLS-BETA	Regression Coefficients	0.01:0.01:1
UVE-PLS	Reliability Cutoff	0.5:0.05:1
SR	Confidence Limit	85, 90, 95, 99%
Lasso	Regression Coefficients	0.05:0.05:5
RIVAL	Regularization Parameter, λ	0.01:0.01:10
GAPLS	Maximum Generation	50:50:300
	Population Size	32:32:160
	Initial Percentage Included	10:10:50
	Mutation Rate	0.0025:0.0025:0.01
	Crossover Rule	Double or Single

5.2.1 Data descriptions

A total of 108 different cases are designed by considering all the possible combinations of the four factors. For each case, the data matrix ***X*** of 1500 sample points is generated. The data matrix is first randomly permuted sample-wise. Then the first 500 samples are used for training, the second 500 samples are for tuning and the last 500 for testing. This MC procedure is carried out 100 times to generate 100 different sets of training, tuning, and testing datasets.

- For convenience, the number of relevant predictors is set to be 10. Different proportion of relevant predictors are achieved by varying the total number of predictors, p, in three levels, 20, 40 and 100, as shown in Equation (5.2), which result in PR in three levels of 0.5, 0.25, and 0.1, respectively.

$$PR = \frac{10}{p} \tag{5.2}$$

- Data matrix $\boldsymbol{X}$ is generated from multivariate normal distribution with zero mean vector and variance-covariance matrix of $\boldsymbol{\Gamma}$, which is defined in Equation (5.3). The elements of matrix $\boldsymbol{\Gamma}$ are function of the magnitude of correlations between predictors (CBP), ρ, which is varied in three levels, 0.5, 0.7 and 0.9.

$$\boldsymbol{\Gamma}_{ij} = \rho^{|i-j|}, \qquad (i, j = 1, 2, \dots, p) \tag{5.3}$$

- The magnitude of signal to noise ratio (SNR) is introduced by manipulating σ, the standard deviation of error terms in y, as defined in Equation (5.4).

$$\sigma = k\sqrt{var(\boldsymbol{X}\beta)} \tag{5.4}$$

where k is the reciprocal of signal to noise ratio, varied in three levels, 0.33, 0.74, 1.22, which results in SNR of 3.03, 1.35, and 0.82, respectively.

- The structure of regression coefficients (SRC) is also considered. 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

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

    - Equal coefficients at the extreme

$$\beta_j = 1, \qquad (j = 1,2,\dots,5, p-4, p-3,\dots,p) \tag{5.6}$$

- Unequal coefficients in the middle of range

$$\beta_j = (5.5 - |j - 0.5(p+1)|)^2, \qquad \left(j = \frac{p}{2} - 4, \frac{p}{2} - 3, \dots \frac{p}{2} + 5\right) \tag{5.7}$$

- Unequal coefficients at the extreme

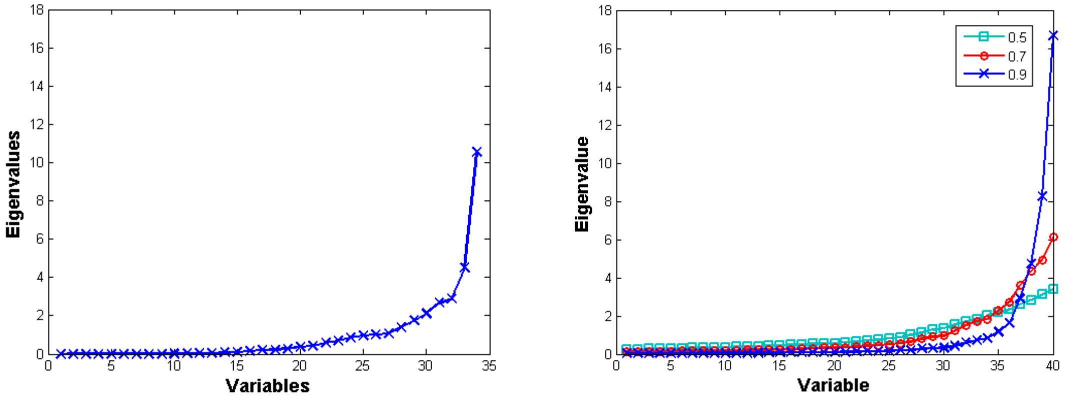
$$\beta_j = \left(|j - 0.5(p+1)| - 0.5(p-11)\right)^2, \qquad (j = 1,2,\dots,5, p-4, p-3,\dots,p) \tag{5.8}$$

5.2.2 Comparison results

Seven methods, PLS-VIP, PLS-BETA, RIVAL, UVE-PLS, SR, CARS-PLS, and GA-PLS, are compared on all 108 cases. Only six representative cases listed in Table 5.2 are presented here. To better mimic industrial data, we compare the patterns of the eigenvalues of the covariance matrix of the simulated data to the industrial polyester data used in the next section. As can be seen from Figure 5.1, when correlation between predictors is high (e.g., 0.9 in this case), the simulated data mimics the industrial data better, which is also observed in [12]. Therefore, more cases with $\rho = 0.9$ (i.e., $CBP = 0.9$) are selected. The results of the other cases are presented in Appendix A.

Table 5.2: Properties of simulation cases. PR, CBP, SNR, and SRC stand for proportion of relevance, correlation between predictors, signal to noise ratio, and structure of regression coefficients, respectively.

Cases	1	2	3	4	5	6
PR	0.25	0.25	0.25	0.5	0.1	0.1
CBP	0.5	0.9	0.9	0.9	0.5	0.7
SNR	3.03	3.03	1.35	0.82	1.35	0.82
SRC	EE	UE	EM	UE	UM	EM



(a) (b)

Figure 5.1: Comparison of eigenvalues between (a) polyester data and (b) simulation data with three different correlations between predictors (CBP) levels

First, the sensitivities of each method with respect to tuning parameters are investigated. In summary, PLS-BETA and RIVAL are the most sensitive; PLS-VIP is somewhat sensitive, while the rest of the methods are relatively insensitive to their respective tuning parameters. The selected sensitivity results of PLS-VIP, PLS-BETA, RIVAL, and UVE-PLS, in terms of change in prediction error in the testing sets with respect to tuning parameter, are shown in Figure 5.2. UVE-PLS is included as an example of insensitive methods to the tuning parameters. Even though the pattern for PLS-VIP and PLS-BETA are similar, there are some key differences. First, the optimal range in PLS-VIP is much wider than PLS-BETA. In addition, "greater than one rule" also falls in the region where the minimum occurs. On the other hand, there is no general guideline for tuning parameter in PLS-BETA. The similarity between PLS-VIP and PLS-BETA is that when the cutoff value is set too high, only a few variables are retained, which leads to high prediction error. However, the "greater than one rule" of PLS-VIP provides good starting point for tuning parameter optimization. The performance of RIVAL is a lot more sensitive to the tuning parameter than PLS-BETA, and a convergent solution is not always guaranteed.

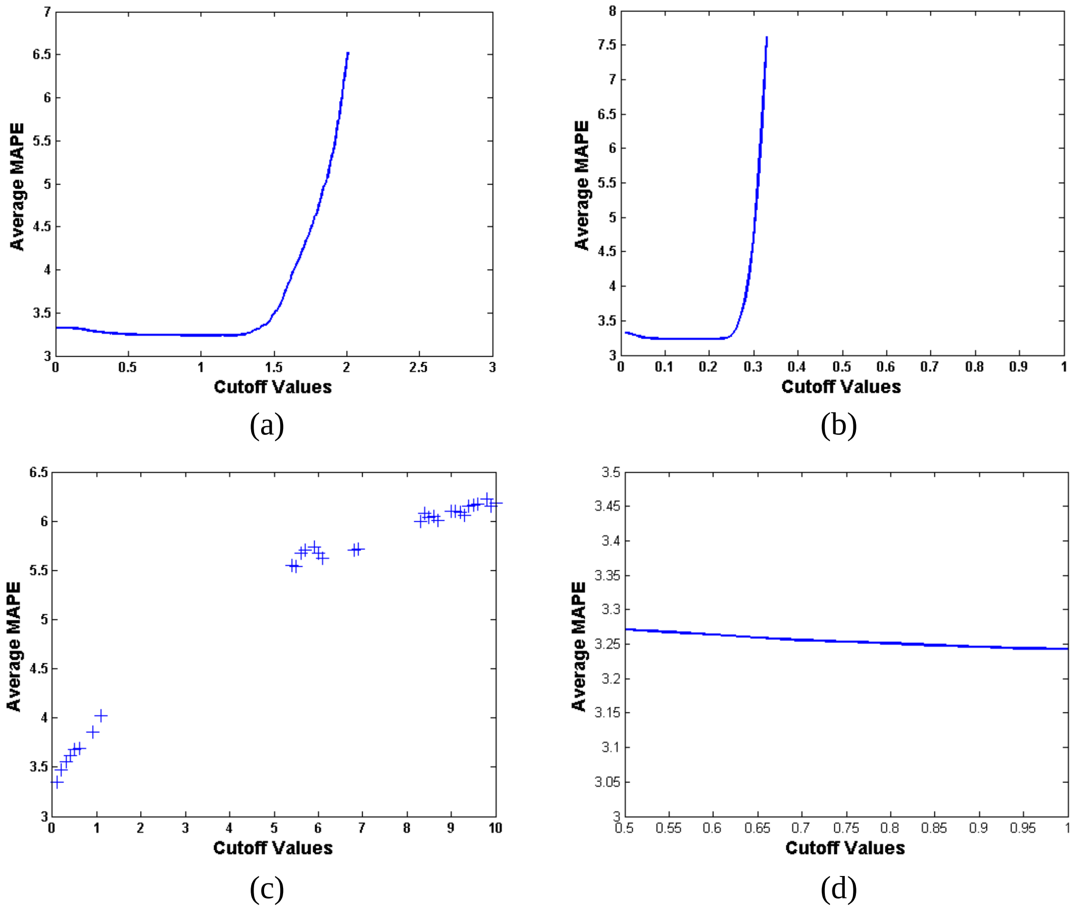


Figure 5.2: Sensitivity in turning parameters in terms of prediction error for (a) PLS-VIP, (b) PLS-BETA, (c) RIVAL, and (d) UVE-PLS in simulated data

The average G and I_C indicators based on 100 MC simulations for the seven variable selection methods are listed in Table 5.3. The good methods are the ones that have both average G and I_C close to 1. The best and the second best performers are highlighted in bold face with and without underline, respectively. It can be seen that for all listed cases, PLS-VIP is either the best performer or the second best performer. To further examine the consistency of different variable selection methods, the frequency of each variable being selected among the 100 MC runs for Case 3 are shown in Figure 5.3, along with the corresponding G and I_C values. Here, Case 3 is selected because its properties are closer to the industrial dataset, e.g., high correlation, low SNR and low PR. By comparing the selection frequency plots for different methods, only PLS-VIP selects all the true relevant variables correctly most of the time (i.e., close to 100% sensitivity) with very few errors on irrelevant variables (i.e., high specificity). In contrast, all the other methods either select true relevant variables less frequency (e.g., PLS-BETA, RIVAL, SR, CARS-PLS, and GA-PLS). Therefore, PLS-VIP is the most consistent variable selection method for the case, which agrees with its high G and I_C values.

Table 5.3: Average G and I_C indicators for the simulation case

		Cases					
Methods	Metrics	1	2	3	4	5	6
PLS-VIP	G	**1.00**	**0.95**	**0.99**	**0.80**	**0.89**	**0.99**
	I_C	**0.91**	**0.72**	**0.83**	**0.64**	**0.81**	**0.84**
PLS-BETA	G	**1.00**	**0.90**	**0.88**	0.61	0.75	0.76
	I_C	**1.00**	**0.91**	**0.69**	0.36	0.81	0.56
RIVAL	G	0.92	0.87	0.79	0.73	0.82	0.87
	I_C	0.46	0.47	0.47	0.33	0.16	0.30
UVE-PLS	G	**0.96**	0.67	0.81	**0.86**	**0.86**	**0.99**
	I_C	0.81	0.61	0.58	**0.57**	**0.83**	**0.82**
SR	G	**1.00**	0.82	0.66	0.60	0.78	0.68
	I_C	0.90	0.60	0.54	0.44	0.80	0.71
CARS-PLS	G	0.65	0.80	0.71	0.45	0.78	0.75
	I_C	0.25	0.53	0.31	0.10	0.43	0.41
GA-PLS	G	0.88	0.77	0.71	0.61	0.77	0.76
	I_C	0.34	0.26	0.18	0.20	0.20	0.19

Next, the variable selection methods are compared by the performance of the PLS soft sensor built using the variables selected. For different variable selection methods, the percentage improvement compared to the full PLS model in terms of MAPE is calculated as follows and provided in Table 5.4.

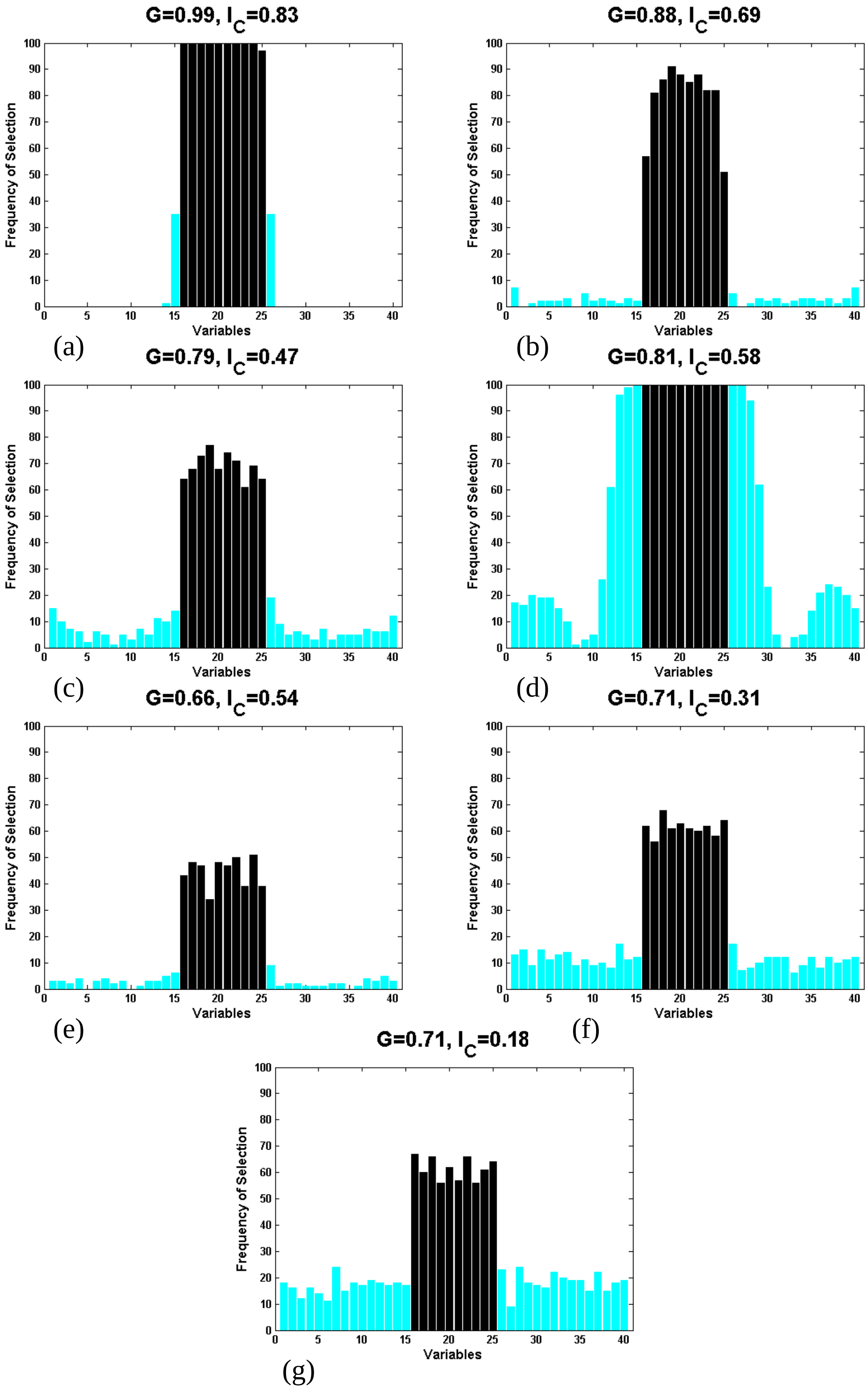


Figure 5.3: Variable selection frequency for Case 3: (a) PLS-VIP, (b) PLS-BETA, (c) RIVAL, (d) UVE-PLS, (e) SR, (f) CARS-PLS, and (g) GA-PLS. The dark-colored bars represent true relevant variables, while the light-colored bars representing true irrelevant variables.

$$\%Imp = \frac{MAPE_{Full} - MAPE_i}{MAPE_{Full}} \times 100\% \qquad (5.9)$$

The positive values indicate improvement in the reduced models, while the negative values indicate deterioration. In summary, different levels of improvement are obtained by variable selection. However, the improvements are not significant in the cases with higher correlation between predictors, i.e., Cases 2, 3, & 4. The best and the second bester performers are also highlighted in bold face with and without underline, respectively. Again, PLS-VIP is either the best or the second best performer in terms of the prediction performance, which is in accordance with the G and I_C values.

Table 5.4: Percentage improvement of average MAPE values from different methods compared to the full model for testing data in the simulated case

		Cases					
Methods	Index	1	2	3	4	5	6
PLS-Full	$MAPE_{Te}$	3.3	3.4	6.0	8.0	6.7	8.6
PLS-VIP	% Imp_{Te}	**2.6**	**1.8**	**1.0**	**0.5**	**8.5**	**7.0**
PLS-BETA	% Imp_{Te}	**2.7**	**1.8**	**0.5**	0.1	6.9	4.5
RIVAL	% Imp_{Te}	-0.6	1.7	0.2	**0.8**	6.0	5.5
UVE-PLS	% Imp_{Te}	2.4	0.9	0.4	0.4	**8.2**	**6.6**
SR	% Imp_{Te}	2.5	0.9	-0.2	0.2	7.2	5.3
CARS-PLS	% Imp_{Te}	1.6	1.2	-0.1	0.0	5.5	3.6
GA-PLS	% Imp_{Te}	-2.1	0.6	-0.3	0.2	4.5	3.7

$MAPE_{Te}$ stands for the average MAPE over 100 MC runs of the full PLS model. % Imp_{Te} stands for percentage improvement for the method compared to the full PLS model.

The prediction performance of the full model (PLS) and PLS-VIP for a specific MC run in Case 5 are shown in Figure 5.4. Only a portion of the testing performance is plotted in order to show the detail. The prediction performances of the soft sensors based on other variable selection methods are somewhat similar to PLS-VIP or between PLS-

VIP and the full PLS model. To reduce clutter, they are not shown in the figure. But overall, the reduced models perform slightly better than the full model.

Besides the MAPE values, the R^2 values are also computed for this simulation case study. The results (shown in Appendix A) indicate that R^2 values are similar among variable selection methods for all the cases. This is due to the way simulation case is generated. In the simulation case, the error term in Equation (5.1) has zero mean and standard deviation specified by SNR, as seen in Equation (5.4). In order to analyze the effect of measurement noise on variable selection performance, three levels of SNR are designed. These levels are generated so that R^2 values of the line of best fit become 0.9, 0.65, and 0.4. The reciprocal of SNR in Equation (5.4), k, can be written as function of R^2, $k = \sqrt{\frac{1-R^2}{R^2}}$. Therefore, different variable selection methods result in similar R^2 values for different cases with same level of SNR.

5.2.3 Remarks

- Consistency: PLS-VIP is the most consistent variable selection method, followed by PLS-BETA and UVE-PLS. High I_C indicates that variables being selected are insensitive to training data being used. High G indicates that the selected variables are mostly true relevant variables.
- Tuning: The performances of PLS-BETA and RIVAL, evaluated by MAPE of the soft sensor, are the most sensitive to their tuning parameters, which is in general not desirable. The tuning for GA-PLS is the most difficult due to the large number of tuning parameters. In addition, no significant improvement is obtained with parameter optimization for GA-PLS.

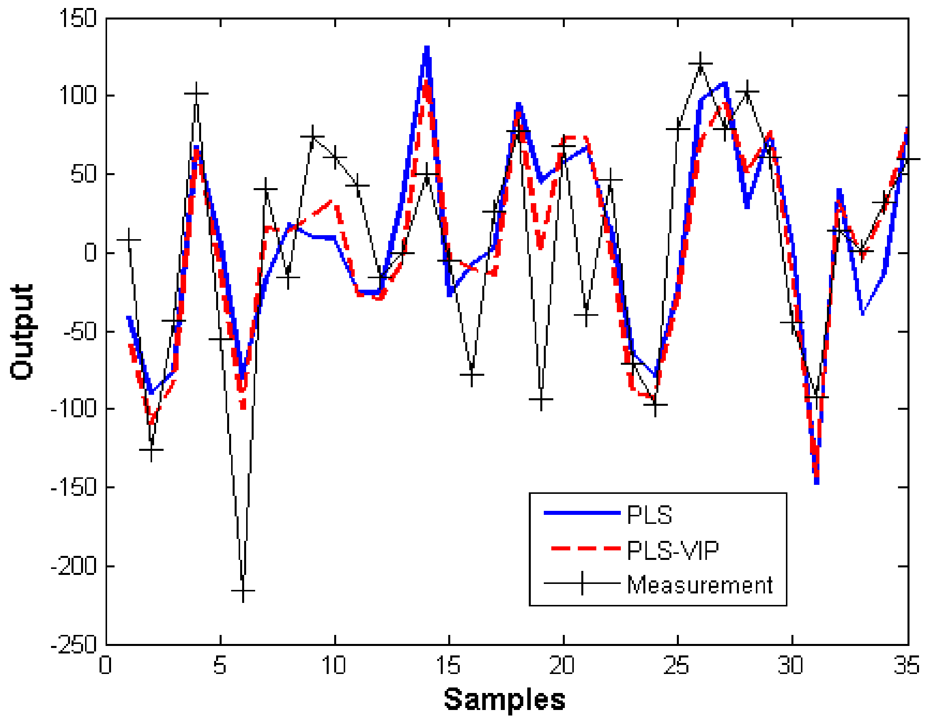


Figure 5.4: Comparison of predicted output for Case 5 in testing set

5.3 Industrial polyester case study

The industrial data was reported in [59], [60], which 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. 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. Thirty-four variables are routinely measured and recorded every 30 seconds. The number of samples in each batch is in the range of 4500 and 7500, varying from batch to batch. Variables are process measurements (e.g., temperature, pressure and valve opening, etc.) and controller settings, which are adjusted manually by the operators. A list of these thirty-four variables is provided in Table 5.5. Product quality is assessed by acidity number and viscosity, which are measured manually and infrequently by the operators. There are only 15 to 25 product quality measurements with uneven intervals available per batch. Thirty-three batches are made available in a 16-month period of time. More process details can be found in [59], [60].

5.3.1 Data preprocessing

For a batch process, the data are stored in a three-dimension array, with dimensions $K \times J \times I_k$, as shown in Figure 5.5. 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. Due to such batch characteristics of the data, preprocessing steps are taken to unfold the three-way array by preserving the variable direction [61], as shown in Figure 5.5. A previous approach proposed by Nomikos and MacGregor [62], [63] is to

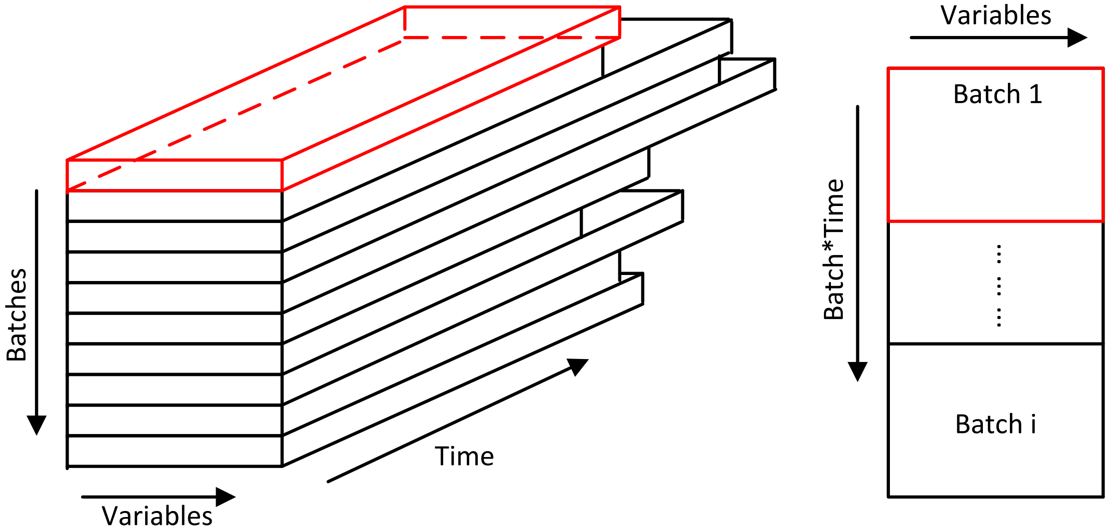


Figure 5.5: Illustration of unfolding three-dimension array to preserve the direction of variables

Table 5.5: List of process variables included in polyester resin dataset

Online Monitored Variable	Variable No.
Mixing rate (%)	1
Mixing rate	2
Mixing rate SP	3
Vacuum line temperature (°C)	4
Inlet dowtherm temperature (°C)	5
Outlet dowtherm temperature (°C)	6
Reactor temperature (sensor 1) (°C)	7
(dummy)	8
Column head temperature (°C)	9
Valve V25 temperature (°C)	10
Scrubber top temperature (°C)	11
Inlet water temperature (°C)	12
Column bottom temperature (°C)	13
Scrubber bottom temperature (°C)	14
Reactor temperature (sensor 2) (°C)	15
Condenser inlet temperature (°C)	16
Valve V14 temperature (°C)	17
Valve V15 temperature (°C)	18
Reactor differential pressure	19
(dummy)	20
Column top temperature PV (°C)	21
Column top temperature SP (°C)	22
V42 way-1 valve opening (%)	23
Inlet dowtherm temperature PV (°C)	24
Inlet dowtherm temperature SP (°C)	25
V42 way-2 valve opening (%)	26
Reactor temperature PV(°C)	27
Reactor temperature SP (°C)	28
(dummy)	29
Valve V25 temperature PV (°C)	30
Valve V25 temperature SP (°C)	31
Valve V42 valve opening (%)	32
Reactor vacuum PV (mbar)	33
Reactor vacuum SP (mbar)	34

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 the objective of this work is to perform variable selection, the approach that preserves the direction of variables is adopted.

The score plot based on principal component analysis (PCA) of the unfolded process variables is shown in Figure 5.6. Only one cluster is formed, with few data points outside of the vicinity, which indicates that the data is collected from a single operating model. The batch dynamics in the product quality variables can be observed in Figure 5.7, where each cycle represents a batch. To synchronize the samples of the process variables and quality variables, two approaches are implemented:

- Method 1: Process measurements collected between two product quality measurements are averaged and utilized as the regressor inputs to be paired up with the product quality measurement. Data preprocessed this way are termed "averaged data".
- Method 2: In contrast to the first method, the integral is taken instead of average. Data preprocessed this way are termed "integrated data".

When new set of measurements become available, same approach is taken to average or integrate the process measurements between two quality variable measurements.

Before modeling, PCA is performed on both averaged data and integrated data to detect outliers. The score plots of the first two principal components are shown Figure 5.8. It was found that the 33 potential outliers identified in the averaged data are all the first samples in each batch. All of them are also classified as outliers in the integrated data, along with 13 other samples. Removing the identified outliers can improve the PLS soft sensor models, which can be observed by comparing Figure 5.10 (a) with (c), and (b)

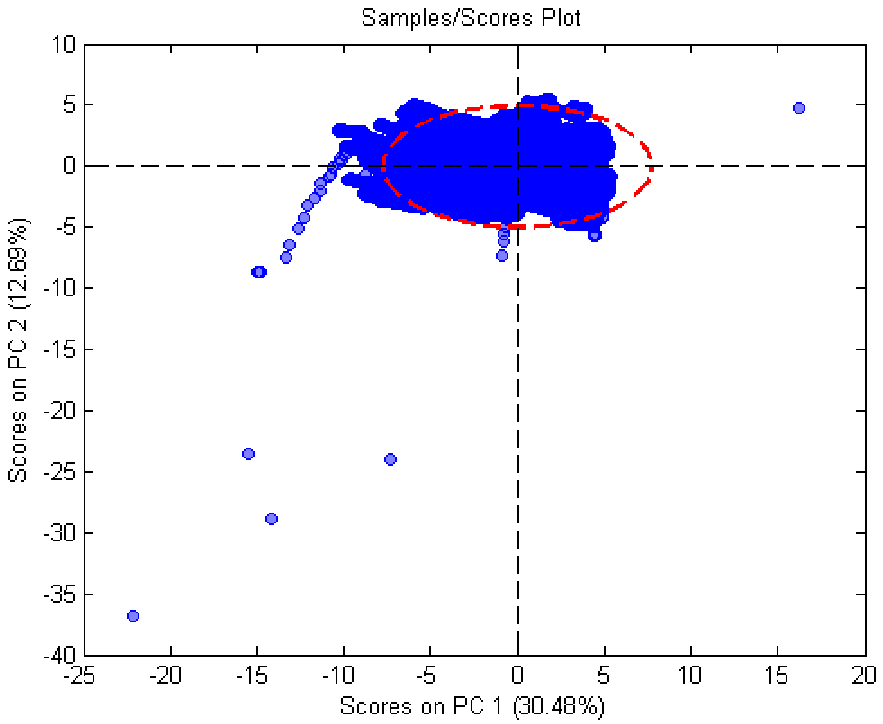


Figure 5.6: Score plot of unfolded original process variables

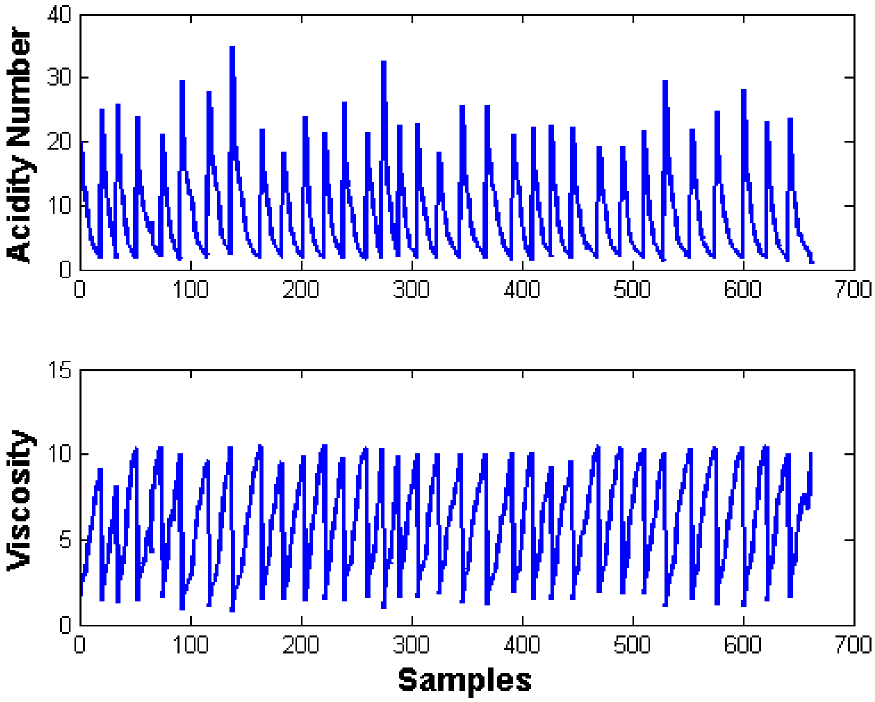
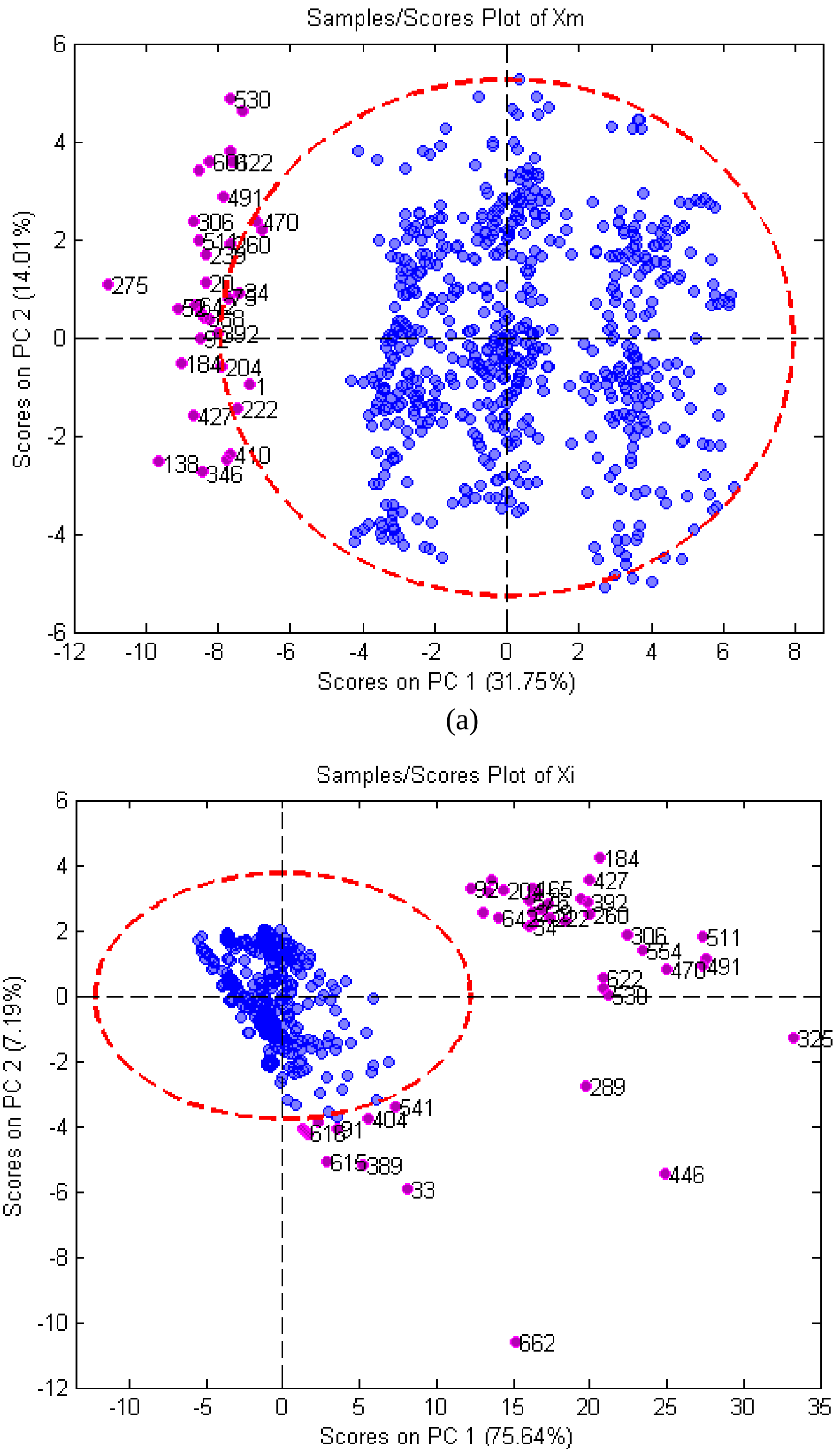


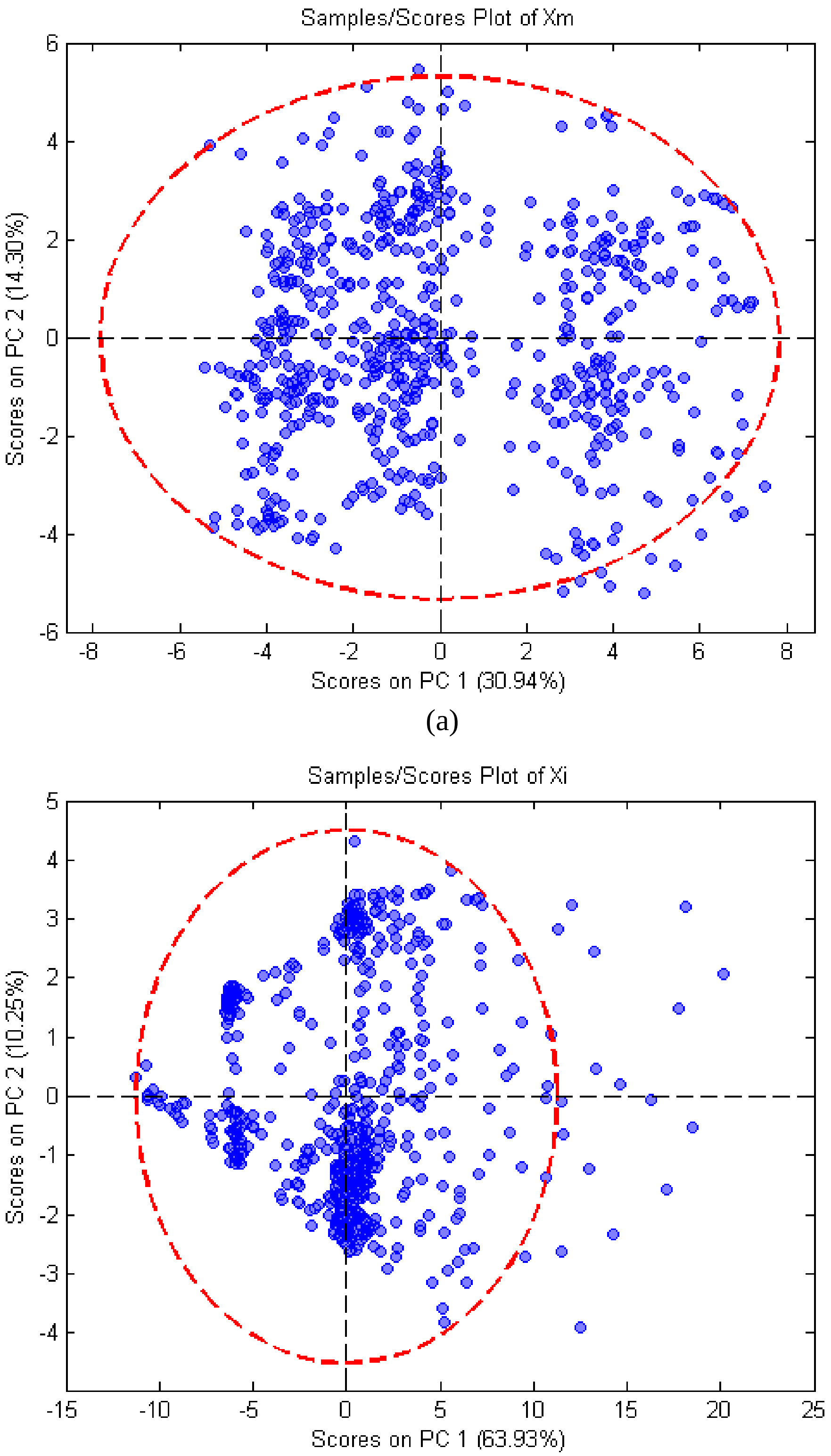
Figure 5.7: Unfolded product quality variables. Each cycle corresponds to samples measured offline during each batch.



(a)

(b)

Figure 5.8: Score plots of (a) averaged data and (b) integrated data

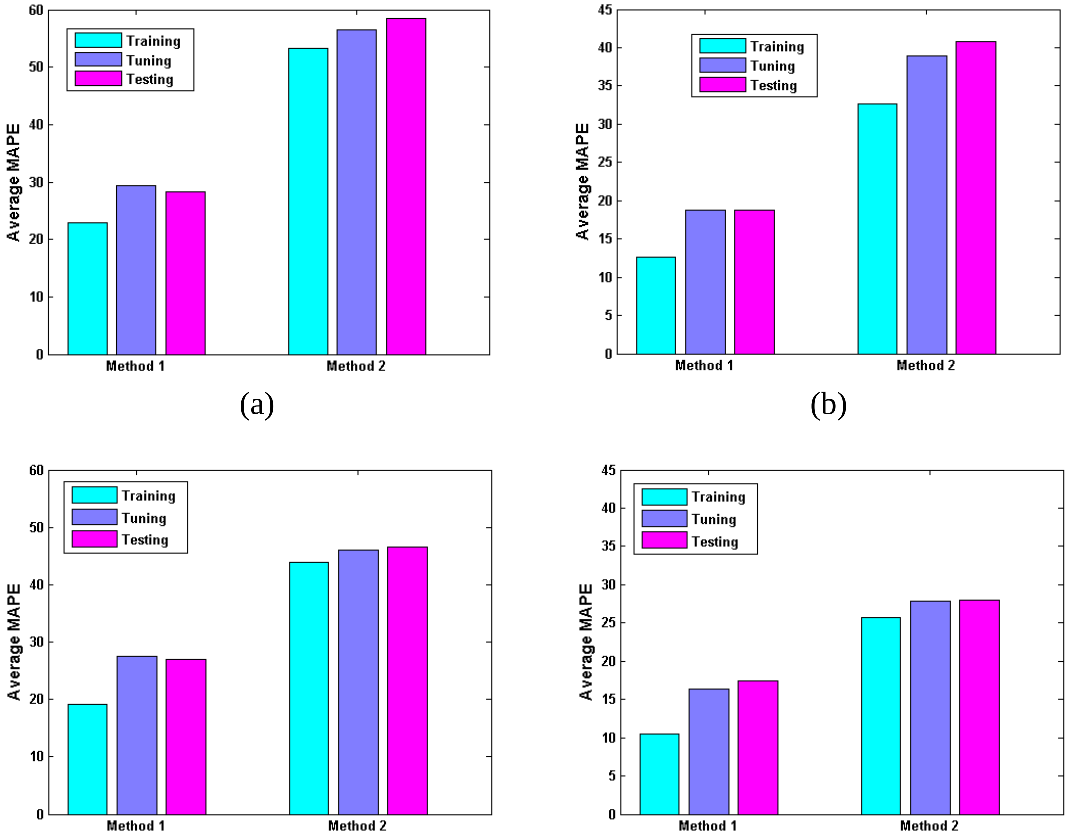


(a)

(b)

Figure 5.9: Score plots of (a) averaged data and (b) integrated data after outlier removal

with (d). The score plots of both the averaged data and integrated data after outlier removal are shown in Figure 5.9.



(a) (b)

(c) (d)

Figure 5.10: Prediction performance of original PLS models on (a) acidity number, (b) viscosity models before outlier removal and (c) acidity number, (d) viscosity models after outlier removal

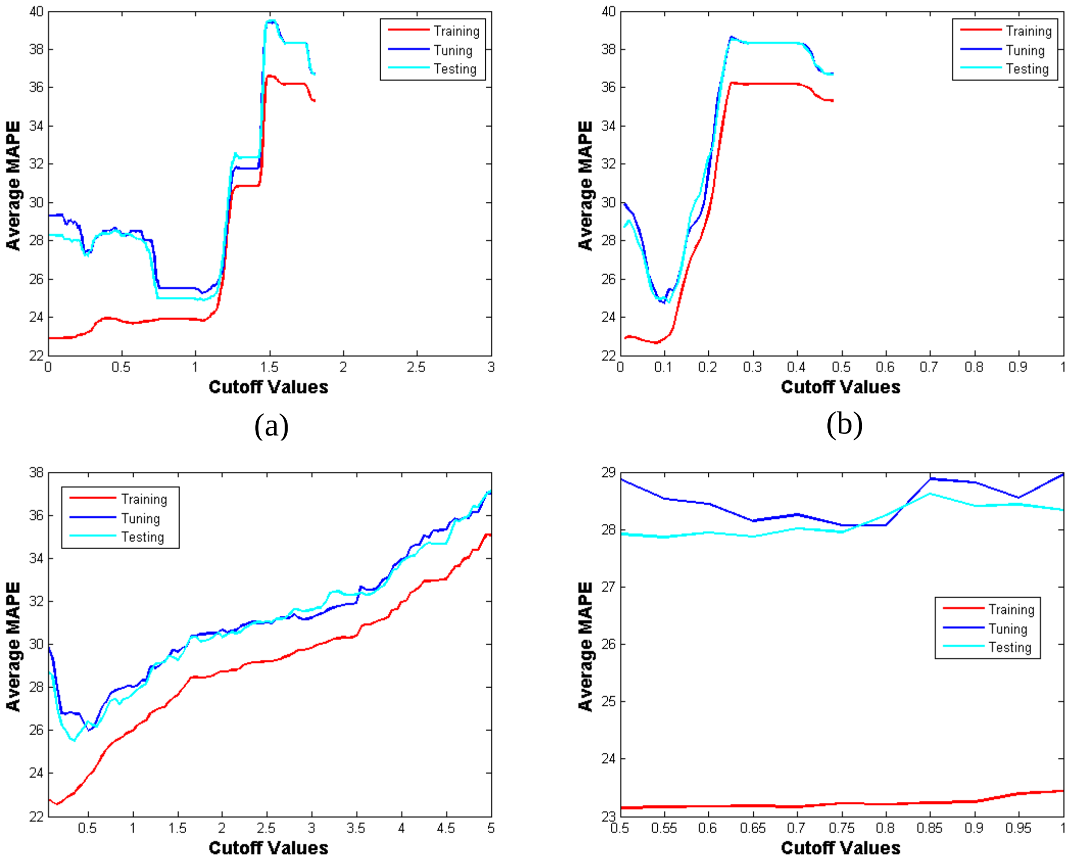
Sample outliers detected by PCA are removed before variable selection and modeling. Twenty-six batches i.e., approximately 80% of total batches, are used for training, three batches are used for parameter tuning, and the remaining four batches are used for testing. 100 MC simulations are performed so that different batches are used for training, tuning and testing in each run, then the average performance is obtained for evaluation.

5.3.2 Comparison results

RIVAL is not evaluated using the industrial dataset since the data do not meet the criterion of positive regression coefficients. Lasso is added in its place. For each variable selection method, separate models are built to predict acidity number and viscosity, respectively.

First, we compare the performance of the full PLS models for both acidity number and viscosity using the two synchronization methods. The results are shown in Figure 5.10, which clearly show that the model developed using the averaged data is significantly better than the integrated data in terms of prediction performance. Even though removing outliers results in more significant improvement for integrated data, the prediction performance is still worse than averaged data. This is true for both acidity number and viscosity models. Comparing Figure 5.8 (a) with (b), it can be observed that the averaged data follow a Gaussian distribution to a much better degree than the integrated data, and the outliers identified by the averaged data are more meaningful than that identified in the integrated data. This is understandable because calculating the average improves the gaussianity of the data according to the central limit theorem [64], [65], while integration does not. Therefore, throughout the rest of the work, we only report the results obtained using the averaged data.

For all the reduced models, the results presented are all optimized or tuned using the independent tuning batches. Unlike the simulation case, all methods are sensitive to their tuning parameters with much larger MAPE in the industrial polyester case. The selected sensitivity results for acidity number model are shown in Figure 5.11. The sensitivity results for viscosity models are not shown here since they are similar to the ones from acidity number models. Compared Figure 5.11 with Figure 5.2, all the methods showed higher sensitivity to the tuning parameter. However, it is worth noting that the "greater than one rule" still applies for PLS-VIP in this industrial polyester case, since it falls in the minimum region. For both PLS-BETA and Lasso, one must be able to pinpoint the narrow optimal region in order to obtain satisfactory results.



(a)

(b)

(c)

(d)

Figure 5.11: Sensitivity in tuning parameters in terms of prediction error for (a) PLS-VIP, (b) PLS-BETA, (c) Lasso, and (d) UVE-PLS in acidity number model

In terms of variable selection consistency when different batches are used as the training batches, the I_C indices of different methods are given in Table 5.6. The best and second best performers are highlighted in bold face with and without underline, respectively. It is worth noting that the G values cannot be assessed because the true relevant variables are unknown. PLS-VIP produces the most consistent variable selection results with significantly higher I_C values than others. As an example, the variable selection frequency plots of PLS-VIP and GA-PLS for both acidity number and viscosity are compared in Figure 5.12. Furthermore, most of the variables selected by PLS-VIP are temperatures. Since both acidity number and viscosity are functions of temperature, the variables selected by PLS-VIP do reflect the process knowledge.

Table 5.6: Comparison of I_C values of different methods in the polyester case

Methods	Acidity Number	Viscosity
PLS-VIP	**0.91**	**0.91**
PLS-BETA	0.65	0.60
Lasso	0.56	0.53
UVE-PLS	**0.72**	**0.62**
SR	0.46	0.29
CARS-PLS	0.48	0.41
GA-PLS	0.28	0.26

However, it was surprising to find out that some of the retained variables are highly correlated with each other, as one would expect variable selection to remove the correlations. Some of the correlated variables are shown in Figure 5.13. Two approaches are taken to remove those variables:

1. To remove highly correlated variables before variable selection.
2. To remove highly correlated variables after variable selection.

This aspect has never been examined before, as one would simply perform variable selection and use the selected variables to build soft sensors; and let PLS to handle the collinearity. Despite the fact that PLS is able to deal with the collinearity among different variables, if we can get a smaller model without sacrificing the performance, principle of parsimony applies. With smaller model, first of all, the model maintenance is cheaper and

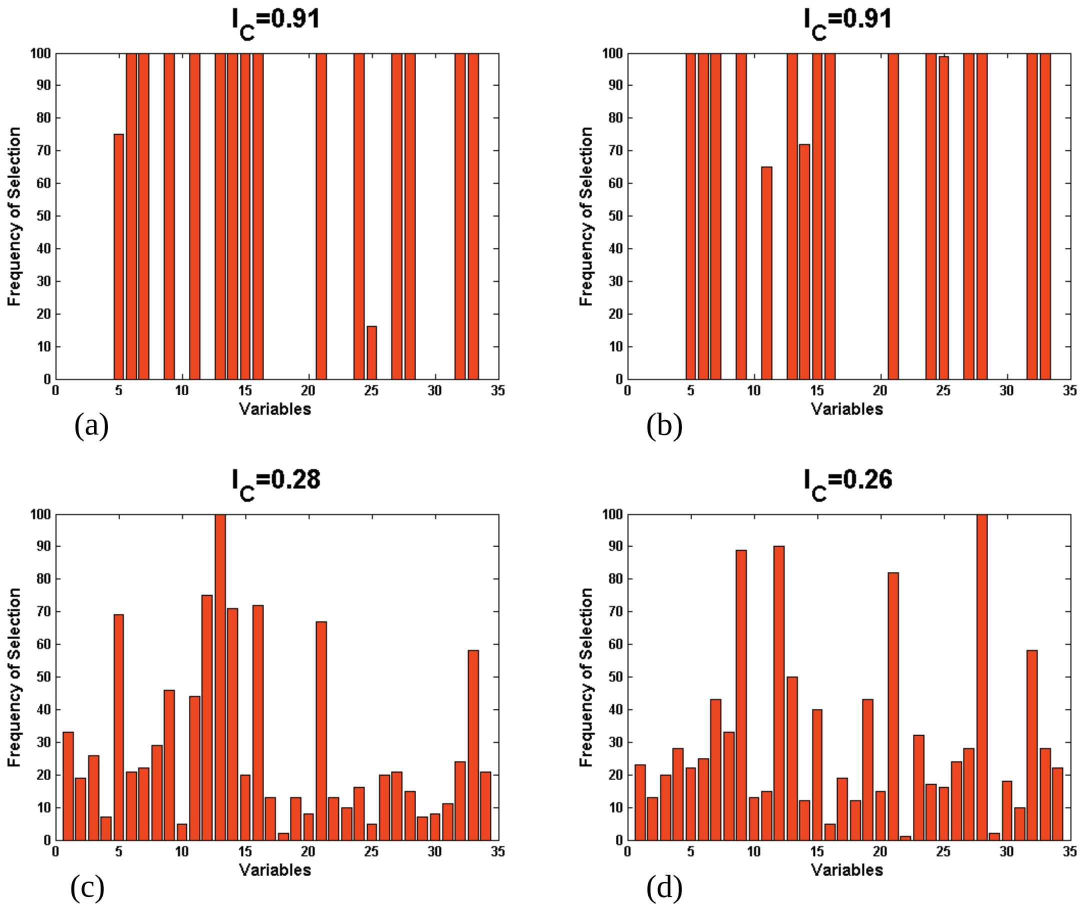


Figure 5.12: Variable selection frequency by PLS-VIP (a) acidity number, (b) viscosity; and by GA-PLS (c) acidity number, (d) viscosity

easier. And more importantly, if the sensor on the variables used to build the model failed, the correlated variables can be substituted to keep the soft sensor working during that kind of circumstances.

Three levels of correlation removal are implemented, at 0.99, 0.95, and 0.90. Results showed that the second removal approach, which removes highly correlated variables after variable selection, performed better than the first approach. As one would expect the first approach to work better than the second one since a smaller model would be obtained before variable selection, which can result in an easier implementation of variable selection. However, if the highly correlated variables, which contain relevant information of primary variables, are removed ahead of time, the covariance explained by those highly correlated variables would be different. This significantly affects the results of variable selection. Hence, the second approach is adopted. To compare the correlation removal at three levels, the prediction error of the testing sets are examined. For removal at 0.99, the prediction errors are similar to the ones without correlation removal. For removal at 0.95 and 0.90, the prediction errors increased, which means the models are oversimplified. Therefore, correlation removal at 0.99 is adopted. Even though the prediction performance does not improve significantly, in terms of practicability, a smaller model is always better without sacrificing the performance. The selected variables after correlation removal are shown in Figure 5.14.

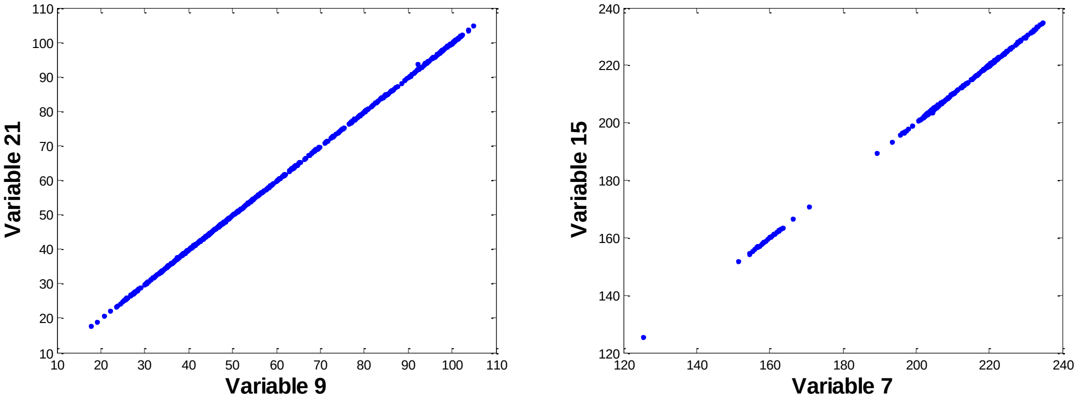


Figure 5.13: Examples of highly correlated variables

The percentage improvement in MAPE of soft sensors built based on the seven variable selection methods for both acidity number and viscosity models are shown in Table 5.7. The best and second best performers are highlighted in bold face with and without underline, respectively. All reduced models yield better performance than the full model. The most significant improvements are obtained by PLS-VIP (28.4%), followed by PLS-BETA (23.9%) in predicting acidity number; and Lasso (33.3%), followed by

PLS-VIP (30.5%) in predicting viscosity. The time series plots of both measured and predicted acidity number and viscosity for one of the batch are shown in Figure 5.15. It can be seen that the predictions made with variable selection follow the true measurement more closely than the original PLS.

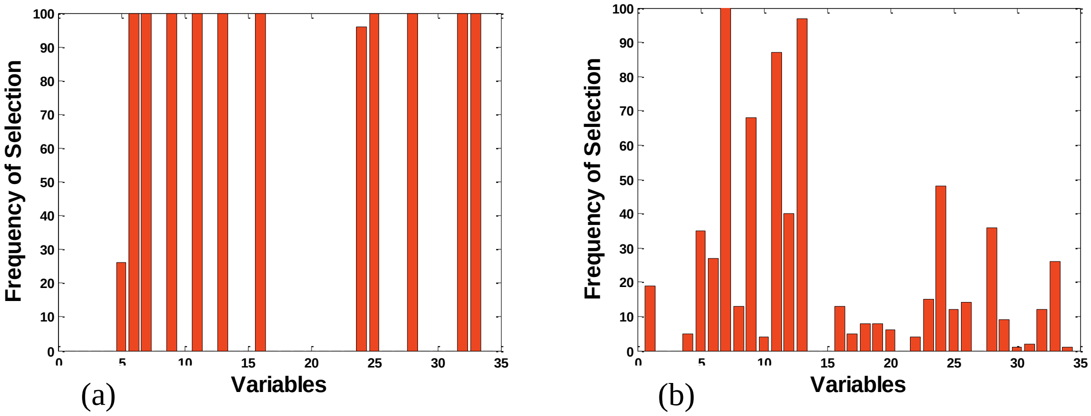


Figure 5.14: Variable selection frequency by (a) PLS-VIP and (b) GA-PLS for acidity number model after correlation removal

For the industrial polyester case study, the R^2 and bias values are also computed. They are listed in Table 5.8 and Table 5.9, respectively. Again, the best and second best performers are highlighted in bold face with and without underline, respectively. There are significant improvements in R^2 values with variable selection, in contrast to no R^2 improvement in the simulated case study. Overall, the comparison results based on R^2 and bias values agree with the ones based on MAPE. Both demonstrate the benefits of applying variable selection before modeling and PLS-VIP performs the best among all methods compared.

5.4 Comparison of the simulation and industrial polyester case studies

It is observed that the improvements obtained in the industrial polyester case study are more significant than the ones in the simulation case study. The main reason for such discrepancy is most likely the difference in sample distribution. For the simulated case study, the process data follows multivariate normal distribution and PLS is expected to handle the white noises in the data well, including the irrelevant variables. As a results, there is not much room for improvement over the full PLS model. On the other hand, for the industrial polyester case study, process data shown clear non-Gaussian distribution and contains outliers. For such cases, variable selection is demonstrated to eliminate variables that are irrelevant, which may contain high noise and outliers, and therefore significantly improve the soft sensor performance.

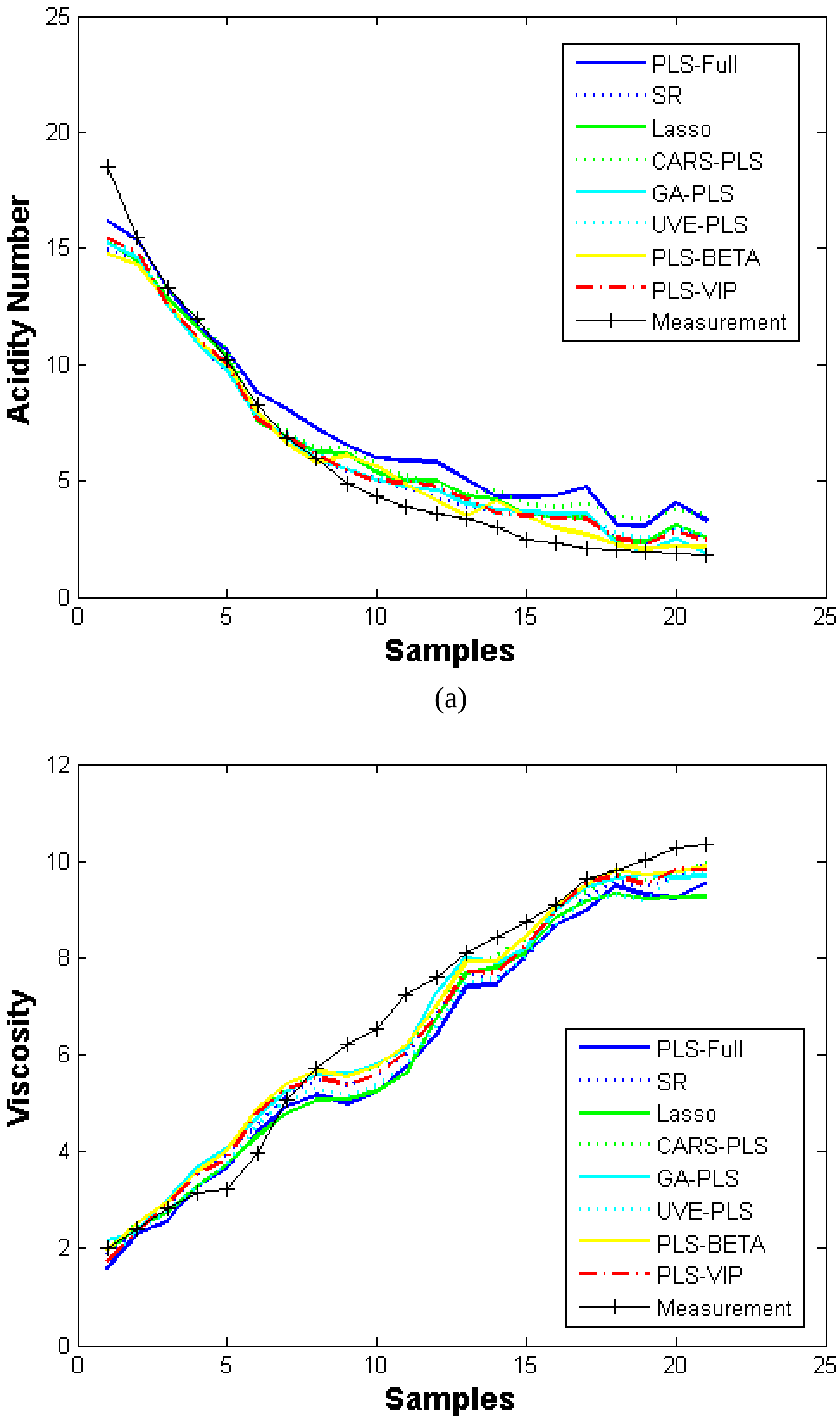
Table 5.7: Comparison of percentage improvement in MAPE of different methods in the polyester case

		Quality Variables	
Methods	Index	Acidity Number	Viscosity
PLS-Full	$MAPE_{Te}$	27.0	17.4
PLS-VIP	% Imp_{Te}	**28.4**	**30.5**
PLS-BETA	% Imp_{Te}	**23.9**	20.9
Lasso	% Imp_{Te}	10.8	**33.3**
UVE-PLS	% Imp_{Te}	10.5	12.9
SR	% Imp_{Te}	19.8	17.3
CARS-PLS	% Imp_{Te}	21.1	22.3
GA-PLS	% Imp_{Te}	21.9	19.9

$MAPE_{Te}$ stands for the average MAPE over 100 MC runs of the full PLS model. % Imp_{Te} stands for percentage improvement for the method compared to the full PLS model.

5.5 Conclusions

In this chapter, seven variable selection methods for PLS-based soft sensor development are compared using a simulated case study and an industrial polyester case study. To address the challenge that there is no published method that directly evaluates the variable selection performance when the true variable relevance is unknown, we propose an



(a)

(b)

Figure 5.15: Comparison of prediction on averaged data testing set for (a) acidity number and (b) viscosity from one batch

information entropy based performance index (I_C) to directly assess the consistency of a variable selection method. It is shown that I_C is consistent with G, but more sensitive, for simulated case studies where the true variable relevance is known. It is also shown that I_C is a good performance indicator for the industrial polyester case study where the true variable relevance is unknown.

Table 5.8: Comparison of R^2 values of different methods in the polyster case

Methods	Acidity Number	Viscosity
PLS-Full	0.06	0.40
PLS-VIP	**0.95**	**0.92**
PLS-BETA	**0.93**	**0.88**
Lasso	0.92	**0.92**
UVE-PLS	0.89	0.83
SR	0.43	0.87
CARS-PLS	**0.93**	**0.88**
GA-PLS	0.45	0.87

Table 5.9: Comparison of bias in predicted variables of different methods in the polyester case

Methods	Acidity Number	Viscosity
PLS-Full	0.13	-0.19
PLS-VIP	**-0.03**	**-0.01**
PLS-BETA	0.05	**-0.10**
Lasso	0.05	**-0.01**
UVE-PLS	0.19	-0.11
SR	0.07	**-0.10**
CARS-PLS	**0.01**	-0.11
GA-PLS	0.07	-0.13

From the simulation case study, we are able to observe how each variable selection method is affected by different characteristics of the data. Overall, PLS-VIP is the most consistent variable selection method. For most of the cases tested, PLS-VIP has the highest or the second highest G and I_C values, which is consistent with the MAPE reduction on the reduced model over the full model where PLS-VIP performs the best or the second best most of the time.

For the industrial polyester case study, independent models are developed for two product quality variables. Two different synchronization methods are used to synchronize the process variables with product quality variables that are sampled at different frequencies. It is found that synchronization by taking the average of the process measurements between production quality variables performs better than synchronization by integration in terms of MAPE from soft sensors. The results also demonstrate the advantages of applying variable selection along with correlation removal before soft sensor development. Prediction errors from the best performing models after variable selection are reduced by 28% and 33% for acidity number and viscosity, respectively. Furthermore, PLS-VIP is the most consistent variable selection method among all the methods compared based on I_C and MAPE. The consistency between I_C and MAPE demonstrates the effectiveness of I_C for evaluating the variable selection performance, when the true variable relevance information of the data is unknown.

The simulation case study indicates that some of the variable selection methods are insensitive to tuning parameters. However, the industrial polyester case study indicates that one should always make the effort to search for the optimal parameter settings when industrial applications are considered. The performance can be greatly improved when optimal tuning parameters are chosen, especially for Lasso and RIVAL, which are sensitive to tuning parameters.

It is also observed that the improvements obtained in the industrial polyester case study are more significant than the ones in the simulation case study. The main reason for such discrepancy is most likely the sample distribution being non-Gaussian. For the simulated case study, the process data follows multivariate normal distribution, and PLS is expected to handle the white noises well in the data, including the irrelevant variables. As a result, there is not much room for improvement over the full PLS model. On the other hand, for the industrial polyester case study, process data shows clear non-Gaussian distribution and contains outliers. For such cases, variable selection could eliminate variables that are irrelevant, which may contain high noise and outliers, and therefore significantly improve the soft sensor performance.

Chapter 6. Applications to Digester Case Study

6.1 Introduction

The performances of different variable selection methods are also compared using digester case studies: a high-yield digester simulated with extended Purdue model [43] and an industrial Kamyr digester [3].

In pulping process, the wood chips are converted into pulp by displacing lignin from cellulose fibers by reacting with a chemical solution (referred to as white liquor). Kraft pulping is one of the most commonly used chemical pulping processes, in which wood chips reacts with an aqueous solution of sodium hydroxide in a continuous Kamyr digester, which is a complex vertical plug flow reactor to remove lignin at high temperature [66]. Most of the continuous digesters consist of three basic zones: an impregnation, one or more cooking zones, and a wash zone. White liquor penetrates and diffuses into the wood chips as they flow through the impregnation zone. The white liquor and woodchips are then heated to reaction temperatures; and the lignin is removed through one or more cooking zones, where the white liquor is either in co-current or counter-current flow with respect to the wood chips. In the wash zone, a counter-current flow of liquor washes the degradation products from the pulp. This also cools the pulp to quench the reaction and reduces damage to the cellulose fibers from continued reaction. The schematic diagram of a single-vessel Kamyr digester is shown in Figure 6.1. Kappa number is used to measure the residual lignin in the pulp, which becomes a direct indicator of pulp quality. It is desired to minimize the variations in Kappa number in the pulp product. [43]

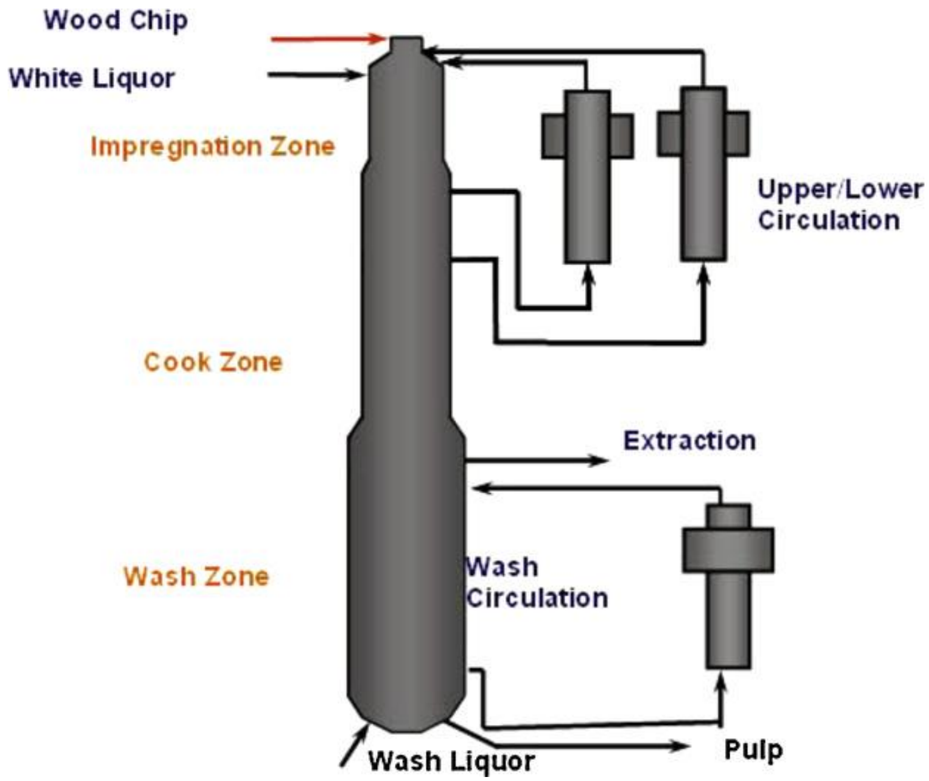


Figure 6.1: Schematic of a high-yield single-vessel Kamyr digester [3]

6.2 Simulated continuous digester

The extended Purdue model developed in [43] is implemented to test the performances of the different variable selection methods. The single-vessel high-yield digester is approximated by 50 continuous stirred tank reactors (CSTRs) in series, which results in 950 nonlinear ordinary differential equations (ODEs). Each CSTR is assumed to contain three phases: solid, entrapped liquor, and free liquor.

6.2.1 Data description

The primary output of this simulator is the Kappa number. The secondary outputs are the effect alkali (EA), hydrosulfide (HS), dissolved lignin (DL), dissolved carbohydrate (DC), and free liquor temperature (T) of upper recirculation, lower recirculation, and extraction flow. [3] The measuring frequencies for both primary and secondary outputs are every 6 minutes during simulation. White noises are added to both measurements during simulation.

Four different types of disturbances are introduced:

1. Integrated white noise with variance of 0.0009 in entering white liquor EA concentration.
2. Integrated white noise with variance of 0.005 in upper and lower heater temperatures.
3. Integrated white noises with variances of 0.0003, 0.0007, 0.007, 0.0001, and 0.002 in the five wood compositions at the inlet.
4. The combination of the above three cases.

For each case, 1500 hours of data are simulated, which is equivalent to 15000 samples. Approximately 800 hours of data are used for training, and the rest are used for testing. There are 23 secondary variables included in the model, as listed in Table 6.1. They comprise the full model. The same procedures for variable selection are implemented for each case.

6.2.2 Comparison results

The selection results for the seven variable selection methods are shown in Figure 6.2 to Figure 6.5, for the four types of disturbances introduced, respectively. The vertical lines separate the groups of variables, e.g., the first group is EA concentrations from 12th, 18th, and 35th CSTR. For disturbance 1, EA concentrations are selected by all the seven methods, since the disturbance is introduced on the EA concentrations at the inlet. For disturbance 2, the fourth group of variables is selected by all the methods, except for PLS-BETA, due to the disturbance introduced in upper and lower heater temperatures.

Table 6.1: Regressors for digester simulator

Variable No.	Description	Location	CSTR index
1	EA_{12}, free liquor effective alkali concentration	Upper cooking	12
2	EA_{18}, free liquor effective alkali concentration	Lower cooking	18
3	EA_{35}, free liquor effective alkali concentration	Extraction	35
4	DL_{12}, free liquor dissolved lignin concentration	Upper cooking	12
5	DL_{18}, free liquor dissolved lignin concentration	Lower cooking	18
6	DL_{35}, free liquor dissolved lignin concentration	Extraction	35
7	T_{12}, free liquor temperature	Upper cooking	12
8	T_{18}, free liquor temperature	Lower cooking	18
9	T_{35}, free liquor temperature	Extraction	35
10	T_U, upper heater exit temperature	Upper heater exit	--
11	T_L, lower heater exit temperature	Lower heater exit	--
12	pEA_{12}, free liquor passive effective alkali concentration	Upper cooking	12
13	pEA_{18}, free liquor passive effective alkali concentration	Lower cooking	18
14	pEA_{35}, free liquor passive effective alkali concentration	Extraction	35
15	HS_{12}, free liquor hydrosulfide concentration	Upper cooking	12
16	HS_{18}, free liquor hydrosulfide concentration	Lower cooking	18
17	HS_{35}, free liquor hydrosulfide concentration	Extraction	35
18	pHS_{12}, free liquor passive hydrosulfide concentration	Upper cooking	12
19	pHS_{18}, free liquor passive hydrosulfide concentration	Lower cooking	18
20	pHS_{35}, free liquor passive hydrosulfide concentration	Extraction	35
21	DC_{12}, free liquor dissolved carbohydrates concentration	Upper cooking	12
22	DC_{18}, free liquor dissolved carbohydrates concentration	Lower cooking	18
23	DC_{35}, free liquor dissolved carbohydrates concentration	Extraction	35

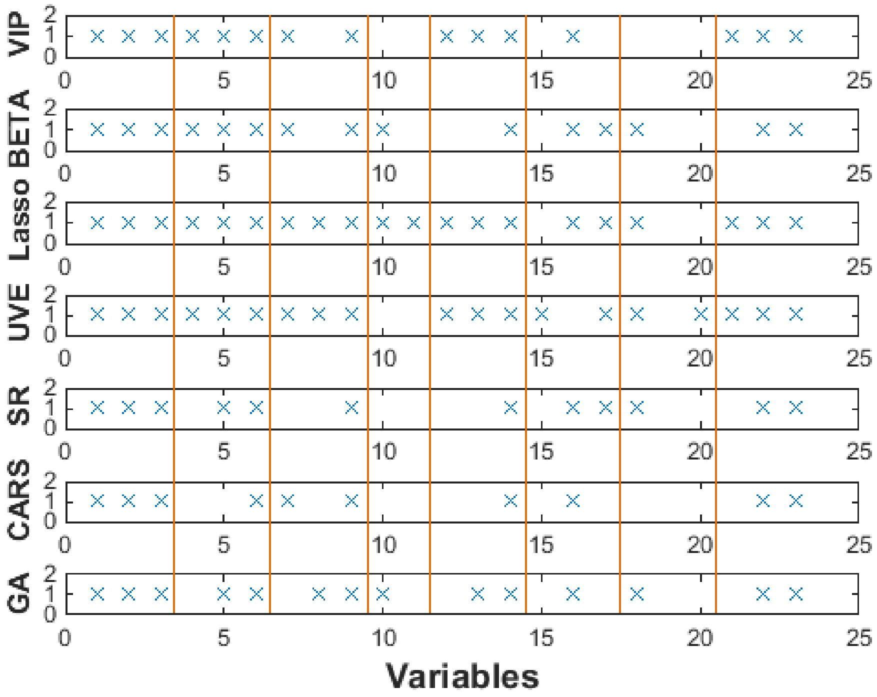


Figure 6.2: Variable selection for digester simulator with disturbance case 1

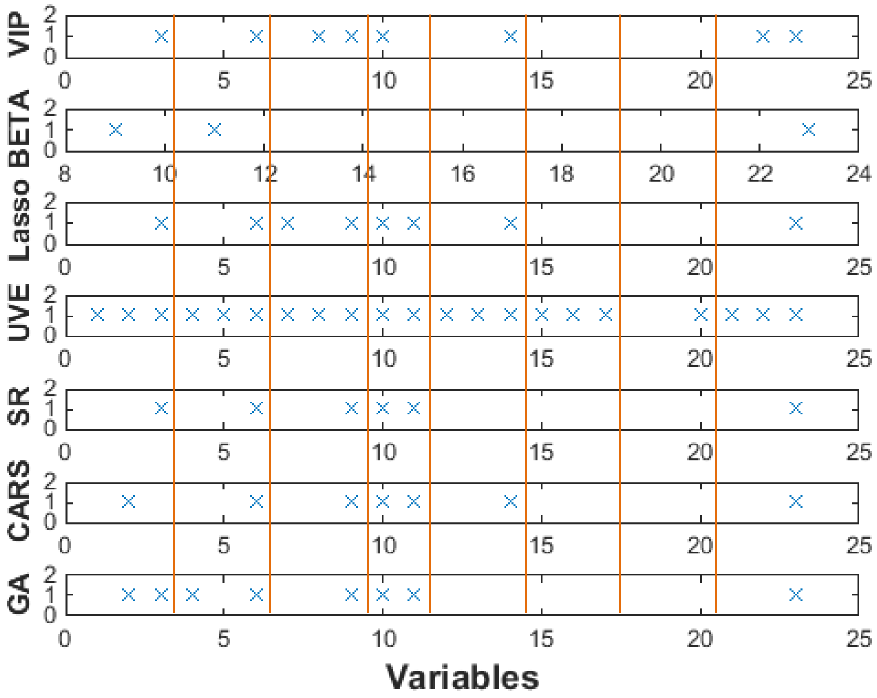


Figure 6.3: Variable selection for digester simulator with disturbance case 2

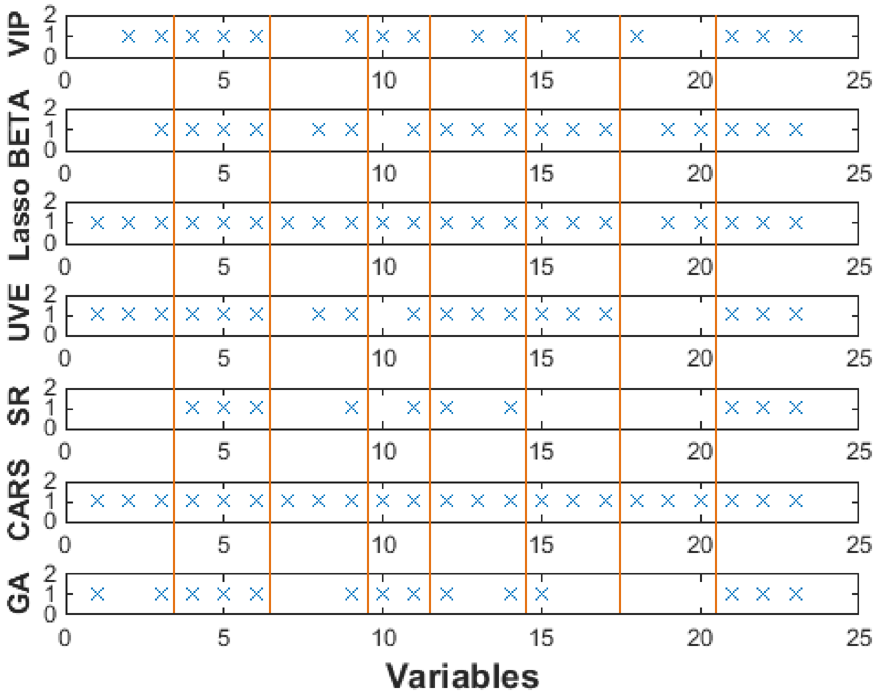


Figure 6.4: Variable selection for digester simulator with disturbance case 3

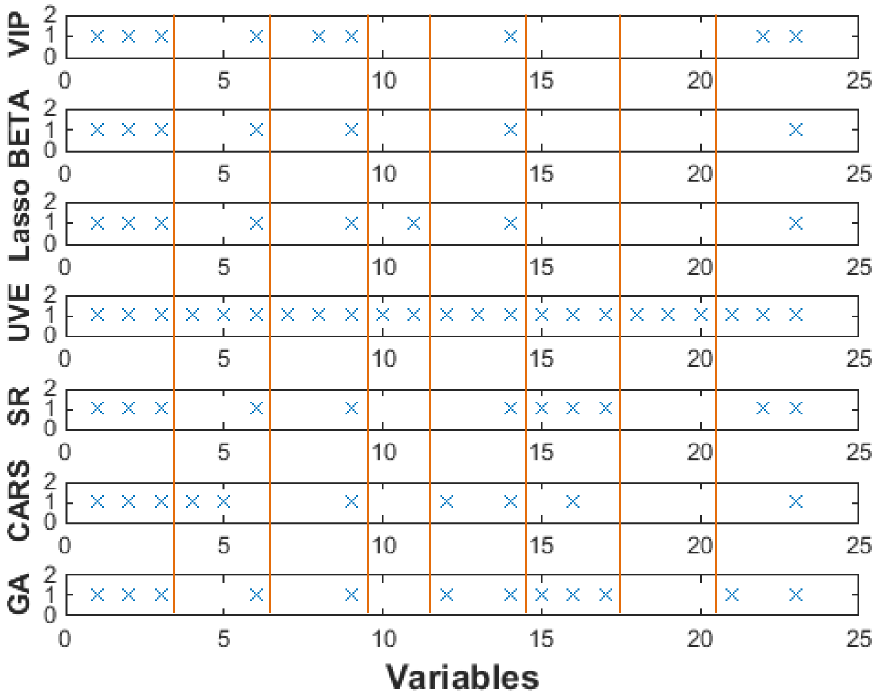


Figure 6.5: Variable selection for digester simulator with disturbance case 4

For disturbance 3, DL concentrations are all selected. Disturbance introduced on wood composition at the inlet directly affects the concentration of DL. Therefore, the selection also reflects the perturbations introduced. For disturbance 4, which is a combination of all the three disturbances, all the EA concentrations are selected as well. One other trend observed is that the variables from 35th CSTR are selected more frequent compared to the ones from the other two CSTR locations. Variables closer to the blow line are more relevant to the Kappa number, since this is where the measurements are taken.

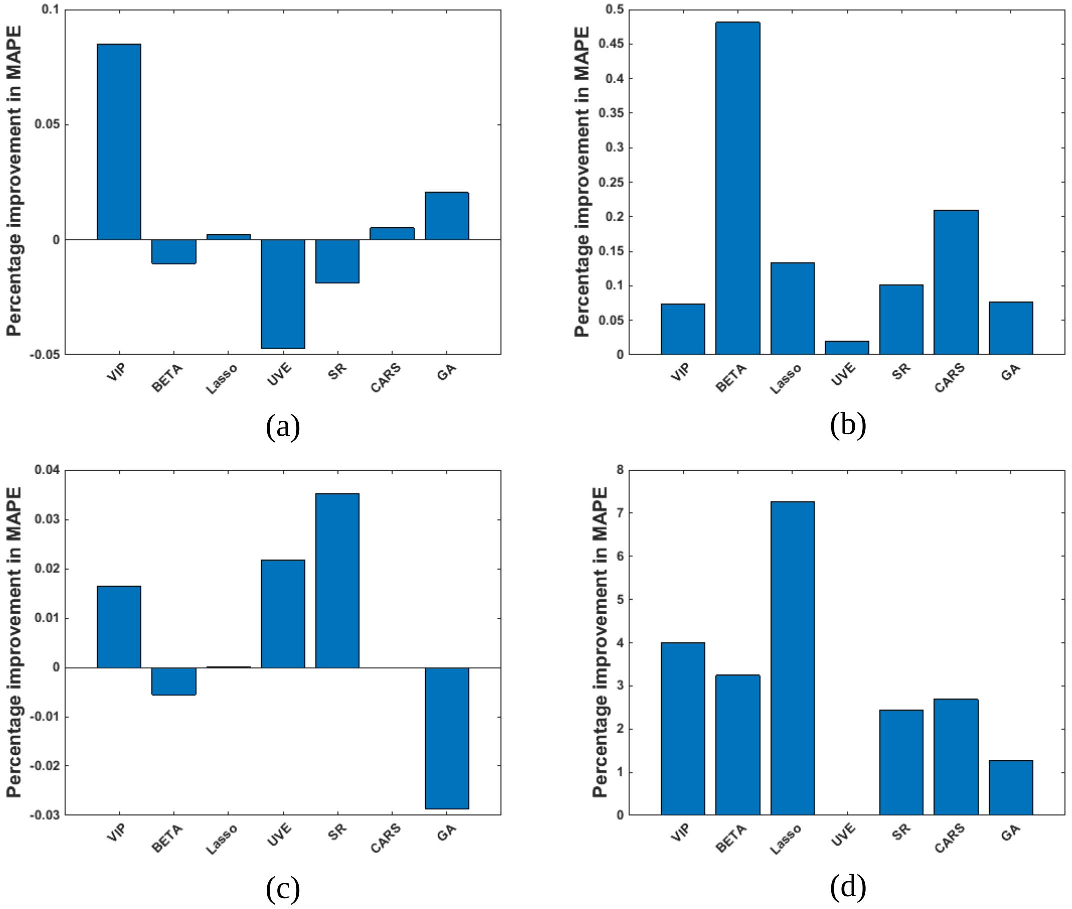


Figure 6.6: Percentage improvements in prediction error for digester simulator with disturbance on (a) EA concentrations, (b) upper and lower heater temperatures, (c) wood chip concentrations, (d) combination of all three

The prediction performances of the reduced models are compared with the full model. The results are shown in Figure 6.6 for the four different types of disturbances introduced. Most of the reduced models yield improvements compared to the full model. The improvements from the first three cases are insignificant. However, the ones from the fourth case, which is the one combines all three disturbances, yield quite significant improvement. The highest improvement is obtained by Lasso with 7.5%, followed by PLS-VIP with 4%. When the process is highly disturbed as in Case 4, variable selection can improve the prediction performance by eliminating the noises and irrelevant information. In addition, even though the prediction improvements from the first three cases are insignificant, the selection results are easily interpretable.

6.3 Industrial Kamyr digester

The industrial Kamyr digester data was obtained from a paper mill located at Mahrt, Alabama, run by MedWestvaco Corporation. The digester has a DCS and a Duralyzer-NIR digester analyzer system to measure secondary variables such as EA, DL, total dissolved solids (TS) and active alikali (AA) for the different zones of the digester [3].

6.3.1 Data preprocessing

84 process variables were made available from two separate datasets. Raw data is treated by visual inspection, extreme point removal and interpolation, and smoothing, etc. Only 49 variables are retained after preprocessing, such as manual outlier removal, with one being Kappa number, resulting in 13000 samples. 16 variables, listed in Table 6.2, are chosen by process knowledge [67]. The sampling frequency for Kappa number is much lower than that for other process variables. For the samples without Kappa measurements, the previous measurement of Kappa number is used. Thus, many of the Kappa

Table 6.2: Selected regressors by process knowledge for industrial Kamyr digester

Variable No.	Description	Zone
1	Top circulation temperature	Impregnation
2	Combine heater temperature	Impregnation
3	Upper cooking [Na_2S]	Upper cook
4	Upper cooking [Na_2CO_3]	Upper cook
5	Upper cooking [Lignin]	Upper cook
6	Upper cooking [Total EA]	Upper cook
7	Lower cooking [Lignin]	Lower cook
8	Lower cooking [Na_2CO_3]	Lower cook
9	Lower cooking [Na_2S]	Lower cook
10	Lower cooking [Total EA]	Lower cook
11	Extraction temperature	Transition
12	Lower extraction [Na_2CO_3]	Transition
13	Lower extraction [Lignin]	Transition
14	Lower extraction [Na_2S]	Transition
15	Lower extraction [Total EA]	Transition
16	Lower extraction [% solids]	Transition

measurements have repeated values.

For continuous process, moving window approach, shown in Figure 6.7, is utilized to generate multiple datasets for MC simulation without corrupting the dynamics of the process. 100 MC runs are generated, with parameters listed in Table 6.3, and used for training. 3000 samples are used for testing.

Different moving window parameters are tested to design MC runs. The preliminary results of variable selection over 100 MC runs are quite inconsistent. The selection frequencies of each method over 100 MC runs, generated by window size of 500, are shown in Figure 6.8, as an example. Only Lasso selects variables with high frequency, but all the variables are selected by Lasso. All the other methods select variables with very low frequency. For instance, GA-PLS selects all the variables less than 40% of the time, while SR yields 35% frequency on average. The I_C values are summarized in Table

6.4. Most of the selections are inconsistent. One of the main reasons is the fast dynamics of this industrial digester data. The dynamics of 100 MC runs are different from one another since MC simulations are generated by moving window approach. As a result, the window size is chosen to be 2000 with step size of 50. The new information contained in the next subset of data would be minimal.

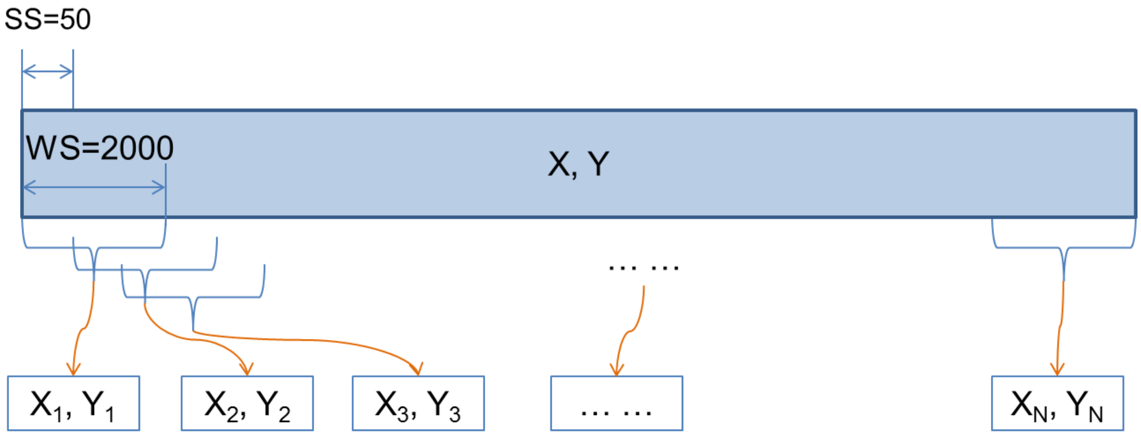


Figure 6.7: Schematic of moving window approach

Table 6.3: Moving window parameters

Step size (SS)	Window size (WS)
50	300
50	500
50	800
50	2000

6.3.2 Comparison results

To cope with the nature of such process, dynamic PLS (DPLS), an extension of the conventional PLS, is implemented along with variable selection. DPLS has been widely implemented in many dynamic applications [68]–[71]. In DPLS, original data matrix is augmented with lagged measurements to capture process dynamics. The amount of lagged measurement to be included is determined by the parameter, past horizon. Based

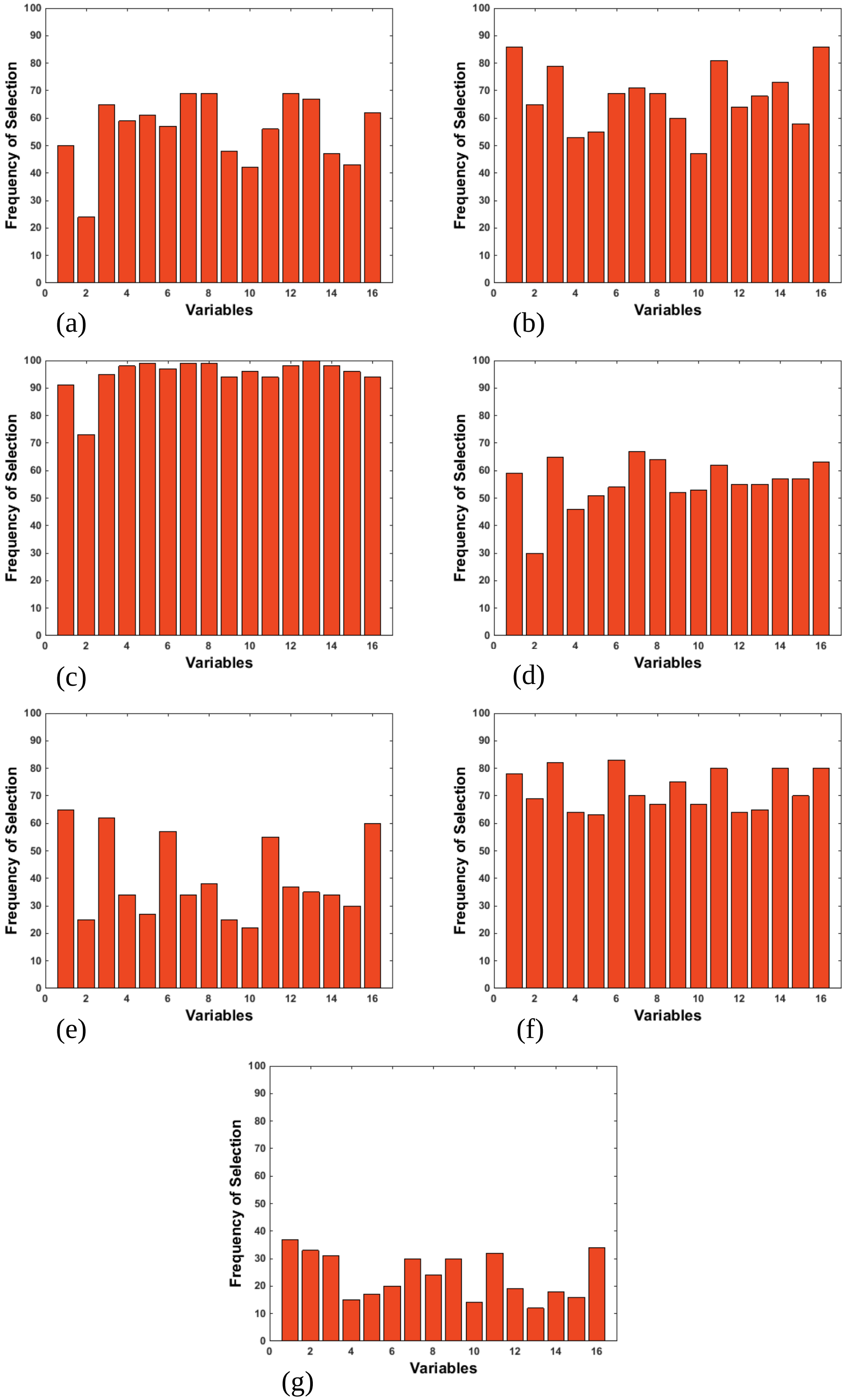


Figure 6.8: Variable selection frequency for industrial digester with WS=500: (a) PLS-VIP, (b) PLS-BETA, (c) Lasso, (d) UVE-PLS, (e) SR, (f) CARS-PLS, and (g) GA-PLS.

on [3], the past horizon parameter is chosen to be 22 for this case study. This results in a regressor matrix with dimension of 2000 by 352.

Table 6.4: Comparison of I_C values of different methods in the industrial Kamyr digester

Methods	I_C
PLS-VIP	0.15
PLS-BETA	0.31
Lasso	0.88
UVE-PLS	0.13
SR	0.07
CARS-PLS	0.37
GA-PLS	0.11

The selection frequencies of all seven methods are shown in Figure 6.9. The selection consistencies of some methods seem to improve compared to the ones with PLS, as shown in Figure 6.8. The vertical lines separate variables from different zones. The consistency index values are listed in Table 6.5. However, the selection performances are still not consistent enough, especially for the methods that performed well in the other case studies. The percentage improvement of reduced models compared to full DPLS models are shown in Figure 6.10. The average MSPE value for the full DPLS is 37.2%. PLS-BETA and PLS-VIP give the highest and second highest improvements of approximately 6%.

Except the high dynamic nature of the digester data, one other reason for such selection inconsistency may be caused by the repeated Kappa measurements used. The recorded Kappa measurements are not the true Kappa values corresponding to the process variables taken at the time instance, which means the information used to train the model are not fully accurate. Therefore, models are also built for unique Kappa measurements. For all the time stamps with repeated Kappa measurements, regressor information is neglected during modeling. This process results in a significant reduction in the number of samples. 1100 samples are used for training, and 450 samples are used for testing.

Table 6.5: Comparison of I_C values of different methods with DPLS in Kamyr digester

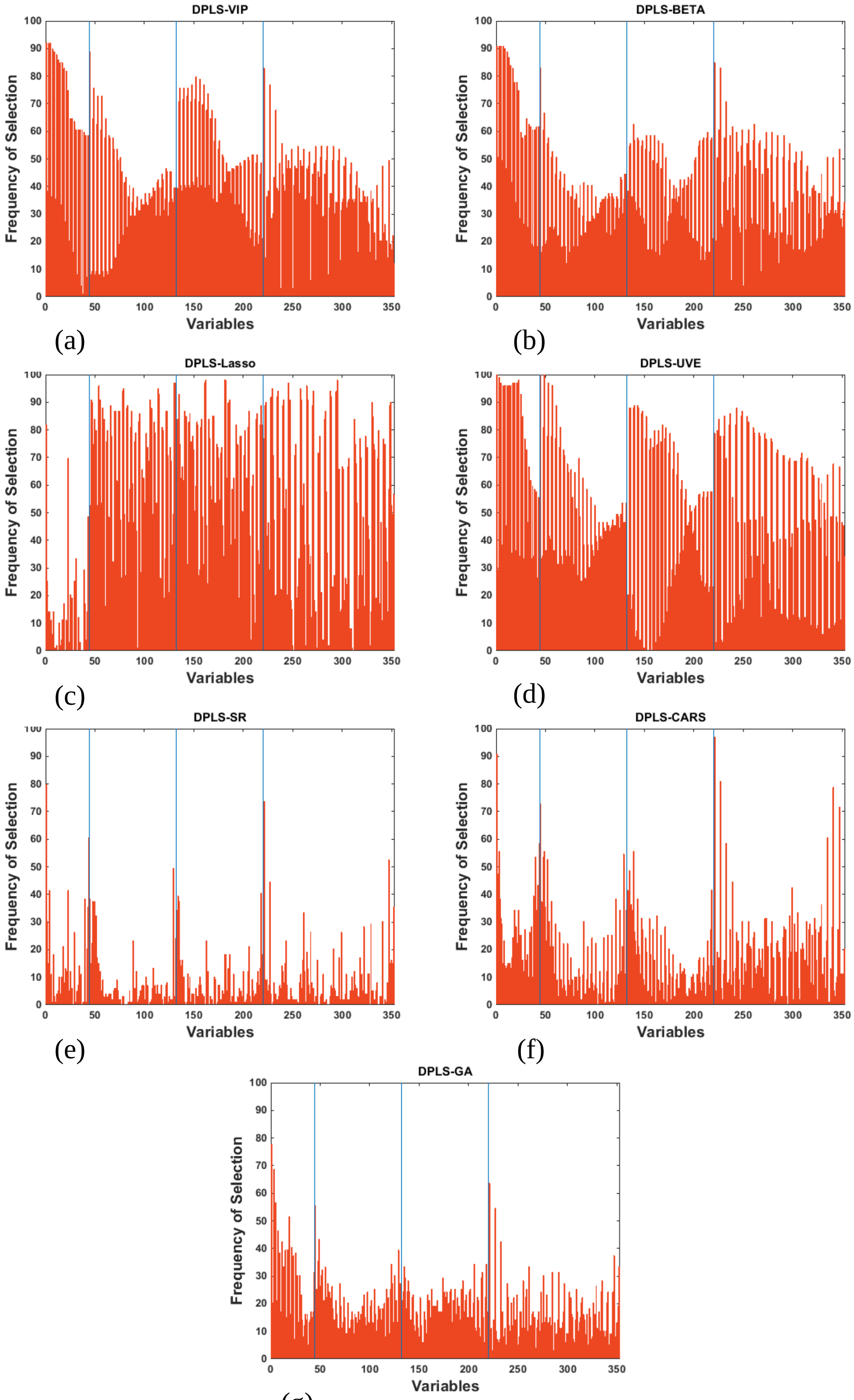
Methods	I_C
PLS-VIP	0.14
PLS-BETA	0.12
Lasso	0.38
UVE-PLS	0.25
SR	0.53
CARS-PLS	0.33
GA-PLS	0.23

Due to the limited availability of the unique Kappa measurements, MC simulation is not performed. The prediction performance has been significantly improved compared to the model using repeated Kappa number. The prediction error in MSPE is 25.2% for the full DPLS model. The selection performances of seven variable selection methods are shown in Figure 6.11. The results are presented in terms of binary number, with ‘1’ being variable selected. Only PLS-VIP and Lasso completely eliminate variables from lower cooking and impregnation zones, respectively. The corresponding prediction performances are shown in Figure 6.12. Lasso yields the highest improvement of 6.5% compared to the full DPLS model.

Even though the selection consistencies for this industrial digester case are not as good as other case studies, results showed that variable selection could still improve the prediction performance and reduce the model size and complexity.

6.4 Conclusions

In this chapter, seven variables selection methods for PLS-based soft sensors are implemented for the digester simulator case. Even though the prediction performances



(a) (b) (c) (d) (e) (f) (g)

Figure 6.9: Variable selection frequency for industrial digester: (a) DPLS-VIP, (b) DPLS-BETA, (c) RIVAL, (d) UVE-DPLS, (e) SR, (f) CARS-DPLS, and (g) GA-DPLS.

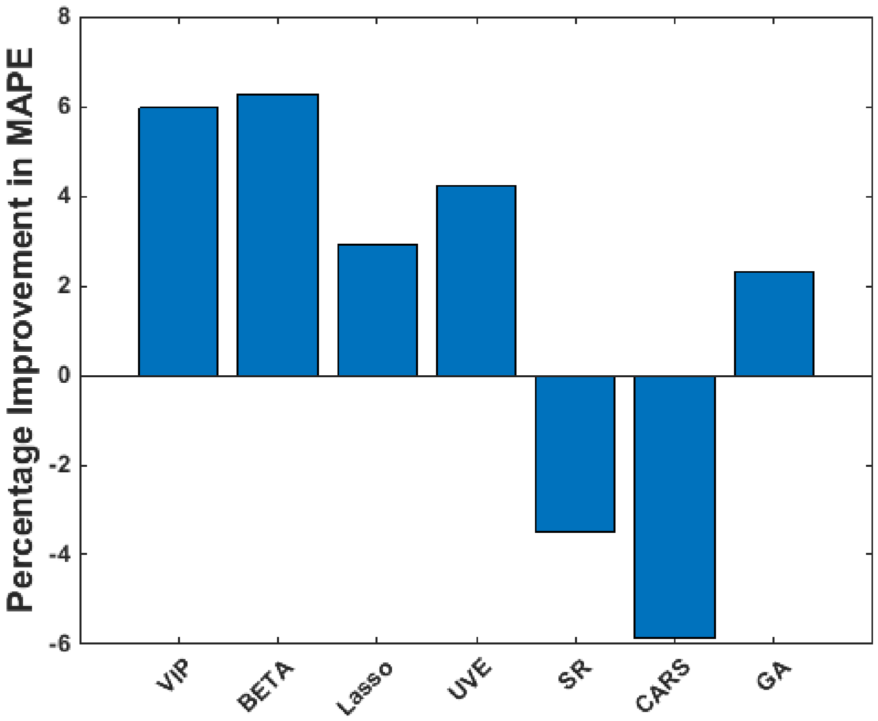


Figure 6.10: Percentage improvements in prediction error for industrial Kamyr digester

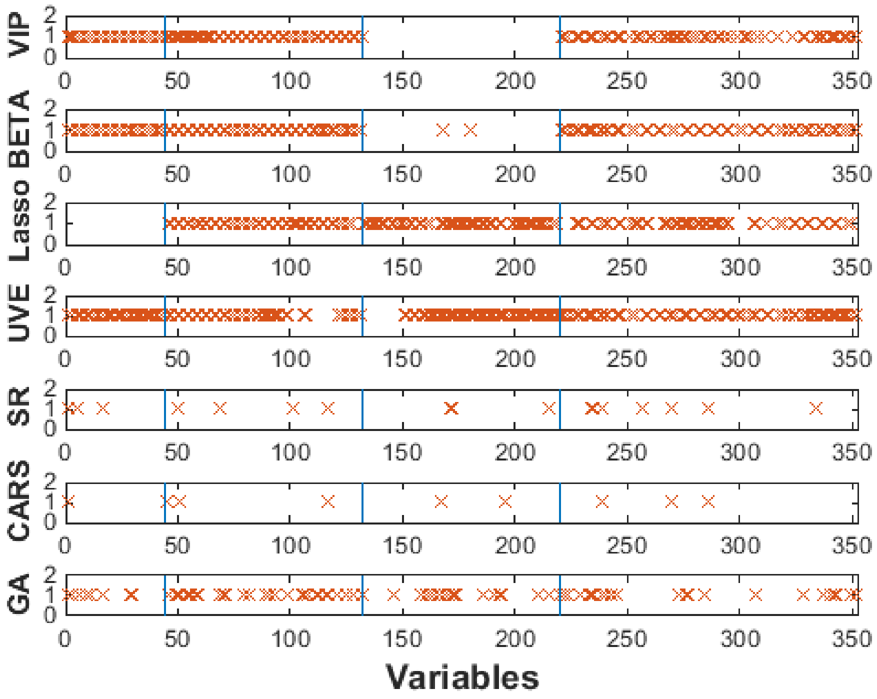


Figure 6.11: Variable selection for industrial digester with unique Kappa measurements

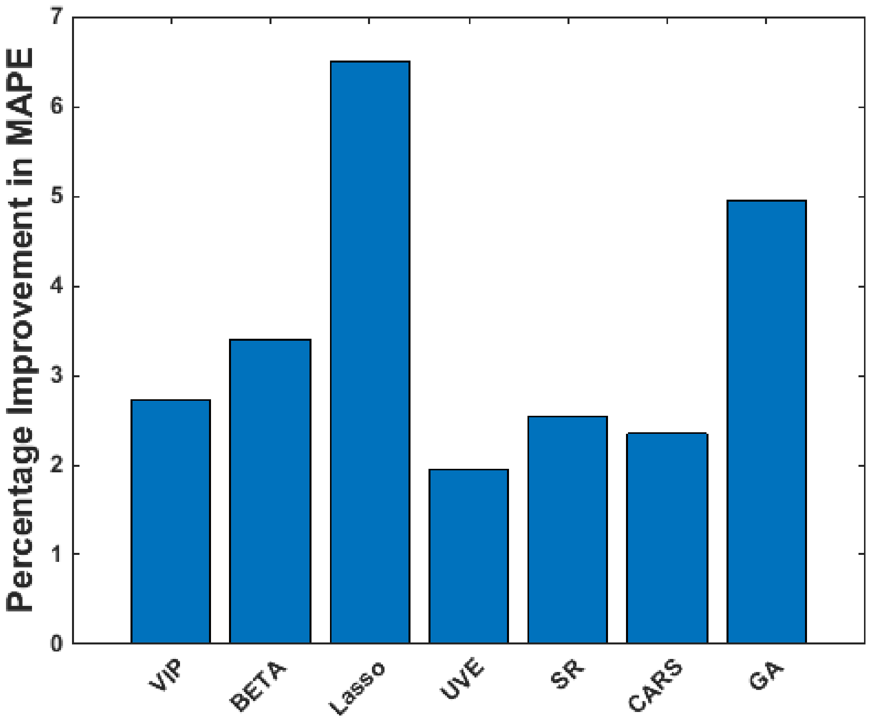


Figure 6.12: Percentage improvements in prediction error for industrial Kamyr digester with unique Kappa measurements

obtained by variable selection are not significant for the first three types of disturbances, but the selection results from the reduced models can be easily interpreted. The selected regressors reflect the type of disturbance introduced. For the combination case, the amount of prediction performance improvements is more notable, with 7.5% by Lasso and 4% by PLS-VIP. Furthermore, selection results also reveal the relevance of variable location. Regressors located closer to the blow line, where the Kappa measurements are taken, are identified to be more relevant compared to others.

For the industrial Kamyr digester case study, the process has fast dynamics. To cope with the nature of such process, DPLS is implemented along with variable selection. The selection performances are still inconsistent, but prediction performances are improved by 6% with PLS-BETA and PLS-VIP. Even with the 6% improvement by variable selection, the prediction performance is still quite poor, given MSPE value of 37% for the full model. Therefore, a new model is built using only the non-repeated Kappa measurements. The prediction error has been reduced to 25% by the full model. This is further boosted by 6.5% with Lasso. Despite the improvement obtained by using only the unique Kappa measurements, there is a drawback associated. The ratio of sample to variable (i.e., 1100 to 352 for training, and 450 to 352 for testing) is not favorable for DPLS with such small dataset. For industrial applications, if there are vast amount of historical data stored, the performances of DPLS could be further improved.

Chapter 7. Conclusions and Future Works

7.1 Conclusions

In this work, seven variable selection methods for PLS/DPLS-based soft sensor development were investigated using four case studies: one simulated case, one industrial polyester case, one digester simulator, and one industrial Kamyr digester case. Current studies usually adopt the readily available performance indicators, such as G, to evaluate variable selection performance. However, all the available performance indicators are only accessible when the ground truth of the data is available. To address the challenge that there is no published method that directly evaluates the variable selection performance when the true variable relevance is unknown, we proposed an information entropy based performance indicator (I_C) to directly assess the consistency of a variable selection method. It was shown that I_C is consistent with G, but more sensitive, for the simulated case studies where the true variable relevance is known. It was also shown that I_C is a good performance indicator for the industrial case study where the true variable relevance is unknown.

From the simulation case study, we were able to observe how each variable selection method was affected by different characteristics of the data. Generally speaking, as the magnitude of CBP increases, magnitude of SNR and PR decreases, more irrelevant variables are selected and less improvement are obtained for all variable selection methods. Overall, PLS-VIP was not only the most consistent variable selection method, but also the most reliable method since the most of the variables selected by PLS-VIP were truly relevant. For most of the cases tested, PLS-VIP had the highest or the second highest G and I_C values, which was consistent with the MAPE reduction on the reduced model over the full model where PLS-VIP performed the best or the second best most of the time.

For the industrial polyester case study, independent models were developed for two product quality variables. Two different synchronization methods were used to synchronize the process variables with product quality variables that were sampled at different frequencies. By comparing the prediction performances of the full models, it was found that synchronization by taking the average of the process measurements between production quality variables performed better than synchronization by integration in terms of MAPE from soft sensors. Outlier detection by PCA clearly identified the first samples taken at process start-up from 33 batches for the averaged data, along with 13 other samples for the integrated data. With outlier removal, the prediction performance obtained on averaged data still exceeded the performance on the integrated data. The results also demonstrated the advantages of applying variable selection before soft sensor development. However, variable selection does not necessarily remove all the highly correlated variables. Correlation removal after variable selection is highly recommended for industrial processes. Prediction errors from the best performing models after variable selection, along with correlation removal at 0.99, were reduced by 28% and 33% for acidity number and viscosity, respectively. Furthermore, PLS-VIP was the most consistent variable selection method among all the methods compared based on multiple criteria, I_C, R^2 and MAPE. The consistency between I_C and prediction performance indices demonstrated the effectiveness of I_C for evaluating the variable selection performance, when the true variable relevance information of the data is unknown. The variables selected by PLS-VIP also reflected the process knowledge.

The sensitivity of each variable selection methods on tuning parameters were also investigated in this work. The simulation case study indicated that most of the variable selection methods were insensitive to tuning parameters. However, the industrial polyester case study indicated that one should always make the effort to search for the optimal parameter settings when industrial applications are considered. The performance can be significantly improved when optimal tuning parameters were chosen, especially for Lasso and RIVAL, which are sensitive to tuning parameters.

It was also observed that the amounts of improvements obtained in the industrial polyester case study were more substantial than the ones in the simulation case study. The main reason for such discrepancy was most likely the difference in sample distribution. For the simulated case study, the process data followed multivariate normal distribution. For such cases, PLS was expected to handle the white noises in the data well, including the irrelevant variables. As a result, there was not much area for continued development over the full PLS model on simulated case. On the other hand, for the industrial polyester case study, process data showed clear non-Gaussian distribution and contained outliers, as shown in the score plots. For such cases, the performance of soft sensors can be greatly improved by implementation of variable selection, in which irrelevant variables that may contain high noise and outliers were eliminated.

For the digester simulated case study, soft sensors were built for data with four different types of disturbances introduced. Even though the prediction performances obtained by variable selection were not significant for the first three types of disturbances, but the selection results from the reduced models can be easily interpreted, as the selected regressors reflected the type of disturbance introduced. For the combination case, the amount of prediction performance improvements was more notable compared to the first three types of disturbances. Implementation of variable selection improved the prediction performance by eliminating the noises and irrelevant information. Consequently, there was 7.5% improvement in prediction performance by Lasso and 4% by PLS-VIP. Furthermore, selection results also revealed the relevance of variable location. Regressors located closer to the blow line (the ones from 35th CSTR), where the Kappa measurements are taken, were identified to be more relevant compared to others by all the variable selection methods.

For the industrial Kamyr digester case study, the fast dynamic nature of the process became problematic. The selection results over 100 MC simulations, which were obtained through moving window approach, were inconsistent. To cope with the nature of such process, DPLS was implemented along with variable selection. Even though the selection performances were still inconsistent, over 6% of prediction performance improvements were obtained by PLS-BETA and PLS-VIP. Since the prediction error of the full model yielded 37% in MSPE value, the best performer after variable selection still gave around 35% in MSPE. Therefore, an alternative approach was proposed to build a new model using only the non-repeated Kappa measurements. Consequently, the prediction error had been reduced to 25% by the full model. When variable selection was applied, the prediction performance was further boosted by 6.5% with Lasso. Despite the improvement obtained by using only the unique Kappa measurements, there is a major drawback associated with this approach. The ratio of sample size to variables is not favorable for DPLS with such small dataset, as discussed previously. For industrial applications, if there are vast amount of historical data stored, the performances of DPLS could be further enhanced, along with implementation of variable selection.

7.2 Future works

The future research directions which require further investigation are summarized in this section.

From literature studies, it is found that 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. This would be beneficial in a case with high dimensionality of data.

Furthermore, multi-criteria based variable selection can be considered as well. In our current work, only the predictive ability is considered and optimized. One possible approach is to apply a modeling power approach, which balances the predictive and descriptive abilities of model. Another approach is to adopt the Pareto Analysis to obtain the balance between model simplicity and predictive power.

As discovered in this work, variable selection can be implemented with dynamic modeling techniques to handle process with fast dynamics. Also, new variable selection methods that can deal with such process should be developed.

The main focus of the future works will be concentrated on variable selection for process monitoring. In soft sensor development, there are product quality variables available for indication of model performance. In process monitoring, fault detection is applied to the process variables only. Once the fault is detected, fault diagnosis must be carried out to find the root cause of the fault in order to take corrective action to prevent process failure. However, there is nothing that can directly relate process variables to model performance. Theoretically, fault detection performance can be greatly improved if only the variables that contribute to separation of normal data and faulty data are included. Furthermore, variable selection can be applied to fault diagnosis to select the subset of variables that are responsible for the abnormality in process. However, all of the current studies on variable selection for process monitoring require known normal and faulty data for model building. This would become a burden to industrial applications, where generation of faulty data is simply unaffordable. Another aspect of our work will focus on simulation of faulty data using the available normal data.

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. More importantly, variable selection can also be employed to select useful statistics in statistics pattern generation.

Bibliography

[1] P. Kadlec, B. Gabrys, and S. Strandt, "Data-driven soft sensors in the process industry," *Comput. Chem. Eng.*, vol. 33, no. 4, pp. 795–814, Apr. 2009.

[2] P. Kadlec, R. Grbić, and B. Gabrys, "Review of adaptation mechanisms for data-driven soft sensors," *Comput. Chem. Eng.*, vol. 35, no. 1, pp. 1–24, Jan. 2011.

[3] H. J. Galicia, Q. P. He, and J. Wang, "A reduced order soft sensor approach and its application to a continuous digester," *J. Process Control*, vol. 21, no. 4, pp. 489–500, Apr. 2011.

[4] H. J. Galicia, Q. Peter He, and J. Wang, "Comparison of the performance of a reduced-order dynamic PLS soft sensor with different updating schemes for digester control," *Control Eng. Pract.*, vol. 20, no. 8, pp. 747–760, Aug. 2012.

[5] J. Reunanen, "Overfitting in making comparisons between variable selection methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1371–1382, 2003.

[6] C. M. Andersen and R. Bro, "Variable selection in regression — a tutorial," *J. Chemom.*, vol. 24, no. 11–12, pp. 728–737, 2010.

[7] F. Alonso-Atienza, J. L. Rojo-Álvarez, A. Rosado-Muñoz, J. J. Vinagre, A. García-Alberola, and G. Camps-Valls, "Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 1956–1967, Feb. 2012.

[8] Q. Liu, A. H. Sung, M. Qiao, Z. Chen, J. Y. Yang, M. Q. Yang, X. Huang, and Y. Deng, "Comparison of feature selection and classification for MALDI-MS data.," *BMC Genomics*, vol. 10 Suppl 1, p. S3, Jan. 2009.

[9] K. Ghosh, M. Ramteke, and R. Srinivasan, "Optimal variable selection for effective statistical process monitoring," *Comput. Chem. Eng.*, vol. 60, pp. 260–276, Jan. 2014.

[10] E. K. Tang, P. N. Suganthan, and X. Yao, "Gene selection algorithms for microarray data based on least squares support vector machine.," *BMC*

Bioinformatics, vol. 7, p. 95, Jan. 2006.

[11] S.-W. Lin, K.-C. Ying, S.-C. Chen, and Z.-J. Lee, “Particle swarm optimization for parameter determination and feature selection of support vector machines,” *Expert Syst. Appl.*, vol. 35, no. 4, pp. 1817–1824, Nov. 2008.

[12] I.-G. Chong and C.-H. Jun, “Performance of some variable selection methods when multicollinearity is present,” *Chemom. Intell. Lab. Syst.*, vol. 78, no. 1–2, pp. 103–112, Jul. 2005.

[13] R. Gosselin, D. Rodrigue, and C. Duchesne, “A bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications,” *Chemom. Intell. Lab. Syst.*, vol. 100, no. 1, pp. 12–21, Jan. 2010.

[14] R. Leardi, R. Boggia, and M. Terrile, “Genetic algorithms as a strategy for feature selection,” *J. Chemom.*, vol. 6, no. 5, pp. 267–281, Sep. 1992.

[15] 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,” *Anal. Chim. Acta*, vol. 348, no. 1–3, pp. 71–86, Aug. 1997.

[16] L. H. Chiang and R. J. Pell, “Genetic algorithms combined with discriminant analysis for key variable identification,” *Anal. Sci.*, vol. 14, pp. 143–155, 2004.

[17] 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.,” *Anal. Chem.*, vol. 68, no. 21, pp. 3851–3858, Nov. 1996.

[18] Q.-J. Han, H.-L. Wu, C.-B. Cai, L. Xu, and R.-Q. Yu, “An ensemble of Monte Carlo uninformative variable elimination for wavelength selection,” *Anal. Chim. Acta*, vol. 612, no. 2, pp. 121–125, Apr. 2008.

[19] W. Cai, Y. Li, and X. Shao, “A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra,” *Chemom. Intell. Lab. Syst.*, vol. 90, no. 2, pp. 188–194, Feb. 2008.

[20] M.-D. Ma, J.-W. Ko, S.-J. Wang, M.-F. Wu, S.-S. Jang, S.-S. Shieh, and D. S.-H. Wong, “Development of adaptive soft sensor based on statistical identification of key variables,” *Control Eng. Pract.*, vol. 17, no. 9, pp. 1026–1034, Sep. 2009.

[21] M. Forina, S. Lanteri, M. Casale, and M. C. Cerrato Oliveros, “Stepwise orthogonalization of predictors in classification and regression techniques: An ‘old’ technique revisited,” *Chemom. Intell. Lab. Syst.*, vol. 87, no. 2, pp. 252–261,

Jun. 2007.

[22] H. Li, Y. Liang, Q. Xu, and D. Cao, “Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration.,” *Anal. Chim. Acta*, vol. 648, no. 1, pp. 77–84, Aug. 2009.

[23] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, Jan. 1996.

[24] P. Kump, E.-W. Bai, K. Chan, and W. Eichinger, “A robust method for detecting nuclear materials when the underlying model is inexact,” *Radiat. Meas.*, pp. 1–7, Jun. 2013.

[25] P. Kump, E.-W. Bai, K. Chan, B. Eichinger, and K. Li, “Variable selection via RIVAL (removing irrelevant variables amidst Lasso iterations) and its application to nuclear material detection,” *Automatica*, vol. 48, no. 9, pp. 2107–2115, Sep. 2012.

[26] P. Kump, “Passive detection of radionuclides from weak and poorly resolved gamma-ray energy spectra,” University of Iowa, 2012.

[27] J.-P. Gauchi and P. Chagnon, “Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data,” *Chemom. Intell. Lab. Syst.*, vol. 58, no. 2, pp. 171–193, Oct. 2001.

[28] 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,” *Anal. Chim. Acta*, vol. 339, pp. 63–77, 1997.

[29] H. Kaneko and K. Funatsu, “A new process variable and dynamics selection method based on a genetic algorithm-based wavelength selection method,” *AIChE J.*, vol. 58, no. 6, pp. 1829–1840, 2012.

[30] 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.

[31] 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,” *Anal. Bioanal. Chem.*, vol. 378, no. 5, pp. 1382–1387, 2004.

[32] 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," *Appl. Spectrosc.*, vol. 54, no. 1, pp. 148–152, Jan. 2000.

[33] 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.," *Anal. Sci.*, vol. 17, no. 2, pp. 319–22, Feb. 2001.

[34] S. Ye, D. Wang, and S. Min, "Successive projections algorithm combined with uninformative variable elimination for spectral variable selection," *Chemom. Intell. Lab. Syst.*, vol. 91, no. 2, pp. 194–199, Apr. 2008.

[35] C. Colombani, a Legarra, S. Fritz, F. Guillaume, P. Croiseau, V. Ducrocq, and C. Robert-Granié, "Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in French Holstein and Montbéliarde breeds.," *J. Dairy Sci.*, vol. 96, no. 1, pp. 575–591, Nov. 2012.

[36] I. Mareels, B. Ninness, and S. L., "A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification," in *14th IFAC Symposium on System Identification, 2006*, 2006, pp. 814–819.

[37] M. A. Rasmussen and R. Bro, "A tutorial on the Lasso approach to sparse modeling," *Chemom. Intell. Lab. Syst.*, vol. 119, pp. 21–31, Oct. 2012.

[38] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, and D.-S. Cao, "Model population analysis for variable selection," *J. Chemom.*, vol. 24, no. 7–8, pp. 418–423, Jul. 2010.

[39] H. Li, Y. Liang, and Q. Xu, "Model population analysis for statistical model comparison," *Chemom. Pract. Appl.*, pp. 3–21, 2012.

[40] T. Mehmood, H. Martens, S. Sæbø, J. Warringer, and L. Snipen, "A partial seast squares based algorithm for parsimonious variable selection.," *Algorithms Mol. Biol.*, vol. 6, no. 1, p. 27, Jan. 2011.

[41] F. Lindgren, B. Hansen, and W. Karcher, "Model validation by permutation tests: applications to variable selection," *J. Chemom.*, vol. 10, pp. 521–532, 1996.

[42] P. P. Roy and K. Roy, "On some aspects of variable selection for partial least squares regression models," *QSAR Comb. Sci.*, vol. 27, no. 3, pp. 302–313, 2008.

[43] P. A. Wisnewski, F. J. Doyle, and F. Kayihan, "Fundamental continuous-pulp-digester model for simulation and control," *AIChE J.*, vol. 43, no. 12, pp. 3175–3192, Dec. 1997.

[44] D. Wang and R. Srinivasan, "Data-driven soft sensor approach for quality prediction in a refining process," *IEEE Trans. Ind. Informatics*, vol. 6, no. 1, pp. 11–17, Feb. 2010.

[45] P. Kadlec and B. Gabrys, "Adaptive local learning soft sensor for inferential control support," in *2008 International Conference on Computational Intelligence for Modelling Control & Automation*, 2008, pp. 243–248.

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

[47] S. Wold, M. Sjostrom, and L. Eriksson, "PLS-regression : a basic tool of chemometrics," *Chemom. Intell. Lab. Syst.*, vol. 58, pp. 109–130, 2001.

[48] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning : data mining , inference and prediction," *2005 Springer Sci. Bus. Media, Inc.*, vol. 27, no. 2, pp. 83–85, 2005.

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

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

[51] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Anal. Chim. Acta*, vol. 185, pp. 1–17, 1986.

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

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

[54] 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.

[55] R. Leardi and A. Lupiáñez González, "Genetic algorithms applied to feature selection in PLS regression: how and when to use them," *Chemom. Intell. Lab. Syst.*, vol. 41, no. 2, pp. 195–207, Jul. 1998.

[56] L. Davis, *Genetic algorithms and simulated annealing*. United States: Morgan Kaufman Publishers, Inc., 1987.

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

[58] D. E. Goldberg and J. H. Holland, “Genetic algorithms and machine learning,” in *Machine Learning*, vol. 3, Kluwer Academic Publishers, 1988, pp. 95–99.

[59] 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,” *J. Process Control*, vol. 19, no. 3, pp. 520–529, Mar. 2009.

[60] P. Facco, F. Bezzo, and M. Barolo, “Nearest-neighbor method for the automatic maintenance of multivariate statistical soft sensors in batch processing,” *Ind. Eng. Chem. Res.*, vol. 49, no. 5, pp. 2336–2347, Mar. 2010.

[61] S. Wold, N. Kettaneh, H. Fridén, and A. Holmberg, “Modelling and diagnostics of batch processes and analogous kinetic experiments,” *Chemom. Intell. Lab. Syst.*, vol. 44, no. 1–2, pp. 331–340, Dec. 1998.

[62] P. Nomikos and J. F. MacGregor, “Multivariate SPC charts for monitoring batch processes,” *Technometrics*, vol. 37, no. 1, pp. 41–59, Feb. 1995.

[63] P. Nomikos and J. F. MacGregor, “Multi-way partial least squares in monitoring batch processes,” *Chemom. Intell. Lab. Syst.*, vol. 30, no. 1, pp. 97–108, Nov. 1995.

[64] Q. P. He and J. Wang, “Statistics pattern analysis: A new process monitoring framework and its application to semiconductor batch processes,” *AIChE J.*, vol. 57, no. 1, pp. 107–121, 2011.

[65] J. Wang and Q. P. He, “Multivariate statistical process monitoring based on statistics pattern analysis,” *Ind. Eng. Chem. Res.*, vol. 49, no. 17, pp. 7858–7869, 2010.

[66] P. A. Wisnewski, “Inferential control using high-order process models with application to a continuous pulp digester,” Purdue University, 1997.

[67] B. Joseph and C. Brosilow, “Inferential control of processes: Part III. Construction of optimal and suboptimal dynamic estimators,” *AIChE J.*, vol. 24, no. 3, pp. 500–509, May 1978.

[68] S. Park and C. Han, “A nonlinear soft sensor based on multivariate smoothing procedure for quality estimation in distillation columns,” *Comput. Chem. Eng.*, vol. 24, no. 2–7, pp. 871–877, Jul. 2000.

[69] L. H. Chiang, E. L. Russell, and R. D. Braatz, “Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and

principal component analysis," *Chemom. Intell. Lab. Syst.*, vol. 50, no. 2, pp. 243–252, Mar. 2000.

[70] B. Lin, B. Recke, J. K. H. Knudsen, and S. B. Jorgensen, "A systematic approach for soft sensor development," *Comput. Chem. Eng.*, vol. 31, no. 5–6, pp. 419–425, May 2007.

[71] J. V. Kresta, T. E. Marlin, and J. F. MacGregor, "Development of inferential process models using PLS," *Comput. Chem. Eng.*, vol. 18, no. 7, pp. 597–611, Jul. 1994.

Appendices

Appendix A: Simulation Case Results

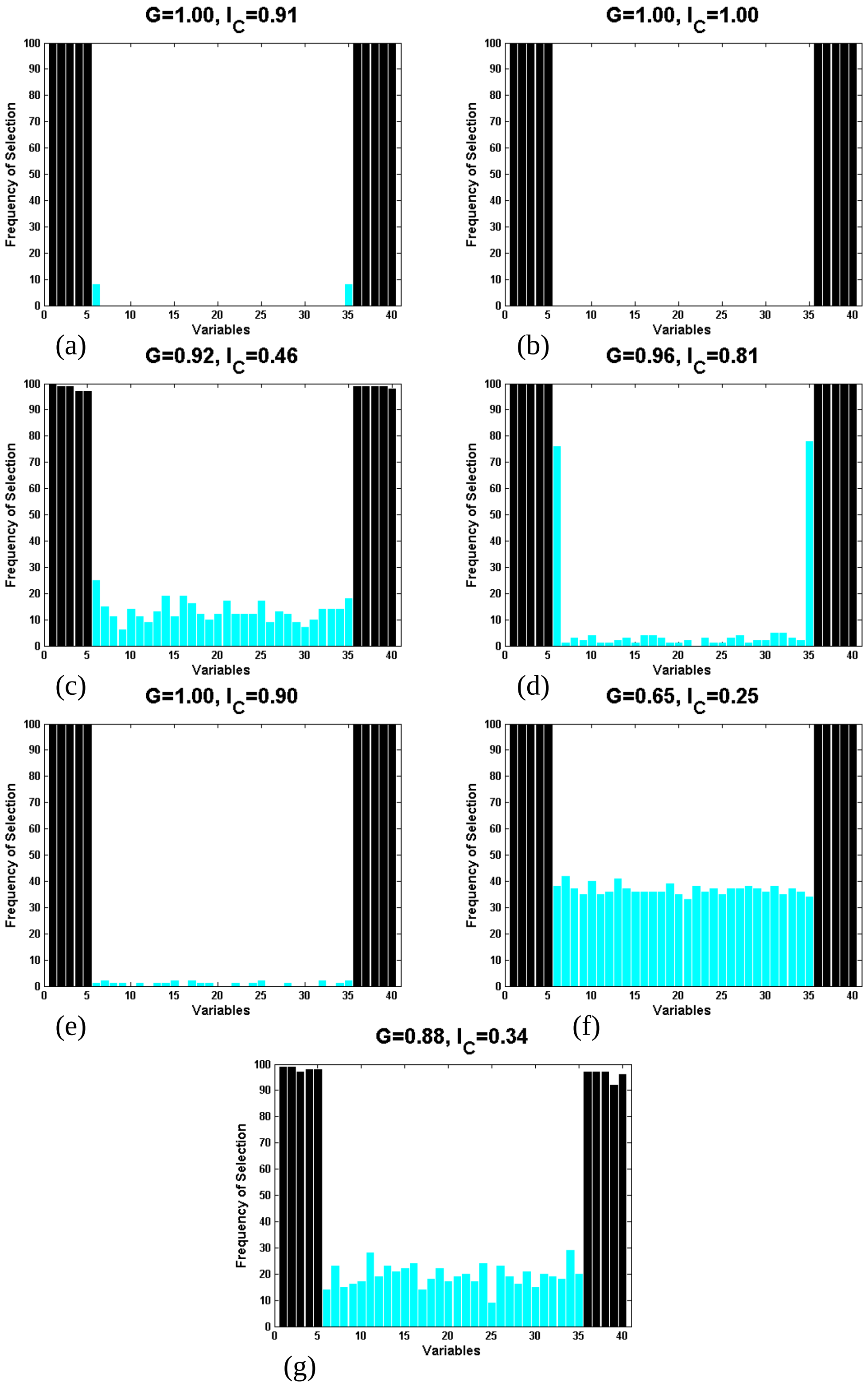
The selection frequency results of each method among 100 MC simulations for the six representative cases for simulated data are shown in Figure A.1 to Figure A.6, respectively. The G and I_C values for each method are also labeled on their respective figures. The results from the other five cases agree with the ones from Case 3 as demonstrated in Section 5.2.2. PLS-VIP is either the best or the second best performer, revealed by both G and I_C values. Furthermore, I_C is a more indicative performance index for the given selection results, as compared to G.

The model size of the six representative cases are summarized in Table A.1, in terms of average number of variables selected among 100 MC simulations, along with the corresponding standard deviations.

Table A.1 Average and standard deviation of numbers of variables selected simulation case.

Methods	Cases					
	1	2	3	4	5	6
PLS-VIP	11 ± 0.6	15 ± 1.2	16 ± 0.7	9 ± 1.4	10 ± 1.1	18 ± 2.3
PLS-BETA	10 ± 0	8 ± 0.4	9 ± 1.3	5 ± 1.3	7 ± 1.1	11 ± 1.8
RIVAL	14 ± 2.0	12 ± 1.7	9 ± 2.0	8 ± 1.4	29 ± 4.4	22 ± 3.6
UVE-PLS	12 ± 1.3	27 ± 3.5	20 ± 3.2	12 ± 2.0	8 ± 1.3	12 ± 1.7
SR	10 ± 0.5	9 ± 1.5	5 ± 1.0	4 ± 1.1	7 ± 1.1	6 ± 1.2
CARS-PLS	21 ± 13.4	9 ± 5.2	10 ± 6.3	8 ± 5.7	15 ± 4.9	14 ± 5.0
GA-PLS	16 ± 2.7	13 ± 2.3	11 ± 2.9	7 ± 1.9	24 ± 4.2	23 ± 4.4

Results for all 108 simulation cases are shown in Table A.2 to Table A.5, in terms of G, I_C, R^2, and percentage improvement of average MAPE values compared to full models, respectively. In Table A.5, only full models are given in average MAPE values, instead of percentage improvement, to provide the magnitude of prediction performance for both full models and reduced models.



(a) (b) (c) (d) (e) (f) (g)

Figure A.1: Variable selection frequency for Case 1: (a) PLS-VIP, (b) PLS-BETA, (c) RIVAL, (d) UVE-PLS, (e) SR, (f) CARS-PLS, and (g) GA-PLS. The dark-colored bars represent true relevant variables, while the light-colored bars representing true irrelevant variables.

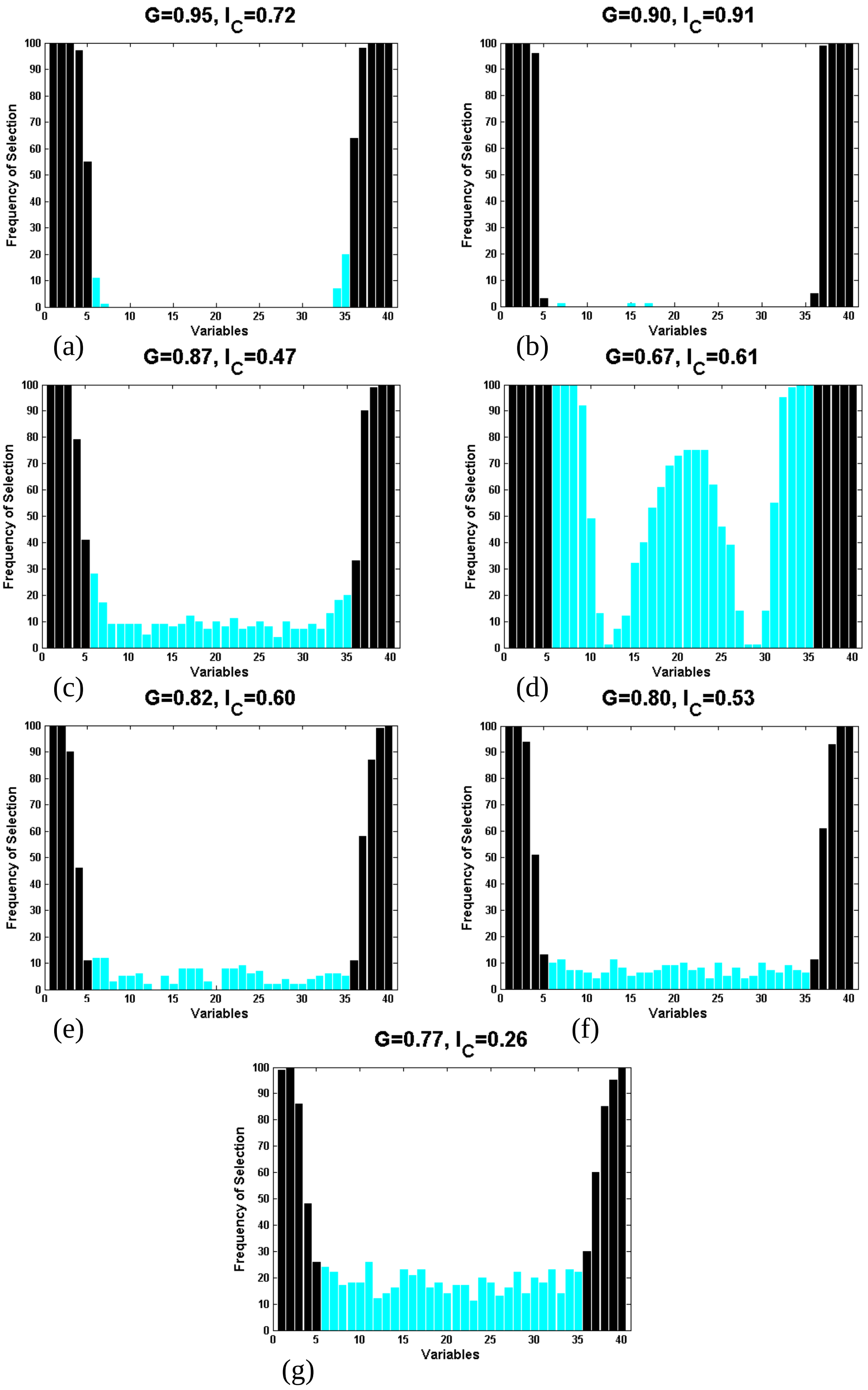
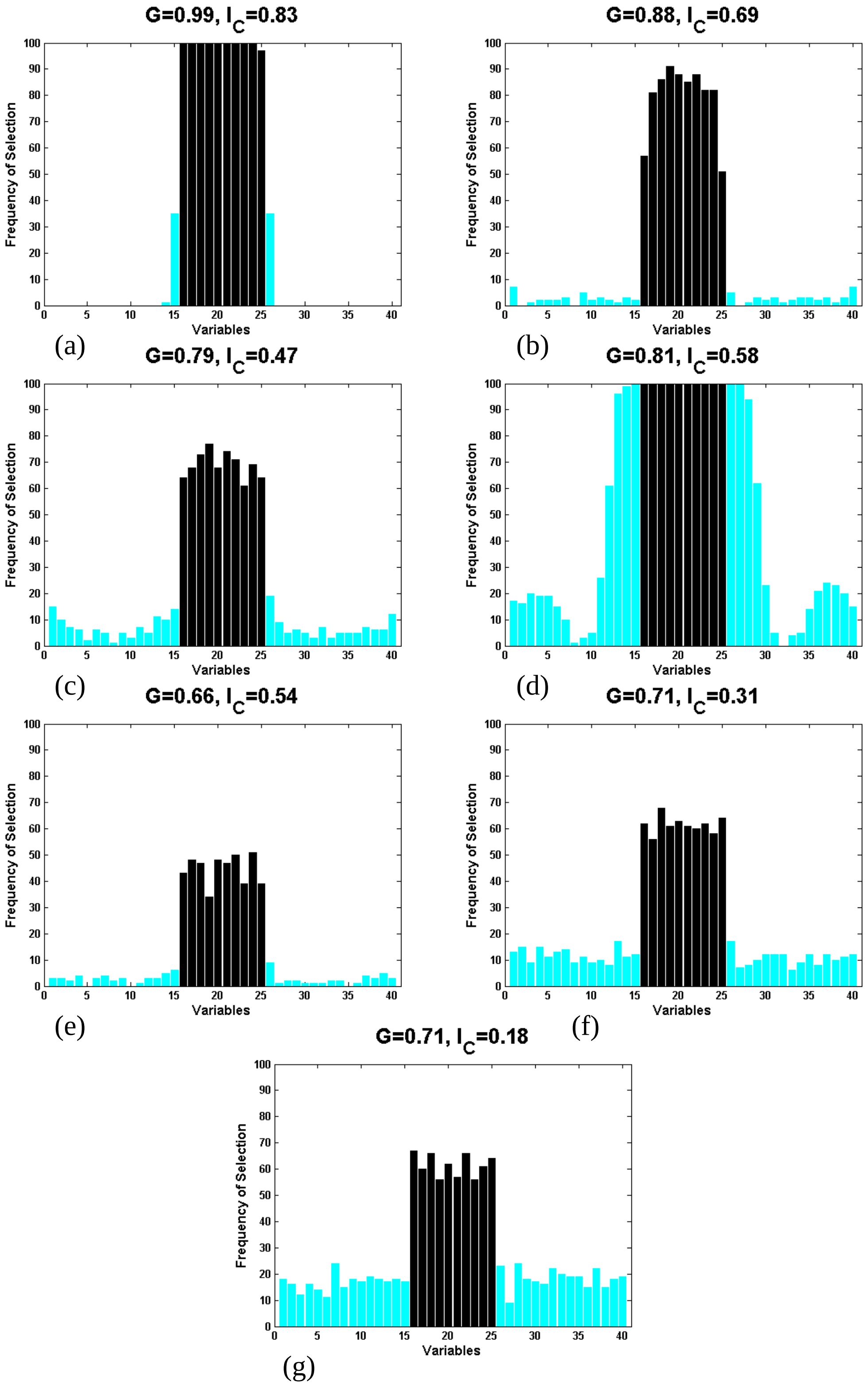


Figure A.2: Variable selection frequency for Case 2: (a) PLS-VIP, (b) PLS-BETA, (c) RIVAL, (d) UVE-PLS, (e) SR, (f) CARS-PLS, and (g) GA-PLS. The dark-colored bars represent true relevant variables, while the light-colored bars representing true irrelevant variables.



(a) G=0.99, I_C=0.83

(b) G=0.88, I_C=0.69

(c) G=0.79, I_C=0.47

(d) G=0.81, I_C=0.58

(e) G=0.66, I_C=0.54

(f) G=0.71, I_C=0.31

(g) G=0.71, I_C=0.18

Figure A.3: Variable selection frequency for Case 3: (a) PLS-VIP, (b) PLS-BETA, (c) RIVAL, (d) UVE-PLS, (e) SR, (f) CARS-PLS, and (g) GA-PLS. The dark-colored bars represent true relevant variables, while the light-colored bars representing true irrelevant variables.

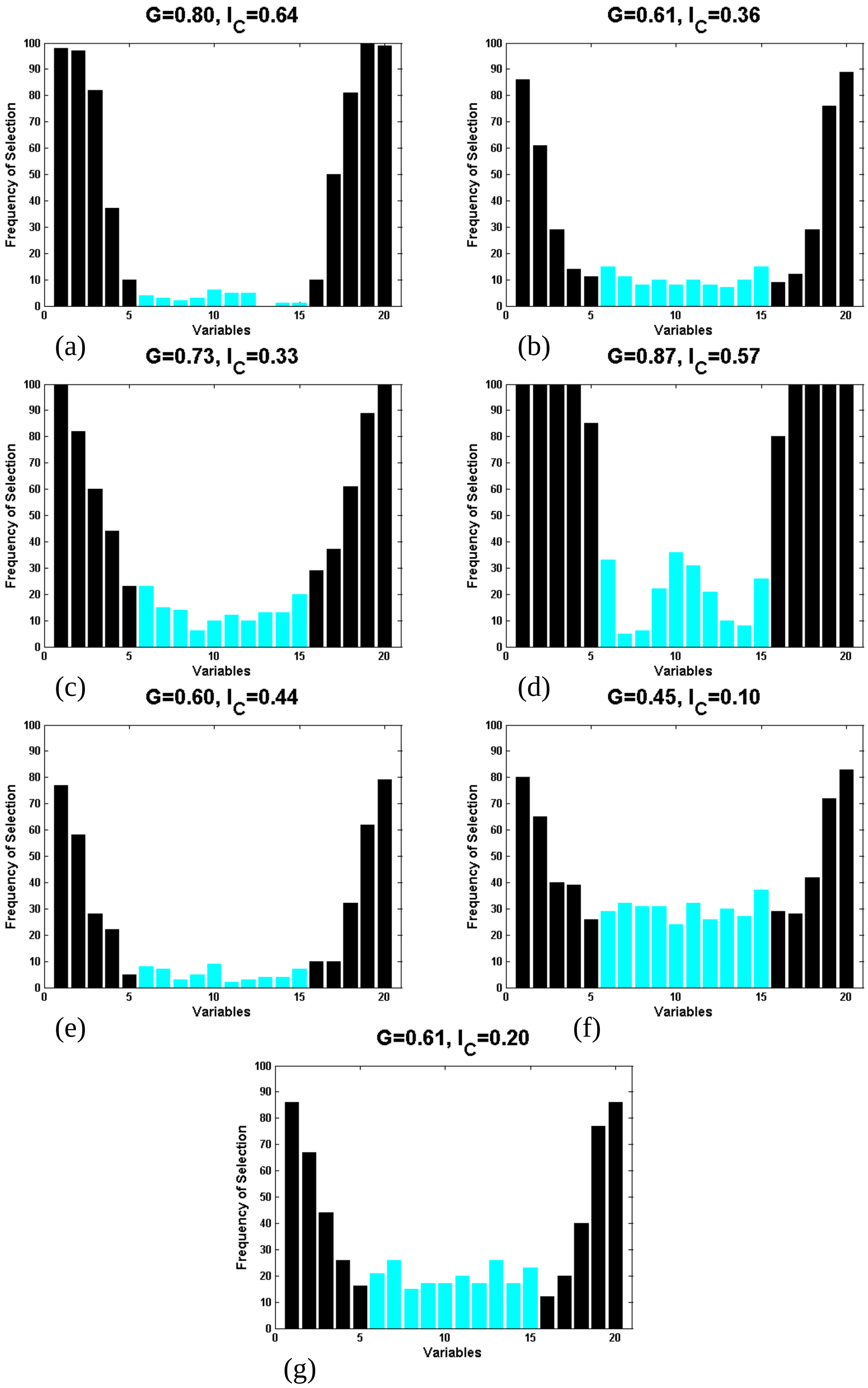


Figure A.4: Variable selection frequency for Case 4: (a) PLS-VIP, (b) PLS-BETA, (c) RIVAL, (d) UVE-PLS, (e) SR, (f) CARS-PLS, and (g) GA-PLS. The dark-colored bars represent true relevant variables, while the light-colored bars representing true irrelevant variables.

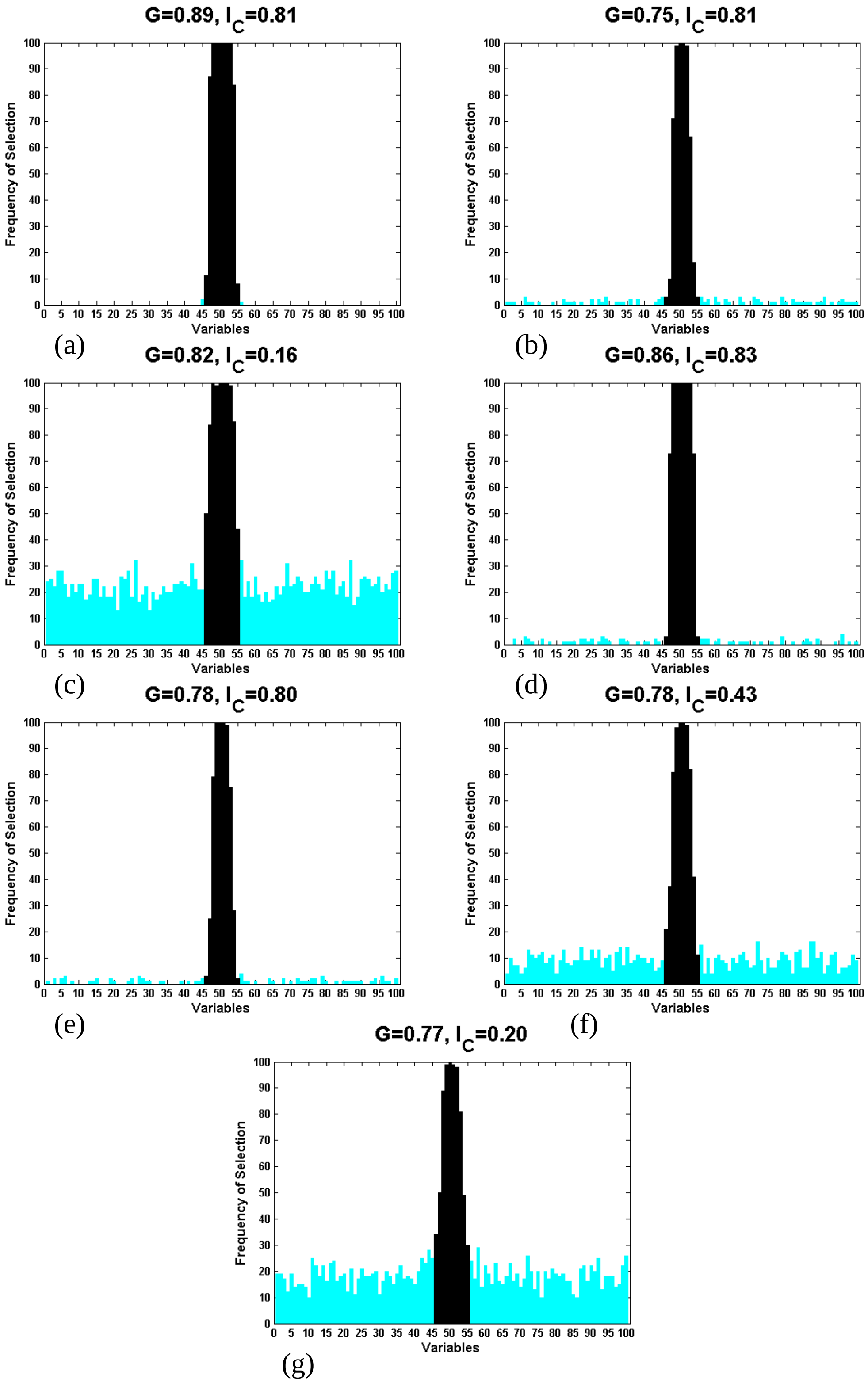


Figure A.5: Variable selection frequency for Case 5: (a) PLS-VIP, (b) PLS-BETA, (c) RIVAL, (d) UVE-PLS, (e) SR, (f) CARS-PLS, and (g) GA-PLS. The dark-colored bars represent true relevant variables, while the light-colored bars representing true irrelevant variables.

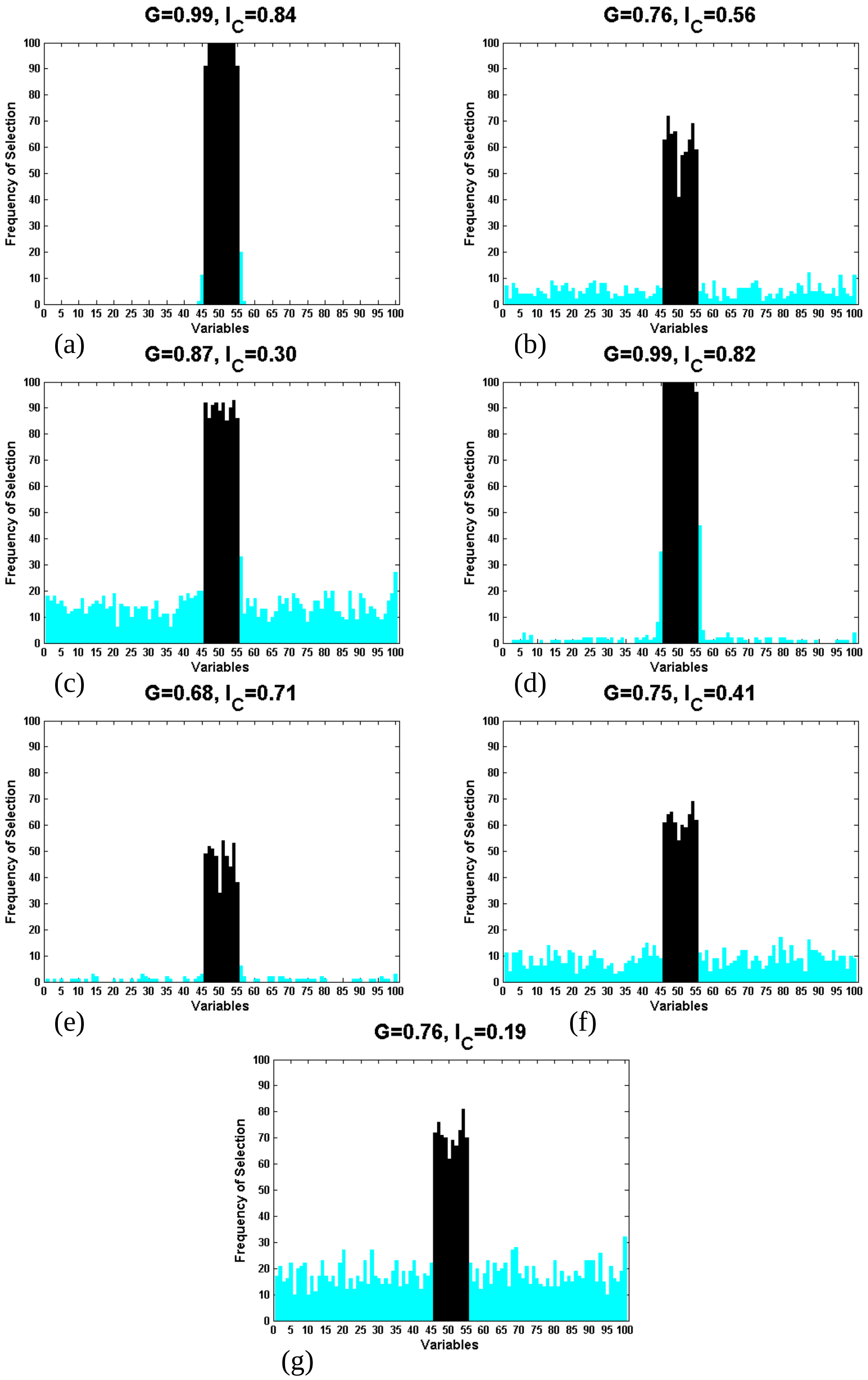


Figure A.6: Variable selection frequency for Case 6: (a) PLS-VIP, (b) PLS-BETA, (c) RIVAL, (d) UVE-PLS, (e) SR, (f) CARS-PLS, and (g) GA-PLS. The dark-colored bars represent true relevant variables, while the light-colored bars representing true irrelevant variables.

Table A.2: Average G values for simulation case. V: PLS-VIP; B: PLS-BETA; R: RI-VAL; U: UVE-PLS; S: SR; C: CARS-PLS; G: GA-PLS.

			$CBP = 0.5$							$CBP = 0.7$							$CBP = 0.9$						
SRC	PR	SNR	V	B	R	U	S	C	G	V	B	R	U	S	C	G	V	B	R	U	S	C	G
EM	0.5	3.03	0.97	1	0.89	0.89	1.00	0.15	0.88	0.97	1	0.90	0.81	1.00	0.07	0.86	0.99	0.95	0.89	0.75	0.88	0.27	0.82
		1.35	0.96	0.98	0.86	0.95	0.98	0.16	0.87	0.96	0.91	0.88	0.86	0.89	0.21	0.84	0.98	0.32	0.79	0.84	0.67	0.46	0.73
		0.82	0.96	0.87	0.82	0.96	0.89	0.17	0.84	0.96	0.73	0.83	0.92	0.78	0.30	0.77	0.95	0.59	0.72	0.90	0.57	0.50	0.65
	0.25	3.03	0.99	1	0.92	0.96	1.00	0.63	0.88	0.97	1	0.94	0.92	1.00	0.72	0.88	0.91	1.00	0.94	0.70	0.89	0.78	0.83
		1.35	0.99	0.99	0.90	0.98	0.99	0.62	0.88	0.97	0.95	0.92	0.94	0.88	0.68	0.84	**0.99**	**0.88**	**0.79**	**0.81**	**0.66**	**0.71**	**0.71**
		0.82	0.99	0.88	0.88	0.98	0.88	0.63	0.85	0.97	0.80	0.86	0.96	0.74	0.65	0.77	0.89	0.69	0.77	0.87	0.56	0.58	0.64
	0.1	3.03	0.99	1	0.92	0.99	1.00	0.96	0.90	0.97	1.00	0.91	0.97	1.00	0.96	0.89	0.90	1.00	0.94	0.88	0.82	0.95	0.83
		1.35	0.99	0.99	0.90	0.99	0.98	0.93	0.88	0.97	0.93	0.84	0.98	0.82	0.88	0.73	0.90	0.99	0.87	0.91	0.61	0.79	0.70
		0.82	0.97	0.83	0.87	0.99	0.81	0.83	0.83	**0.99**	**0.76**	**0.87**	**0.99**	**0.68**	**0.75**	**0.76**	0.89	0.82	0.77	0.93	0.52	0.65	0.66
EE	0.5	3.03	0.99	1	0.92	0.87	0.99	0.09	0.88	0.99	1	0.79	0.71	1.00	0.13	0.87	0.90	0.98	0.93	0.65	0.94	0.17	0.83
		1.35	0.99	0.99	0.87	0.93	0.99	0.08	0.88	0.98	0.95	0.90	0.81	0.94	0.15	0.85	0.83	0.14	0.85	0.79	0.71	0.38	0.74
		0.82	0.97	0.90	0.86	0.95	0.91	0.17	0.84	0.96	0.40	0.86	0.87	0.79	0.25	0.78	0.73	0.11	0.76	0.87	0.60	0.45	0.68
	0.25	3.03	**1.00**	**1**	**0.92**	**0.96**	**1.00**	**0.65**	**0.88**	0.95	1	0.61	0.92	1.00	0.59	0.88	0.90	1.00	0.95	0.62	0.95	0.72	0.85
		1.35	0.99	0.99	0.90	0.97	0.99	0.61	0.87	0.95	0.98	0.92	0.94	0.94	0.64	0.86	0.90	0.96	0.88	0.74	0.75	0.77	0.76
		0.82	0.98	0.91	0.88	0.98	0.90	0.67	0.84	0.95	0.84	0.89	0.96	0.78	0.66	0.80	0.87	0.84	0.82	0.82	0.62	0.67	0.67
	0.1	3.03	0.98	1	0.90	0.99	1.00	0.96	0.90	0.96	1	0.89	0.98	1.00	0.96	0.90	0.88	1.00	0.94	0.90	0.93	0.98	0.86
		1.35	0.98	0.99	0.90	0.99	0.99	0.93	0.89	0.96	0.98	0.89	0.98	0.89	0.91	0.86	0.87	0.99	0.89	0.92	0.69	0.88	0.75
		0.82	0.96	0.87	0.87	0.99	0.85	0.85	0.86	0.94	0.79	0.86	0.99	0.72	0.78	0.81	0.86	0.94	0.84	0.93	0.57	0.73	0.68
UM	0.5	3.03	0.77	0.87	0.81	0.88	0.88	0.14	0.80	0.84	0.85	0.81	0.85	0.86	0.23	0.79	0.90	0.78	0.81	0.75	0.78	0.39	0.75
		1.35	0.78	0.79	0.77	0.84	0.81	0.39	0.76	0.85	0.74	0.78	0.88	0.76	0.47	0.75	0.91	0.65	0.75	0.87	0.61	0.51	0.65
		0.82	0.78	0.72	0.74	0.80	0.73	0.44	0.71	0.85	0.67	0.75	0.86	0.66	0.52	0.67	0.92	0.56	0.71	0.88	0.54	0.50	0.61
	0.25	3.03	0.86	0.87	0.86	0.89	0.89	0.64	0.81	0.97	0.85	0.86	0.96	0.84	0.73	0.80	0.92	0.87	0.85	0.71	0.77	0.79	0.74
		1.35	0.87	0.77	0.80	0.84	0.78	0.65	0.76	0.97	0.72	0.82	0.93	0.73	0.69	0.73	0.92	0.75	0.79	0.83	0.57	0.67	0.65
		0.82	0.87	0.68	0.77	0.78	0.70	0.66	0.71	0.97	0.63	0.77	0.88	0.66	0.63	0.67	0.91	0.65	0.72	0.89	0.49	0.57	0.59
	0.1	3.03	0.95	0.86	0.86	0.91	0.88	0.85	0.84	0.98	0.86	0.87	0.99	0.84	0.83	0.80	0.90	0.90	0.87	0.89	0.75	0.83	0.73
		1.35	**0.89**	**0.75**	**0.82**	**0.86**	**0.78**	**0.78**	**0.77**	0.98	0.72	0.82	0.95	0.73	0.75	0.73	0.90	0.81	0.80	0.92	0.56	0.71	0.63
		0.82	0.94	0.65	0.78	0.79	0.69	0.70	0.72	0.96	0.59	0.78	0.89	0.61	0.67	0.67	0.89	0.74	0.74	0.94	0.48	0.61	0.58
UE	0.5	3.03	0.82	0.89	0.83	0.87	0.89	0.14	0.81	0.89	0.87	0.84	0.83	0.86	0.27	0.79	0.83	0.55	0.84	0.63	0.79	0.41	0.73
		1.35	0.82	0.80	0.78	0.85	0.81	0.30	0.77	0.89	0.78	0.81	0.88	0.77	0.47	0.75	0.82	0.70	0.77	0.78	0.70	0.47	0.65
		0.82	0.82	0.73	0.77	0.81	0.75	0.49	0.71	0.89	0.70	0.79	0.88	0.70	0.53	0.70	**0.80**	**0.61**	**0.73**	**0.86**	**0.60**	**0.45**	**0.61**
	0.25	3.03	0.91	0.88	0.84	0.89	0.89	0.61	0.81	0.97	0.87	0.87	0.96	0.87	0.70	0.80	**0.95**	**0.90**	**0.87**	**0.67**	**0.82**	**0.80**	**0.77**
		1.35	0.91	0.79	0.81	0.86	0.81	0.72	0.78	0.97	0.79	0.83	0.95	0.77	0.70	0.76	0.91	0.82	0.82	0.77	0.65	0.72	0.68
		0.82	0.91	0.72	0.78	0.81	0.73	0.67	0.74	0.96	0.69	0.80	0.91	0.67	0.65	0.69	0.89	0.74	0.77	0.84	0.59	0.64	0.63
	0.1	3.03	0.97	0.88	0.85	0.90	0.89	0.86	0.84	0.97	0.87	0.87	0.99	0.87	0.85	0.82	0.87	0.90	0.88	0.91	0.78	0.86	0.77
		1.35	0.96	0.77	0.82	0.87	0.80	0.79	0.77	0.97	0.77	0.83	0.96	0.77	0.78	0.75	0.87	0.86	0.82	0.93	0.66	0.76	0.69
		0.82	0.94	0.69	0.78	0.81	0.74	0.73	0.74	0.95	0.67	0.79	0.91	0.68	0.71	0.69	0.85	0.73	0.78	0.95	0.53	0.70	0.63

Table A.3: I_C values for simulation case. V: PLS-VIP; B: PLS-BETA; R: RIVAL; U: UVE-PLS; S: SR; C: CARS-PLS; G: GA-PLS.

			$CBP = 0.5$							$CBP = 0.7$							$CBP = 0.9$						
SRC	PR	SNR	V	B	R	U	S	C	G	V	B	R	U	S	C	G	V	B	R	U	S	C	G
EM	0.5	3.03	0.87	1	0.55	0.74	0.94	0.82	0.55	0.86	1	0.58	0.71	0.96	0.92	0.52	0.92	0.73	0.57	0.72	0.54	0.63	0.40
		1.35	0.85	0.86	0.52	0.73	0.82	0.81	0.52	0.85	0.58	0.53	0.71	0.56	0.71	0.41	0.87	0.73	0.32	0.71	0.29	0.16	0.22
		0.82	0.82	0.49	0.49	0.75	0.54	0.72	0.42	0.83	0.29	0.42	0.68	0.32	0.46	0.28	0.79	0.05	0.20	0.65	0.27	0.04	0.17
	0.25	3.03	0.88	1	0.44	0.82	0.90	0.25	0.32	0.83	1	0.57	0.73	0.91	0.26	0.33	0.83	0.97	0.60	0.56	0.60	0.28	0.28
		1.35	0.87	0.86	0.42	0.76	0.88	0.22	0.33	0.82	0.78	0.51	0.74	0.64	0.19	0.29	**0.83**	**0.69**	**0.47**	**0.58**	**0.54**	**0.31**	**0.18**
		0.82	0.84	0.51	0.36	0.77	0.60	0.16	0.29	0.81	0.30	0.33	0.73	0.55	0.12	0.23	0.80	0.30	0.36	0.65	0.55	0.18	0.18
	0.1	3.03	0.84	1	0.51	0.85	0.86	0.56	0.26	0.79	0.99	0.61	0.83	0.87	0.58	0.24	0.80	0.98	0.67	0.60	0.74	0.71	0.22
		1.35	0.84	0.86	0.28	0.84	0.85	0.46	0.23	0.78	0.80	0.54	0.82	0.76	0.45	0.22	0.77	0.84	0.55	0.69	0.69	0.57	0.20
		0.82	0.71	0.67	0.32	0.85	0.75	0.32	0.19	**0.84**	**0.56**	**0.30**	**0.82**	**0.71**	**0.41**	**0.19**	0.77	0.65	0.51	0.73	0.70	0.48	0.22
EE	0.5	3.03	0.96	1	0.64	0.71	0.93	0.89	0.54	0.91	1	0.41	0.68	0.95	0.84	0.50	0.62	0.85	0.66	0.76	0.72	0.76	0.42
		1.35	0.93	0.91	0.55	0.72	0.93	0.89	0.52	0.86	0.70	0.59	0.69	0.68	0.80	0.46	0.46	0.92	0.45	0.70	0.33	0.29	0.22
		0.82	0.84	0.56	0.53	0.72	0.61	0.74	0.43	0.77	0.73	0.51	0.68	0.39	0.55	0.30	0.27	0.91	0.30	0.68	0.27	0.08	0.18
	0.25	3.03	**0.91**	**1**	**0.46**	**0.81**	**0.90**	**0.25**	**0.34**	0.81	1	0.43	0.71	0.91	0.26	0.34	0.74	0.96	0.66	0.64	0.75	0.25	0.30
		1.35	0.83	0.92	0.41	0.77	0.89	0.24	0.31	0.79	0.81	0.44	0.73	0.71	0.21	0.32	0.71	0.83	0.55	0.60	0.48	0.32	0.21
		0.82	0.81	0.58	0.35	0.78	0.63	0.17	0.29	0.76	0.44	0.37	0.73	0.56	0.14	0.26	0.66	0.53	0.43	0.63	0.49	0.36	0.18
	0.1	3.03	0.83	1	0.52	0.86	0.86	0.58	0.27	0.77	1	0.62	0.82	0.87	0.61	0.26	0.69	0.96	0.75	0.66	0.83	0.77	0.22
		1.35	0.82	0.89	0.24	0.85	.85	0.47	0.24	0.77	0.84	0.52	0.83	0.79	0.45	0.24	0.67	0.89	0.64	0.72	0.71	0.63	0.19
		0.82	0.66	0.71	0.39	0.86	0.76	0.38	0.23	0.73	0.67	0.44	0.83	0.73	0.39	0.23	0.68	0.76	0.54	0.76	0.70	0.56	0.19
UM	0.5	3.03	0.91	0.63	0.44	0.70	0.82	0.76	0.39	0.78	0.59	0.43	0.55	0.68	0.60	0.37	0.88	0.69	0.40	0.60	0.54	0.35	0.30
		1.35	0.87	0.59	0.38	0.66	0.64	0.38	0.34	0.79	0.44	0.36	0.57	0.59	0.26	0.31	0.83	0.40	0.32	0.57	0.46	0.13	0.21
		0.82	0.80	0.47	0.34	0.64	0.58	0.25	0.28	0.79	0.30	0.31	0.60	0.52	0.15	0.26	0.77	0.20	0.26	0.56	0.39	0.11	0.19
	0.25	3.03	0.82	0.81	0.32	0.78	0.81	0.18	0.28	0.80	0.81	0.32	0.66	0.76	0.28	0.27	0.81	0.86	0.49	0.51	0.69	0.52	0.22
		1.35	0.78	0.68	0.23	0.74	0.75	0.17	0.23	0.79	0.78	0.29	0.70	0.70	0.29	0.23	0.80	0.79	0.38	0.53	0.60	0.41	0.20
		0.82	0.73	0.70	0.19	0.71	0.72	0.22	0.23	0.75	0.75	0.23	0.69	0.62	0.24	0.22	0.78	0.71	0.36	0.59	0.62	0.35	0.20
	0.1	3.03	0.82	0.81	0.21	0.84	0.84	0.51	0.24	0.75	0.83	0.31	0.81	0.82	0.49	0.22	0.75	0.87	0.50	0.58	0.79	0.67	0.19
		1.35	**0.81**	**0.81**	**0.16**	**0.83**	**0.80**	**0.43**	**0.20**	0.75	0.81	0.27	0.82	0.79	0.42	0.20	0.76	0.74	0.47	0.69	0.73	0.59	0.21
		0.82	0.70	0.77	0.12	0.81	0.79	0.34	0.19	0.74	0.80	0.21	0.81	0.77	0.42	0.21	0.75	0.71	0.49	0.73	0.72	0.52	0.21
UE	0.5	3.03	0.75	0.83	0.46	0.67	0.86	0.76	0.40	0.81	0.69	0.48	0.57	0.62	0.58	0.37	0.71	0.53	0.48	0.77	0.53	0.35	0.28
		1.35	0.75	0.54	0.41	0.67	0.63	0.48	0.35	0.80	0.51	0.41	0.57	0.59	0.28	0.31	0.60	0.49	0.39	0.62	0.50	0.19	0.20
		0.82	0.74	0.50	0.36	0.65	0.56	0.24	0.28	0.76	0.48	0.37	0.57	0.57	0.019	0.27	**0.64**	**0.36**	**0.33**	**0.57**	**0.44**	**0.10**	**0.20**
	0.25	3.03	0.76	0.84	0.27	0.79	0.84	0.17	0.27	0.74	0.87	0.37	0.67	0.79	0.23	0.26	**0.72**	**0.91**	**0.47**	**0.61**	**0.59**	**0.53**	**0.26**
		1.35	0.74	0.78	0.26	0.75	0.77	0.28	0.27	0.71	0.79	0.34	0.71	0.74	0.29	0.26	0.66	0.78	.43	0.56	0.63	0.50	0.20
		0.82	0.75	0.64	0.21	0.71	0.74	0.25	0.24	0.70	0.73	0.29	0.72	0.67	0.26	0.20	0.63	0.76	0.38	0.58	0.62	0.44	0.19
	0.1	3.03	0.80	0.82	0.21	0.84	0.82	0.55	0.26	0.77	0.87	0.26	0.81	0.82	0.56	0.24	0.66	0.94	0.64	0.67	0.79	0.75	0.21
		1.35	0.76	0.82	0.17	0.82	0.79	0.46	0.21	0.77	0.83	0.28	0.81	0.79	0.48	0.21	0.64	0.84	0.53	0.72	0.76	0.65	0.20
		0.82	0.59	0.78	0.13	0.82	0.78	0.37	0.22	0.66	0.80	0.25	0.82	0.77	0.43	0.19	0.66	0.73	0.54	0.74	0.74	0.55	0.22

Table A.4: R^2 values of variable selection methods for simulation case. F: PLS-Full; V: PLS-VIP; B: PLS-BETA; R: RIVAL; U: UVE-PLS; S: SR; C: CARS-PLS; G: GA-PLS.

			$CBP = 0.5$								$CBP = 0.7$								$CBP = 0.9$							
SRC	PR	SNR	F	V	B	R	U	S	C	G	F	V	B	R	U	S	C	G	F	V	B	R	U	S	C	G
EM	0.5	3.03	0.90	0.88	0.90	0.89	0.90	0.90	0.90	0.90	0.90	0.89	0.90	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.89	0.90	0.90	0.90	0.90
		1.35	0.63	0.62	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.62	0.63	0.62	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63
		0.82	0.37	0.37	0.37	0.37	0.38	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.38	0.36	0.37	0.37	0.37	0.38	0.37	0.38	0.38	0.37	0.37	0.38
	0.25	3.03	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
		1.35	0.61	0.64	0.63	0.63	0.63	0.63	0.62	0.62	0.62	0.64	0.63	0.63	0.63	0.62	0.62	0.62	**0.63**	**0.64**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**
		0.82	0.35	0.39	0.37	0.37	0.38	0.36	0.35	0.36	0.35	0.38	0.37	0.38	0.38	0.37	0.36	0.37	0.37	0.38	0.37	0.39	.38	0.38	0.37	0.37
	0.1	3.03	0.88	0.90	0.90	0.87	0.90	0.90	0.90	0.89	0.89	0.90	0.90	0.88	0.90	0.90	0.90	0.89	0.89	0.90	0.90	0.90	0.90	0.89	0.90	0.89
		1.35	0.57	0.63	0.63	0.61	0.63	0.63	0.61	0.60	0.58	0.63	0.63	0.61	0.63	0.62	0.61	0.60	0.62	0.63	0.64	0.63	0.63	0.63	0.62	0.62
		0.82	0.26	0.37	0.34	0.34	0.38	0.35	0.32	0.33	**0.29**	**0.38**	**0.35**	**0.36**	**0.38**	**0.36**	**0.34**	**0.34**	0.36	0.37	0.38	0.38	0.37	0.37	0.36	0.36
EE	0.5	3.03	0.90	0.89	0.90	0.89	0.90	0.90	0.90	0.89	0.90	0.90	0.90	0.82	0.90	0.90	0.90	0.89	0.90	0.89	0.90	0.90	0.90	0.90	0.90	0.90
		1.35	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.62	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.62	0.63	0.63	0.63	0.63	0.63	0.63
		0.82	0.37	0.38	0.37	0.38	0.38	0.37	0.37	0.37	0.37	0.38	0.37	0.38	0.38	0.37	0.37	0.37	0.37	0.37	0.37	0.38	0.38	0.37	0.37	0.38
	0.25	3.03	**0.89**	**0.90**	**0.90**	**0.89**	**0.90**	**0.90**	**0.90**	**0.89**	0.90	0.90	0.90	0.70	0.90	0.90	0.90	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
		1.35	0.62	0.64	0.64	0.63	0.64	0.64	0.62	0.62	0.62	0.64	0.64	0.64	0.64	0.63	0.63	0.63	0.64	0.64	0.64	0.63	0.64	.63	0.63	0.63
		0.82	0.36	0.39	0.37	0.38	0.39	0.37	0.36	0.36	0.36	0.39	0.37	0.38	0.39	0.37	0.37	0.37	0.38	0.38	0.38	0.39	0.38	0.38	0.38	0.38
	0.1	3.03	0.88	0.90	0.90	0.85	0.90	0.90	0.90	0.89	0.89	0.90	0.90	0.86	0.90	0.90	0.90	0.89	0.89	0.90	0.90	0.89	0.90	0.90	0.90	0.89
		1.35	0.56	0.63	0.63	0.61	0.63	0.63	0.61	0.60	0.58	0.63	0.63	0.60	0.63	0.62	0.61	0.60	0.62	0.63	0.63	0.63	0.63	0.62	0.62	0.61
		0.82	0.26	0.36	0.35	0.34	0.38	0.35	0.32	0.33	0.29	0.37	0.35	0.36	0.38	0.35	0.33	0.33	0.37	0.37	0.38	0.38	0.37	0.37	0.36	0.35
UM	0.5	3.03	0.90	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
		1.35	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63
		0.82	0.37	0.38	0.37	0.38	0.38	0.37	0.37	0.37	0.37	0.38	0.37	0.38	0.38	0.37	0.37	0.37	0.37	0.38	0.37	0.38	0.38	0.38	0.37	0.38
	0.25	3.03	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
		1.35	0.61	0.63	0.63	0.63	0.63	0.63	0.62	0.62	0.62	0.64	0.63	0.63	0.63	0.63	0.63	0.62	0.63	0.64	0.64	0.64	0.63	0.63	0.63	0.63
		0.82	0.35	0.38	0.37	0.37	0.38	0.37	0.36	0.36	0.36	0.38	0.37	0.38	0.38	0.37	0.37	0.37	0.38	0.38	0.39	0.39	0.38	0.38	0.38	0.37
	0.1	3.03	0.88	0.90	0.90	0.89	0.90	0.90	0.89	0.89	0.89	0.90	0.90	0.90	0.90	0.90	0.89	0.89	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.89
		1.35	**0.57**	**0.64**	**0.62**	**0.62**	**0.63**	**0.63**	**0.61**	**0.61**	0.59	0.63	0.63	0.63	0.63	0.63	0.61	0.61	0.62	0.63	0.64	0.63	0.63	0.63	0.62	0.62
		0.82	0.26	0.37	0.36	0.35	0.38	0.37	0.34	0.34	0.29	0.37	0.37	0.37	0.38	0.37	0.34	0.34	0.36	0.38	0.39	0.38	0.38	0.38	0.36	0.36
UE	0.5	3.03	0.90	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
		1.35	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.64	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.64	.63	0.64	0.63	0.63	0.63	0.63
		0.82	0.37	0.38	0.37	0.38	0.38	0.37	0.37	0.37	0.37	0.39	0.38	0.38	0.38	0.38	0.37	0.38	**0.38**	**0.38**	**0.38**	**0.39**	**0.38**	**0.38**	**0.38**	**0.38**
	0.25	3.03	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	**0.90**	**0.90**	**0.90**	**0.90**	**0.90**	**0.90**	**0.90**	**0.90**
		1.35	0.62	0.64	0.63	0.63	0.64	0.63	0.63	0.63	0.63	0.64	0.64	0.64	0.64	0.63	0.63	0.63	0.64	0.64	0.64	0.64	0.64	0.63	0.63	0.63
		0.82	0.35	0.39	0.37	0.38	0.39	0.38	0.37	0.37	0.37	0.39	0.38	0.39	0.39	0.38	0.37	0.37	0.38	0.39	0.39	0.39	0.39	0.38	0.38	0.38
	0.1	3.03	0.88	0.90	0.90	0.89	0.90	0.90	0.89	0.89	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.89	0.88	0.90	0.90	0.90	0.90	0.90	0.90	0.89
		1.35	0.57	0.63	0.63	0.62	0.63	0.63	0.61	0.60	0.59	0.63	0.63	0.62	0.63	0.63	0.62	0.61	0.61	0.63	0.64	0.63	0.63	0.63	0.63	0.62
		0.82	0.26	0.36	0.36	0.35	0.38	0.37	0.33	0.33	0.30	0.37	0.37	0.37	0.38	0.37	0.35	0.34	0.36	0.37	0.39	0.38	0.38	0.37	0.36	0.35

Table A.5: Percentage improvement of average MAPE from different methods compared to full model for simulation case. Only full PLS models are given in average MAPE.

			$CBP = 0.5$								$CBP = 0.7$								$CBP = 0.9$							
SRC	PR	SNR	F	V	B	R	U	S	C	G	F	V	B	R	U	S	C	G	F	V	B	R	U	S	C	G
EM	0.5	3.03	3.22	-6.2	0.85	-0.9	0.57	0.78	0.13	-0.5	3.23	-2.8	0.59	-1.8	0.30	0.54	0.04	-0.5	3.23	0.04	0.14	-1.1	0.09	-0.6	-0.2	-0.5
		1.35	6.15	-1.1	0.54	0.13	0.66	0.51	-0.0	0.10	6.16	-0.1	0.08	-0.1	0.40	-0.4	-0.2	-0.4	6.19	0.53	-0.0	0.26	0.33	-0.1	-0.2	0.24
		0.82	89.08	0.12	-0.0	0.42	0.63	-0.2	-0.1	-0.1	8.08	0.46	-0.2	0.41	0.53	-0.4	-0.2	0.04	8.11	0.57	0.05	0.88	0.47	0.34	0.13	0.56
	0.25	3.03	3.24	2.84	2.86	1.04	2.61	2.63	1.47	-1.3	3.17	1.88	1.96	0.37	1.60	1.74	0.82	0.13	3.15	0.72	0.87	-0.3	0.32	-0.5	-0.2	-0.5
		1.35	6.13	2.89	2.60	1.29	2.57	2.40	0.67	1.24	6.09	2.11	1.65	1.21	1.79	0.54	0.20	0.51	**6.02**	**0.99**	**0.45**	**0.21**	**0.36**	**-0.2**	**-0.1**	**-0.3**
		0.82	8.05	2.86	1.23	1.94	2.49	1.24	0.34	1.07	8.03	2.29	0.85	1.79	1.96	0.90	0.67	0.90	7.91	0.80	0.25	1.09	0.62	0.39	0.11	0.14
	0.1	3.03	3.50	7.91	8.14	-2.2	7.72	7.54	6.39	3.22	3.39	5.50	5.95	-1.3	5.42	5.44	4.19	2.06	3.21	2.15	2.60	0.69	1.89	0.51	1.55	-0.4
		1.35	6.68	8.14	7.80	5.87	7.91	7.45	5.58	4.26	6.49	6.30	5.63	3.19	6.17	4.43	3.67	2.99	6.12	0.88	1.55	0.99	0.74	0.12	-0.1	-1.1
		0.82	8.73	7.26	5.59	5.75	8.02	5.86	4.24	4.35	**8.55**	**6.97**	**4.46**	**5.47**	**6.61**	**5.26**	**3.64**	**3.66**	8.09	0.92	1.80	1.53	1.10	1.12	-0.0	-0.3
EE	0.5	3.03	3.15	-1.2	1.06	-2.3	0.79	0.94	0.08	-1.9	3.17	0.63	0.85	-27	0.40	0.80	0.04	-1.5	3.23	-1.5	0.40	-0.2	0.19	-0.1	-0.0	-0.7
		1.35	5.99	0.24	0.88	0.79	0.84	0.84	0.01	-0.4	6.03	0.65	0.41	0.35	0.49	0.19	0.03	-0.1	6.15	-0.8	0.02	0.56	0.33	-0.3	-0.2	0.02
		0.82	7.83	0.51	0.27	0.80	0.78	0.15	-0.1	-0.0	7.91	0.66	-0.0	0.51	0.63	-0.2	-0.2	-0.1	8.04	-0.6	-0.0	0.81	0.51	0.15	0.06	0.45
	0.25	3.03	**3.32**	**2.64**	**2.67**	**-0.6**	**2.43**	**2.47**	**1.57**	**-2.1**	3.31	1.98	2.12	-65	1.74	1.96	0.88	-0.5	3.31	0.57	0.70	-0.9	0.02	-0.2	-0.1	-0.8
		1.35	6.31	2.61	2.46	1.03	2.39	2.37	0.60	0.56	6.29	1.93	1.74	1.56	1.72	0.93	0.40	0.39	6.22	0.16	0.24	-0.3	-0.1	-0.9	-0.6	-0.8
		0.82	8.34	2.52	1.30	1.62	2.36	1.12	0.53	0.76	8.29	2.09	1.03	1.73	1.98	0.77	0.44	0.76	8.13	0.14	0.22	0.57	0.17	-0.2	-0.2	-0.4
	0.1	3.03	3.49	7.87	8.07	-8.7	7.83	7.46	6.61	2.93	3.44	4.82	5.41	-8.9	5.02	4.84	3.82	1.26	3.43	3.23	3.75	1.34	3.09	2.36	2.92	1.13
		1.35	6.67	7.94	7.77	5.83	8.11	7.56	5.18	4.41	6.58	5.60	5.59	2.19	5.81	3.94	3.17	2.49	6.31	0.97	1.29	0.47	0.59	-0.4	-0.1	-1.5
		0.82	8.80	7.14	5.76	5.76	8.12	5.95	4.16	4.58	8.70	5.61	4.26	4.74	6.42	4.71	3.15	3.21	98.19	0.47	1.18	0.79	0.34	-0.1	-0.6	-1.3
UM	0.5	3.03	3.20	-2.3	0.43	0.54	0.73	0.41	-0.1	-0.4	3.17	-0.1	0.03	0.34	0.25	0.08	-0.1	-0.3	3.18	0.28	-0.1	0.33	0.04	-0.3	-0.2	-0.4
		1.35	6.17	0.58	0.32	0.72	0.77	0.46	0.18	0.23	6.15	0.81	0.09	0.64	0.56	0.17	-0.0	0.15	6.15	0.46	-0.0	0.59	0.15	-0.2	-0.2	0.00
		0.82	8.06	1.01	0.30	0.83	0.86	0.41	0.09	0.32	8.09	1.00	0.10	0.84	0.81	0.20	-0.0	0.41	8.07	0.52	0.06	0.83	0.38	0.31	0.18	0.35
	0.25	3.03	3.31	1.68	1.86	1.87	2.39	2.12	0.69	0.35	3.20	1.54	0.89	1.00	1.35	0.52	0.07	-0.3	3.14	0.80	0.91	0.76	0.50	0.07	0.28	-0.3
		1.35	6.26	2.77	1.61	2.06	2.53	1.89	1.10	1.35	6.16	1.91	0.94	1.40	1.67	0.92	0.53	0.23	6.04	0.34	0.61	0.53	0.26	-0.5	-0.1	-0.6
		0.82	8.19	2.93	1.68	2.16	2.53	1.84	1.28	1.37	8.18	2.24	1.26	1.83	2.03	1.23	0.92	0.95	7.95	0.38	0.81	0.79	0.34	0.13	0.10	-0.1
	0.1	3.03	3.49	7.74	6.85	5.60	7.60	6.97	5.43	3.86	3.37	4.58	4.43	3.34	4.64	3.78	2.26	1.06	3.33	5.42	5.79	5.29	5.15	4.68	4.77	3.22
		1.35	**6.74**	**8.45**	**6.87**	**6.03**	**8.15**	**7.15**	**5.46**	**4.50**	6.51	5.46	4.77	4.43	5.54	4.56	3.16	2.41	6.17	1.92	2.41	1.85	1.73	1.16	0.69	-0.5
		0.82	8.72	7.50	6.92	6.27	8.16	7.25	4.95	4.93	8.51	5.78	5.40	5.27	6.44	5.34	3.64	3.59	8.10	1.29	2.21	1.65	1.31	1.29	0.12	-0.4
UE	0.5	3.03	3.25	-1.3	0.67	0.77	0.85	0.69	-0.0	-0.9	3.23	0.37	0.38	0.47	0.37	0.15	-0.2	-0.5	3.24	-0.1	-0.0	0.27	0.00	-0.3	-0.1	-0.5
		1.35	6.11	0.72	0.36	0.82	0.81	0.41	-0.0	-0.2	6.09	0.69	0.09	0.52	0.42	-0.1	-0.2	-0.3	6.16	0.47	0.06	0.48	0.17	0.10	-0.1	0.03
		0.82	7.94	1.06	0.34	0.88	0.91	0.37	0.21	0.18	7.92	0.84	0.24	0.73	0.59	0.31	-0.1	0.11	**8.00**	**0.54**	**0.10**	**0.79**	**0.35**	**0.24**	**0.03**	**0.24**
	0.25	3.03	3.29	2.27	2.00	1.67	2.29	2.04	0.39	-0.8	3.31	1.87	1.68	1.42	1.77	1.45	0.42	-0.3	**3.35**	**1.83**	**1.79**	**1.65**	**0.89**	**0.94**	**1.21**	**0.63**
		1.35	6.22	2.62	1.67	1.84	2.32	1.74	0.99	0.83	6.23	1.47	1.17	1.16	1.39	0.80	0.38	-0.1	6.23	0.38	0.53	0.47	0.28	-0.4	-0.2	-0.6
		0.82	8.16	2.64	1.45	1.95	2.40	1.63	1.02	1.07	8.15	1.65	1.03	1.44	1.72	0.85	0.34	0.32	8.10	0.10	0.50	0.45	0.18	-0.3	-0.2	-0.6
	0.1	3.03	3.50	7.65	6.95	4.20	7.61	7.00	5.48	2.90	3.44	5.35	5.16	4.03	5.40	4.57	3.72	1.98	3.67	7.91	9.85	9.37	8.64	8.71	9.08	7.24
		1.35	6.69	7.67	6.95	5.71	7.78	6.81	5.12	3.93	6.49	4.76	4.59	3.75	4.99	4.09	2.75	1.34	6.40	2.32	3.04	2.57	2.40	1.85	1.58	0.61
		0.82	8.75	6.93	6.88	6.07	8.10	7.19	4.71	4.57	8.59	4.70	5.12	4.53	5.59	4.67	3.12	2.42	8.24	1.10	2.09	1.57	1.26	0.63	0.11	-0.7

Appendix B: Industrial Polyester Case Results

The average and standard deviation of numbers of variables selected by different variable selection methods are summarized in Table A.6, for both acidity number and viscosity. Acidity number seems to require slightly larger model, as compared to viscosity models.

Table A.6 Average and standard deviation of numbers of variables selected for acidity number and viscosity models

	Cases	
Methods	Acidity Number	Viscosity
PLS-VIP	15.22 ± 0.50	10.98 ± 0.20
PLS-BETA	14.15 ± 1.15	10.41 ± 1.61
Lasso	16.29 ± 2.35	15.94 ± 4.29
UVE-PLS	23.33 ± 2.48	19.41 ± 3.81
SR	8.26 ± 1.13	8.61 ± 1.35
CARS-PLS	8.37 ± 2.27	9.29 ± 3.90
GA-PLS	7.68 ± 1.81	6.90 ± 1.79