Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 78 (2005) 103–112

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Performance of some variable selection methods when multicollinearity is present

Il-Gyo Chong, Chi-Hyuck Jun*

Department of Industrial Engineering, Pohang University of Science and Technology, San 31 Hyoja-dong, Pohang 790-784, Republic of Korea

Received 22 April 2004; accepted 22 December 2004
Available online 7 March 2005

Abstract

Variable selection is one of the important practical issues for many scientific engineers. Although the PLS (partial least squares) regression combined with the VIP (variable importance in the projection) scores is often used when the multicollinearity is present among variables, there are few guidelines about its uses as well as its performance. The purpose of this paper is to explore the nature of the VIP method and to compare with other methods through computer simulation experiments. We design 108 experiments where observations are generated from true models considering four factors–the proportion of the number of relevant predictors, the magnitude of correlations between predictors, the structure of regression coefficients, and the magnitude of signal to noise. Confusion matrix is adopted to evaluate the performance of PLS, the Lasso, and stepwise method. We also discuss the proper cutoff value of the VIP method to increase its performance. Some practical hints for the use of the VIP method are given as simulation results.
© 2005 Elsevier B.V. All rights reserved.

Keywords: Variable selection; VIP (Variable Importance in the Projection) scores; Partial least squares regression; The lasso; Stepwise regression; Multicollinearity

1. Introduction

The quality of a final product in a process industry is believed to be determined by a lot of process variables. Process engineers are often interested in finding vital few process variables that would be most influential on the quality of the product. With only several variables in hand, their control problem for the quality improvement would become much easier. Although stepwise regression methods are often used for this purpose due to their simplicity, there are several reasons why process engineers are often not satisfied with the results. One of them is its poor performance when the multicollinearity exists among variables. Under this situation, the VIP (Variable Importance in the Projection) scores obtained by the partial least squares (PLS) regression, has been paid an increasing attention these days as an importance measure of each explanatory variable or predictor [1]. However, the performance and the use of the VIP scores are not well discovered.

The objective of this study is to investigate the performance of the VIP scores for selecting the relevant process variables which "really" have an effect on the response or have nonzero coefficients. For this purpose, we used computer simulation experiments where some true models are assumed and data sets are generated so as to mimic the typical manufacturing process which consists of consecutive unit processes. We compare the performance of VIP scores under PLS (called PLS-VIP method) with the PLS regression (called PLS-BETA method), the Lasso regression [2] and the stepwise regression [3]. We also aim to discuss the proper cutoff value of the PLS-VIP method.

The rest of the paper is organized as follows. A brief review of variable selection methods using PLS regression,

* Corresponding author. Tel.: +82 54 279 2197; fax: +82 54 279 2870.
E-mail address: chjun@postech.ac.kr (C.-H. Jun).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2004.12.011

the Lasso regression and the stepwise regression is given in Section 2. Section 3 describes the simulation design and performance measure using confusion matrix. The simulation results and the discussion are given in Section 4. Finally, Section 5 concludes the paper with a summary.

2. Variable selection methods

2.1. Partial least squares regression

In case of single response y and p predictors, PLS regression model with h ($h \le p$) latent variables can be expressed as follows [4,5].

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^t + \boldsymbol{E} \tag{1a}$$

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{b} + \boldsymbol{f} \tag{1b}$$

In Eq. (1a,b), $\boldsymbol{X}$ ($n \times p$), $\boldsymbol{T}$ ($n \times h$), $\boldsymbol{P}$ ($p \times h$), $\boldsymbol{y}$ ($n \times 1$), and $\boldsymbol{b}$ ($h \times 1$) are respectively used for predictors, $\boldsymbol{X}$ scores, $\boldsymbol{X}$ loadings, a response, and regression coefficients of $\boldsymbol{T}$. The k-th element of column vector $\boldsymbol{b}$ explains the relation between $\boldsymbol{y}$ and $\boldsymbol{t}_k$, the k-th column vector of $\boldsymbol{T}$. Meanwhile, $\boldsymbol{E}$ ($n \times p$) and $\boldsymbol{f}$ ($n \times 1$) stand for random errors of $\boldsymbol{X}$ and $\boldsymbol{y}$, respectively. Generally, by using the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm, a weight matrix $\boldsymbol{W}$ ($p \times h$) is obtained to make $\|\boldsymbol{f}\|$ (Euclidian norm) as small as possible and, at the same time, to derive a useful relation between $\boldsymbol{X}$ and $\boldsymbol{y}$. Here, unlike many other applications, $n > p$ is assumed due to the easiness of data availability in process industries.

NIPALS algorithm: in case of single $\boldsymbol{y}$.

Assume that the $n \times p$ matrix $\mathbf{X}$ and the column vector $\boldsymbol{y}$ have been standardized to have mean 0 and unit variance. In the following, $\boldsymbol{t}_k$, $\boldsymbol{p}_k$, and $\boldsymbol{w}_k$ respectively stand for the k-th column vector of $\boldsymbol{T}$, $\boldsymbol{P}$, and $\boldsymbol{W}$. The k-th latent variable is obtained iteratively as follows (k=1,2,...,h). Thus, model parameters in Eq. (1a,b) are determined accordingly.

Step 1 $\boldsymbol{y}_{(k)} \leftarrow \boldsymbol{y}_{(k-1)} - \boldsymbol{b}_{k-1}\boldsymbol{t}_{k-1}$; $\boldsymbol{y}_{(1)} \leftarrow \boldsymbol{y}$ and $\mathbf{X}_{(k)} \leftarrow \mathbf{X}_{(k-1)} - \boldsymbol{t}_{k-1}\boldsymbol{p}_{k-1}^t$; $\mathbf{X}_{(1)} \leftarrow \mathbf{X}$
Step 2 $\boldsymbol{w}_k^t = \boldsymbol{y}_{(k)}^t \mathbf{X}_{(k)} / \boldsymbol{y}_{(k)}^t \boldsymbol{y}_{(k)}$
Step 3 $\boldsymbol{w}_k \leftarrow \boldsymbol{w}_k / \| \boldsymbol{w}_k \|$
Step 4 $\boldsymbol{t}_k = \mathbf{X}_{(k)} \boldsymbol{w}_k / \boldsymbol{w}_k^t \boldsymbol{w}_k$
Step 5 $\boldsymbol{p}_k^t = \boldsymbol{t}_k^t \mathbf{X}_{(k)} / \boldsymbol{t}_k^t \boldsymbol{t}_k$
Step 6 $\boldsymbol{t}_k \leftarrow \boldsymbol{t}_k \bullet \| \boldsymbol{p}_k \|$
Step 7 $\boldsymbol{w}_k \leftarrow \boldsymbol{w}_k \bullet \| \boldsymbol{p}_k \|$
Step 8 $\boldsymbol{p}_k \leftarrow \boldsymbol{p}_k / \| \boldsymbol{p}_k \|$
Step 9 $b_k = \boldsymbol{y}_{(k)}^t \boldsymbol{t}_k / \boldsymbol{t}_k^t \boldsymbol{t}_k$

Here, two variable selection methods using PLS regression will be considered. The first one is to use VIP scores (PLS-VIP method) and the other is to use regression coefficients estimated by PLS regression (called PLS-BETA method).

2.1.1. PLS-VIP method

The VIP score of a predictor, first published in [6], is a summary of the importance for the projections to find h latent variables. The VIP score for the j-th variable can be calculated by Eq. (2). On the other hand, since the average of squared VIP scores equals 1, 'greater than one rule' is generally used as a criterion for variable selection.

$$\mathrm{VIP}_j = \sqrt{p \sum_{k=1}^{h} \left(SS(b_k \boldsymbol{t}_k) \left(\boldsymbol{w}_{jk} / \| \boldsymbol{w}_k \| \right)^2 \right) \Big/ \sum_{k=1}^{h} SS(b_k \boldsymbol{t}_k)},$$

$$\text{where} \quad SS(b_k \boldsymbol{t}_k) = b_k^2 \boldsymbol{t}_k^t \boldsymbol{t}_k \tag{2}$$

2.1.2. PLS-BETA method

The relation of $\boldsymbol{T}$ and $\boldsymbol{W}$ obtained by the NIPALS algorithm is given by Eq. (3) [7].

$$\boldsymbol{T} = \mathbf{X}\boldsymbol{W}^* \qquad \text{where} \qquad \boldsymbol{W}^* = (\boldsymbol{P}^t \boldsymbol{W})^{-1} \tag{3}$$

From this, the predicted values can be directly calculated by Eq. (4). The relevant predictors could be selected according to the magnitude of the absolute values of regression coefficients.

$$\hat{\boldsymbol{y}} = \boldsymbol{T}(\boldsymbol{T}^t \boldsymbol{T})^{-1} \boldsymbol{T}^t \boldsymbol{y} = \mathbf{X} \boldsymbol{b}_{\mathrm{pls}} \tag{4}$$

$$\text{where} \quad \boldsymbol{b}_{\mathrm{pls}} = \boldsymbol{W}(\boldsymbol{P}^t \boldsymbol{W})^{-1} (\boldsymbol{T}^t \boldsymbol{T})^{-1} \boldsymbol{T}^t \boldsymbol{y}$$

2.2. Least absolute shrinkage and selection operator (Lasso)

The Lasso [2] minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant s. In view of shrinking the regression coefficients by imposing a penalty on their size, the Lasso is similar in spirit to Ridge regression. If the data are standardized to have mean 0, the Lasso estimate is defined by Eq. (5). Here a tuning parameter, $s \ge 0$, can be determined by the cross-validation. Because of the nature of the constraint it tends to produce some coefficients as zero and it may improve the overall prediction accuracy by sacrificing a little bias to reduce the variance of the predicted values.

$$\hat{\boldsymbol{\beta}}^{\mathrm{lasso}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\text{subject to} \quad \sum_{j-1}^{p} |\beta_j| \le s \tag{5}$$

Although the solution to Eq. (5) can be obtained by the standard quadratic programming with linear inequality

constraints, the use of Least Angle Regression (LARS) algorithm reduces the computation burden [8].

LARS algorithm for the Lasso estimate.

Assume that the predictors have been standardized to have a mean 0 and unit length, and that the response has a mean 0. In the k-th iteration, the algorithm is roughly described as follows.

Step 1 Update the active set. Calculate the absolute current correlations.

$$\hat{c}_{kj} = \boldsymbol{x}_j^t(\boldsymbol{y} - \hat{\boldsymbol{y}}_{k-1}); \quad \hat{\boldsymbol{y}}_0 = 0 \quad \text{and} \quad \hat{C}_k = \max_j \left\{|\hat{c}_{kj}|\right\}$$

Update the active set $A(k)$.

$$A(k) = A(k-1) + \left\{\hat{j}\right\}; \quad A(0) = \phi$$
$$\text{and} \quad \hat{j} = \underset{j \notin A(k-1)}{\arg\max} \left\{|\hat{c}_{kj}|\right\}$$

Step 2 Determine the least angle direction ($\boldsymbol{u}_k$). Define $X_k = (\cdots s_j \boldsymbol{x}_j \cdots)_{j \in A(k)}$ where $s_j = \text{sign}\{\hat{c}_{kj}\}$ and $\boldsymbol{w}_k = A_k(X_k^t X_k)^{-1}\mathbf{1}_k$ where $A_k = (\mathbf{1}_k^t(X_k^t X_k)^{-1}\mathbf{1}_k)^{-0.5}$ (Here, $\mathbf{1}_k$ is a vector of 1's of length equaling $|A|$, the size of A.) Calculate the least angle direction.

$$\boldsymbol{u}_k = \mathrm{X}_k \boldsymbol{w}_k$$

Step 3 Calculate the step size. Define $a_{kj} = \boldsymbol{x}_j^t \boldsymbol{u}_k$ for $j \notin A(k)$. Determine the step size.

If $|A|$ equals the number of predictors,

$$\hat{\gamma}_k = \hat{C}_k / A_k \text{ and the algorithm is terminated}$$

else

$$\hat{\gamma}_k = \min{}^{+}_{j \in A(k)} \left\{ \left(\hat{C}_k - \hat{c}_{kj}\right) / \left(A_k - a_{kj}\right), \left(\hat{C}_k + \hat{c}_{kj}\right) / \left(A_k + a_{kj}\right) \right\}$$

(Here, "$\min^{+}$" indicates that the minimum is taken over only positive components within each choice of j.)

Step 4 Predict the response.

$$\tilde{\gamma}_k = \min_{\gamma_j > 0,\, j \in A(k)} \left\{\gamma_j\right\} \quad \text{where} \quad \gamma_j = -\hat{\beta}_j / \left(s_j w_{kj}\right); \quad \tilde{\gamma}_1 = \infty$$

If $\tilde{\gamma}_k < \hat{\gamma}_k$

$$\hat{\boldsymbol{y}}_k = \hat{\boldsymbol{y}}_{k-1} + \tilde{\gamma}_k \boldsymbol{u}_k$$

$$\text{if} \quad j \in A, \hat{\beta}_j \leftarrow \hat{\beta}_j + \tilde{\gamma}_k w_{kj} s_j. \quad \text{Otherwise,} \quad \hat{\beta}_j = 0$$

$$A(k+1) = A(k) - \left\{\tilde{j}\right\} \qquad \text{where} \qquad \tilde{j} = \underset{\gamma_j > 0,\, j \in A(k)}{\arg\min} \left\{\gamma_j\right\}$$

$$\hat{c}_{k+1\,j} = \boldsymbol{x}_j^t(\boldsymbol{y} - \hat{\boldsymbol{y}}_k) \quad \text{and} \quad \hat{C}_{k+1} = \max_j \left\{|\hat{c}_{k+1\,j}|\right\}$$

Go to Step 2.

else

$$\hat{\boldsymbol{y}}_k = \hat{\boldsymbol{y}}_{k-1} + \hat{\gamma}_k \boldsymbol{u}_k$$

$$\text{if} \quad j \in A, \hat{\beta}_j \leftarrow \hat{\beta}_j + \hat{\gamma}_k w_{kj} s_j. \quad \text{Otherwise,} \quad \hat{\beta}_j = 0$$

Go to Step 1.

2.3. Stepwise regression

Stepwise regression is a standard procedure for variable selection, which is based on the procedure of sequentially introducing the predictors into the model one at a time. The stepwise regression is classified into three methods: forward selection, backward elimination and stepwise method. The forward selection adds predictors to the model one at a time. In contrast to the forward selection, the backward elimination begins with the full model and successively eliminates one predictor at a time. An advantage of a forward selection for a large number of correlated variables, as opposed to backward elimination, is that the $\mathbf{X}^t\mathbf{X}$ matrix does not need to be inverted. Meanwhile, the stepwise method starts as the forward selection, but at each stage the possibility of deleting a predictor, as backward elimination, is considered.

In these methods, the number of predictors retained in the final model is determined by the levels of significance assumed for inclusion and exclusion of predictors from the model. In view of the "rule-of-thumb", the significance levels of 0.15 give equation with low Mallows-C_p [9]. On the other hand, the three methods are expected to perform similarly, so in this study only the stepwise method will be considered for the comparison. We use the equal significance level as entry and removal criteria and select the proper one from 0.05, 0.1, 0.15, and 0.2 by cross-validation.

3. Experimental

3.1. Design of simulations

We generate datasets by assuming that true response follows a linear model having p predictors defined as Eq. (6).

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i,$$

$$\text{where} \qquad \varepsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right), \ (i = 1, 2, .., 500) \qquad (6)$$

Here, the data matrix $\mathbf{X}=(\mathrm{x}_{ij})$ is generated by assuming a special correlation structure described in Section 3.1.2. For convenience, we fix the number of relevant predictors as 10 and therefore the rest of predictors ($p-10$) are irrelevant to

the response over all cases. We design 108 (=3×3×4×3) different cases with four factors–the proportion of the number of relevant predictors among total predictors (3 levels), the magnitude of correlations between predictors (3 levels), the structure of regression coefficients (4 levels), and the magnitude of signal to noise (3 levels). In each case, 100 replications are made and performance measures are calculated. At each replication, a different dataset of 500 observations is generated from Eq. (6).

3.1.1. The proportion of relevant predictors among total predictors

This factor defined as Eq. (7) has three levels of 0.5, 0.25 and 0.1. These levels correspond to p=20, 40, and 100, respectively.

$$proportion = 10/p \tag{7}$$

3.1.2. Magnitude of correlations between predictors

At each replication, a new set of 500 rows of $\mathbf{X}$ is generated from multivariate normal distribution with zero mean vector and variance–covariance matrix of $\mathbf{\Gamma}$. The elements of matrix $\mathbf{\Gamma}$ are chosen as Eq. (8) since neighboring process variables (temperatures, e.g.) tend to be strongly correlated in a real manufacturing process which consists of consecutive unit processes.

$$\mathbf{\Gamma}_{ij} = \rho^{|i-j|}, \quad (i,j = 1,2,..,p) \tag{8}$$

Here, ρ is the magnitude of correlations between predictors which has three levels of 0.5, 0.7 and 0.9. Note that Eq. (8) gives a very specific pattern to the eigenvalues and eigenvectors of $\mathbf{\Gamma}$ [10]. Fig. 1 shows a comparison of eigenvalues of the covariance matrix from a real data encountered in a steel process with those from $\mathbf{\Gamma}$ in Eq. (8) for p=145 and ρ=0.9. In case of $\mathbf{\Gamma}$, a sequence of 21 eigenvalues is gradually decreasing from 18 to 1, explaining 85.5% of the variance with the remaining eigenvalues all small (but larger than 0.05). Meanwhile, in case of real data, a sequence of 32 eigenvalues is gradually decreasing from 30 to 1 explaining 81% of the variance with the remaining eigenvalues, all small (but larger than 0.0001 except two values near zero). Two patterns of eigenvalues are not exactly the same, but the decreasing patterns are similar to each other.

3.1.3. Structure of regression coefficients

The third factor is the structure of regression coefficients. Two types of equal and unequal coefficients are compared. It is intended to know performance of selection methods according to whether relevant predictors have similar effects on the response or not. Each type has two levels according to the location of relevant predictors; in the middle of the range and at the extremes.

In case of equal type, the coefficients of 10 relevant predictors are chosen as Eq. (9.a).

- In the middle of range

$$\beta_j = 1, \quad (j = p/2-4, p/2-3,..., p/2+5) \tag{9.a}$$

- At the extremes

$$\beta_j = 1, \quad (j = 1,2,..,5; p-4, p-3,..., p) \tag{9.b}$$

In case of unequal type, the coefficients of 10 relevant predictors are constructed as Eq. (10.a).

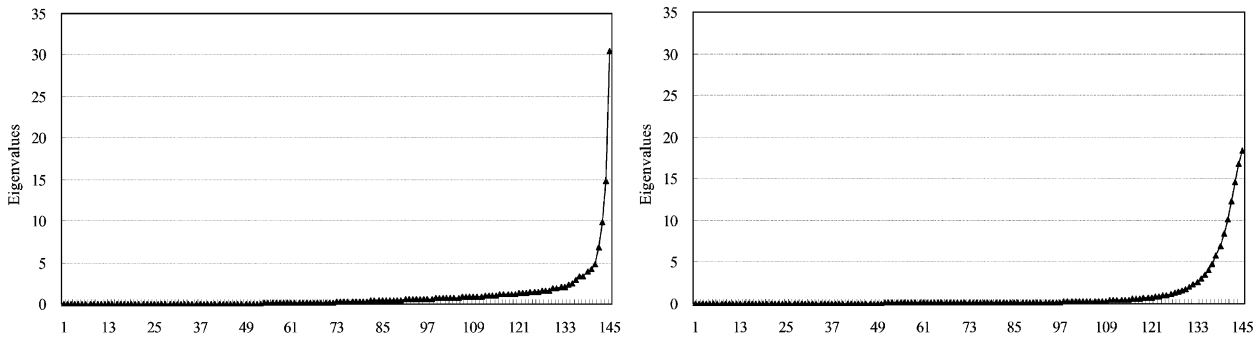
- In the middle of the range

$$\beta_j = (5.5 - |j - 0.5(p+1)|)^2, \quad (j = p/2-4, p/2-3,..., p/2+5) \tag{10.a}$$

- At the extremes

$$\beta_j = (|j - 0.5(p+1)| - 0.5(p-11))^2, \quad (j = 1,2,..,5; p-4, p-3,..., p) \tag{10.b}$$

All irrelevant predictors have zero coefficients in both types. For example, in case of unequal type, when the relevant coefficients are in the middle for p=20, β_j's are (0, 0, 0, 0, 0, 1, 4, 9, 16, 25, 25, 16, 9, 4, 1, 0, 0, 0, 0, 0,



(a) real data (b) artificial data (p = 145, ρ = 0.9)

Fig. 1. Comparison of eigenvalues between real data and artificial data.

Table 1
Confusion matrix and the descriptions of its entries

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

0). When the relevant coefficients are at the extremes, they are (25, 16, 9, 4, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 4, 9, 16, 25).

3.1.4. Magnitude of signal to noise

We are interested in knowing whether the performance of variable selection methods is affected by the model fitness. To investigate this, when generating y_i we select the standard deviation of error terms through Eq. (11) where k, the reciprocal of signal to noise ratio, has three levels of 0.33, 0.74, and 1.22. These levels are set so that R-square of the multiple linear regression with an intercept becomes 0.9, 0.65 and 0.4, respectively, when infinite observations are assumed. Some simple calculations using the formula for R-square give $k=((1\text{-}R^2)/R^2)^{1/2}$.

$$\sigma = k\sqrt{\mathrm{var}(X\beta)}, \quad (k = 0.33, 0.74, 1.22) \tag{11}$$

where var(·) is the sample variance.

3.2. Performance measure

For the evaluation of different selection methods we adopt the confusion matrix which contains information

Table 2
Mean G of each method along the cases

			k=0.33			k=0.74			k=1.22		
			V	L	S	V	L	S	V	L	S
Equal coefficients-middle	Prop.=0.5	ρ=0.5	0.964	0.755	0.966	0.967	0.746	0.951	0.958	0.749	0.875
		ρ=0.7	0.961	0.804	0.975	0.960	0.793	0.881	0.948	0.791	0.737
		ρ=0.9	0.987	0.821	0.876	0.964	0.755	0.660	0.950	0.715	0.550
	Prop.=0.25	ρ=0.5	0.998	0.863	0.973	0.993	0.848	0.959	0.992	0.859	0.880
		ρ=0.7	0.980	0.891	0.978	0.978	0.883	0.874	0.978	0.862	0.723
		ρ=0.9	0.921	0.907	0.857	0.923	0.831	0.650	0.921	0.779	0.556
	Prop.=0.1	ρ=0.5	0.990	0.930	0.979	0.989	0.924	0.971	0.988	0.924	0.872
		ρ=0.7	0.974	0.946	0.980	0.974	0.942	0.867	0.975	0.908	0.721
		ρ=0.9	0.912	0.958	0.869	0.912	0.891	0.649	0.911	0.807	0.541
Equal coefficients-extreme	Prop.=0.5	ρ=0.5	0.989	0.748	0.966	0.981	0.741	0.957	0.958	0.742	0.899
		ρ=0.7	0.988	0.780	0.962	0.970	0.773	0.915	0.949	0.800	0.778
		ρ=0.9	0.883	0.777	0.917	0.826	0.757	0.698	0.758	0.734	0.619
	Prop.=0.25	ρ=0.5	0.990	0.842	0.965	0.988	0.845	0.971	0.987	0.851	0.902
		ρ=0.7	0.959	0.882	0.977	0.962	0.873	0.925	0.960	0.872	0.784
		ρ=0.9	0.942	0.898	0.945	0.942	0.853	0.731	0.942	0.802	0.622
	Prop.=0.1	ρ=0.5	0.987	0.915	0.976	0.984	0.912	0.973	0.978	0.915	0.899
		ρ=0.7	0.966	0.937	0.981	0.967	0.936	0.933	0.963	0.919	0.763
		ρ=0.9	0.901	0.952	0.957	0.903	0.922	0.718	0.902	0.847	0.616
Unequal coefficients-middle	Prop.=0.5	ρ=0.5	0.772	0.746	0.863	0.767	0.746	0.790	0.771	0.727	0.737
		ρ=0.7	0.819	0.803	0.850	0.826	0.754	0.749	0.829	0.741	0.678
		ρ=0.9	0.897	0.769	0.770	0.895	0.728	0.621	0.883	0.670	0.520
	Prop.=0.25	ρ=0.5	0.861	0.841	0.870	0.852	0.815	0.797	0.852	0.777	0.728
		ρ=0.7	0.964	0.856	0.843	0.958	0.816	0.761	0.956	0.760	0.654
		ρ=0.9	0.940	0.831	0.764	0.941	0.780	0.610	0.937	0.707	0.518
	Prop.=0.1	ρ=0.5	0.941	0.885	0.877	0.940	0.834	0.804	0.912	0.799	0.729
		ρ=0.7	0.983	0.896	0.852	0.982	0.848	0.749	0.980	0.786	0.650
		ρ=0.9	0.916	0.862	0.764	0.916	0.806	0.608	0.915	0.745	0.519
Unequal coefficients-extreme	Prop.=0.5	ρ=0.5	0.798	0.745	0.869	0.809	0.749	0.810	0.797	0.717	0.755
		ρ=0.7	0.854	0.781	0.861	0.855	0.756	0.767	0.852	0.750	0.722
		ρ=0.9	0.790	0.766	0.789	0.796	0.728	0.682	0.784	0.711	0.590
	Prop.=0.25	ρ=0.5	0.898	0.824	0.883	0.891	0.802	0.815	0.893	0.766	0.752
		ρ=0.7	0.970	0.849	0.866	0.973	0.801	0.777	0.971	0.791	0.710
		ρ=0.9	0.960	0.847	0.797	0.956	0.780	0.691	0.945	0.755	0.603
	Prop.=0.1	ρ=0.5	0.966	0.881	0.885	0.960	0.840	0.814	0.944	0.816	0.756
		ρ=0.7	0.976	0.895	0.873	0.973	0.857	0.777	0.971	0.806	0.701
		ρ=0.9	0.908	0.883	0.802	0.909	0.828	0.681	0.909	0.779	0.601

V: PLS-VIP, L: Lasso, S: Stepwise.

about true and predicted classes. Table 1 shows the confusion matrix and the meanings of its entries in the context of our study.

From Table 1 accuracy, sensitivity, and specificity are respectively defined as Eqs. (12)–(14).

$$\text{Accuracy} = (a+d)/(a+b+c+d) \tag{12}$$

$$\text{Sensitivity} = d/(c+d) \tag{13}$$

$$\text{Specificity} = a/(a+b) \tag{14}$$

In the classification area, the usual performance measure is accuracy which is the proportion of predictors correctly classified. However, it may not be suitable when there is an imbalance between the numbers of irrelevant and relevant predictors. For example, consider the case where the proportion of relevant predictors equals 0.1. A method that always classifies all predictors as irrelevant will achieve an accuracy of 90%. Although this looks high, the method would be useless because it totally fails to select relevant predictors.

Thus, instead of using accuracy as the overall performance measure, we suggest using G, the geometric mean of sensitivity (the proportion of selected relevant predictors among relevant predictors) and specificity (the proportion of unselected irrelevant predictors among irrelevant predictors) [11] as in Eq. (15).

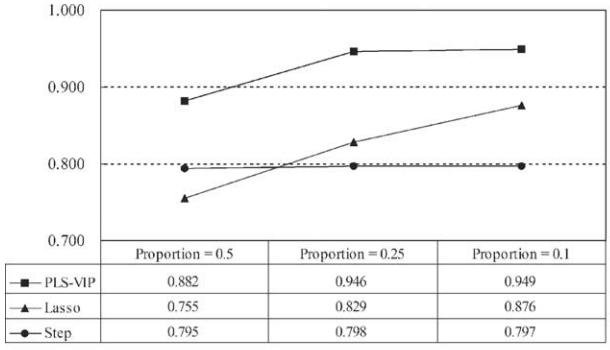
$$G = (\text{Sensitivity} \times \text{Specificity})^{1/2} \tag{15}$$

The value of G ranges between 0 and 1. The values close to 1 imply that most predictors are classified correctly. As mentioned in ref. [11], this measure has the distinctive property of being independent of the numbers of relevant and irrelevant predictors, and is thus robust regardless of the level of proportion.

4. Results and discussion

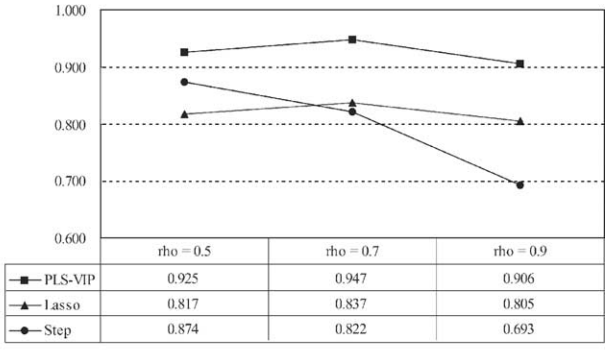
4.1. PLS-VIP method vs. the Lasso or Stepwise method

The number of latent variables for PLS regression, the tuning parameter for the Lasso and the significant levels for stepwise regression are determined by five-fold cross-validation which is widely used for estimating prediction error [12]. As mentioned before, 100 replications for each of 108 cases are made to evaluate the performance of the variable selection methods. At each replication, performance measure of G was calculated. In addition, the root mean squared error (RMSE) of predicted response for each method was also obtained to examine prediction performance.



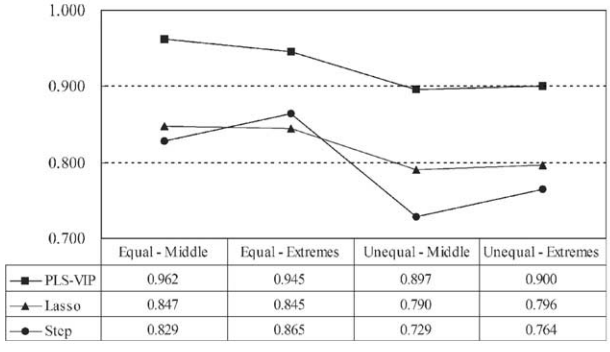
	Proportion = 0.5	Proportion = 0.25	Proportion = 0.1
PLS-VIP	0.882	0.946	0.949
Lasso	0.755	0.829	0.876
Step	0.795	0.798	0.797

(a) Proportion



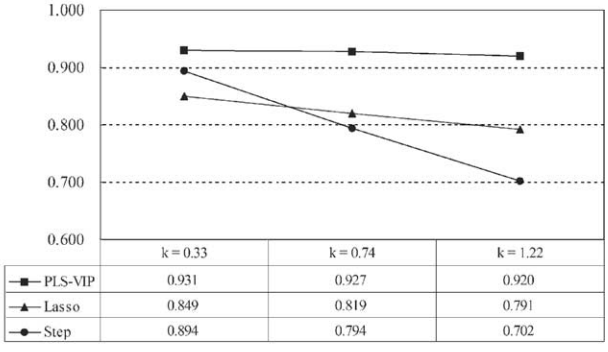
	rho = 0.5	rho = 0.7	rho = 0.9
PLS-VIP	0.925	0.947	0.906
Lasso	0.817	0.837	0.805
Step	0.874	0.822	0.693

(b) Correlation



	Equal - Middle	Equal - Extremes	Unequal - Middle	Unequal - Extremes
PLS-VIP	0.962	0.945	0.897	0.900
Lasso	0.847	0.845	0.790	0.796
Step	0.829	0.865	0.729	0.764

(c) Coefficients



	k = 0.33	k = 0.74	k = 1.22
PLS-VIP	0.931	0.927	0.920
Lasso	0.849	0.819	0.791
Step	0.894	0.794	0.702

(d) Noise

Fig. 2. Mean G of each method according to factors.

4.1.1. Comparison based on G

Table 2 summarizes the simulation results of variable selection performance by using the average G over 100 replications along the cases. The bold figures denote the best ones. As seen in Table 2, in most cases, the PLS-VIP method outperforms the other methods particularly when the error variance is large in data set or when the model fitness (R-square) is relatively low. On the other hand, Fig. 2 shows the average G of each method according to factors. We confirm again that the PLS-VIP method outperforms the other methods over all factors. Besides, the PLS-VIP method seems to be insensitive to noise while the others seem to be sensitive.

4.1.2. Comparison based on RMSE

Although the objective of this study is not to compare performance of response prediction, we provide estimates of the RMSEs as in Eq. (16) for different methods as supplement information.

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{500} (y_i - \hat{y}_i)^2 / 500} \tag{16}$$

Table 3 summarizes the simulation results of prediction performance by using the average RMSE over 100 runs along the cases. The best method is also shown in bold type. Unlike variable selection performance, PLS does not outperform the other methods in some cases. This means that there may not be a strong relation between variable selection and prediction performance.

4.2. PLS-VIP method vs. PLS-BETA method

Now, we also compare the PLS-VIP method and the PLS-BETA method. Assuming that the number of relevant

Table 3
Mean RMSE of each method along the cases

			k=0.33			k=0.74			k=1.22		
			P	L	S	P	L	S	P	L	S
Equal coefficients-middle	Prop.=0.5	ρ=0.5	1.650	1.654	1.656	3.722	3.707	3.710	6.114	6.069	6.074
		ρ=0.7	2.086	2.083	2.087	4.712	4.682	4.688	7.720	7.681	7.701
		ρ=0.9	2.792	2.773	2.776	6.202	6.172	6.187	10.39	10.35	10.37
	Prop.=0.25	ρ=0.5	1.640	1.655	1.655	3.655	3.667	3.663	6.150	6.124	6.121
		ρ=0.7	2.081	2.081	2.082	4.716	4.689	4.684	7.760	7.741	7.758
		ρ=0.9	2.812	2.790	2.795	6.231	6.205	6.221	10.39	10.37	10.36
	Prop.=0.1	ρ=0.5	1.597	1.657	1.643	3.536	3.646	3.612	6.014	6.081	6.035
		ρ=0.7	2.031	2.071	2.051	4.586	4.639	4.611	7.611	7.656	7.609
		ρ=0.9	2.722	2.743	2.734	6.190	6.233	6.214	10.16	10.18	10.15
Equal coefficients-extreme	Prop.=0.5	ρ=0.5	1.522	1.526	1.528	3.438	3.427	3.430	5.659	5.639	5.644
		ρ=0.7	1.815	1.810	1.812	4.087	4.058	4.062	6.700	6.667	6.686
		ρ=0.9	2.405	2.383	2.386	5.342	5.310	5.320	8.786	8.756	8.769
	Prop.=0.25	ρ=0.5	1.504	1.519	1.519	3.403	3.412	3.416	5.656	5.668	5.670
		ρ=0.7	1.792	1.796	1.797	4.057	4.056	4.053	6.652	6.644	6.649
		ρ=0.9	2.158	2.155	2.155	4.870	4.842	4.855	7.978	7.960	7.970
	Prop.=0.1	ρ=0.5	1.446	1.510	1.498	3.234	3.343	3.319	5.505	5.607	5.562
		ρ=0.7	1.737	1.783	1.771	3.912	3.983	3.950	6.498	6.622	6.586
		ρ=0.9	2.063	2.091	2.086	4.676	4.724	4.726	7.765	7.811	7.793
Unequal coefficients-middle	Prop.=0.5	ρ=0.5	22.40	22.43	22.46	50.22	50.25	50.29	82.71	82.71	82.71
		ρ=0.7	26.81	26.81	26.85	60.16	60.16	60.21	99.17	99.10	99.18
		ρ=0.9	32.57	32.51	32.58	72.54	72.43	72.56	121.6	121.3	121.4
	Prop.=0.25	ρ=0.5	22.08	22.36	22.37	49.59	50.11	50.08	82.62	83.19	82.90
		ρ=0.7	26.63	26.91	26.88	60.10	60.41	60.34	98.69	99.03	98.98
		ρ=0.9	32.26	32.38	32.40	72.86	72.91	72.97	119.7	119.7	119.9
	Prop.=0.1	ρ=0.5	21.22	22.32	22.13	48.04	50.09	49.43	82.29	83.76	82.72
		ρ=0.7	25.74	26.79	26.57	58.46	60.03	59.53	96.08	98.14	97.23
		ρ=0.9	31.48	32.37	32.25	71.28	72.31	72.05	118.4	119.1	118.5
Unequal coefficients-extreme	Prop.=0.5	ρ=0.5	19.12	19.20	19.21	42.99	43.11	43.17	70.83	70.95	71.04
		ρ=0.7	21.40	21.43	21.47	48.35	48.35	48.43	79.33	79.43	79.47
		ρ=0.9	26.33	26.28	26.35	58.58	58.56	58.63	96.97	96.66	96.87
	Prop.=0.25	ρ=0.5	18.82	19.13	19.15	42.42	42.92	42.92	70.20	71.02	70.96
		ρ=0.7	21.09	21.34	21.35	47.89	48.24	48.28	78.33	78.77	78.72
		ρ=0.9	23.84	24.03	24.06	54.36	54.45	54.61	90.29	90.26	90.36
	Prop.=0.1	ρ=0.5	17.91	19.04	18.86	40.62	42.52	42.07	67.91	70.73	70.00
		ρ=0.7	20.08	21.15	20.99	46.33	47.88	47.55	77.15	79.16	78.39
		ρ=0.9	23.09	23.83	23.80	52.68	53.74	53.65	87.39	88.42	88.33

P: PLS, L: Lasso, S: Stepwise.

Table 4
Comparison between PLS-VIP and PLS-BETA method

			Equal coefficients						Unequal coefficients					
			k=0.33		k=0.74		k=1.22		k=0.33		k=0.74		k=1.22	
			V	B	V	B	V	B	V	B	V	B	V	B
Location of coefficients: middle	Prop.=0.5	ρ=0.5	9.99	10	10	9.99	9.91	9.91	9.41	8.5	9.47	8.15	9.32	7.98
		ρ=0.7	9.96	10	9.94	9.91	9.9	9.83	9.65	8.35	9.72	7.93	9.65	8
		ρ=0.9	9.99	9.9	9.9	9.69	9.76	9.07	9.93	8.23	9.83	8.13	9.69	8.15
	Prop.=0.25	ρ=0.5	10	10	9.98	10	9.98	9.98	9.4	8.27	9.36	7.94	9.03	7.59
		ρ=0.7	9.99	10	9.95	9.98	9.91	9.9	9.77	8.12	9.68	8.07	9.61	7.88
		ρ=0.9	9.98	9.98	9.9	9.82	9.71	9.41	9.95	8.18	9.85	8.3	9.68	8.26
	Prop.=0.1	ρ=0.5	10	10	9.96	10	9.96	9.95	9.28	8.04	9.13	7.72	8.51	8.15
		ρ=0.7	9.98	10	9.94	9.99	9.88	9.87	9.82	8.14	9.76	8.21	9.53	8.14
		ρ=0.9	10	10	9.9	9.89	9.75	9.66	9.93	8.35	9.86	8.92	9.72	8.99
Location of coefficients: extreme	Prop.=0.5	ρ=0.5	9.99	10	9.96	9.99	9.86	9.94	9.44	8.66	9.46	7.92	9.24	7.93
		ρ=0.7	9.95	10	9.82	9.97	9.68	9.88	9.64	8.42	9.56	7.96	9.4	7.84
		ρ=0.9	8.7	9.96	8.14	9.82	7.36	9.55	7.75	8.16	7.65	8	7.58	8.38
	Prop.=0.25	ρ=0.5	9.99	10	9.99	10	9.89	9.95	9.52	8.2	9.33	7.69	9.24	7.33
		ρ=0.7	9.99	10	9.84	10	9.68	9.9	9.58	8.01	9.62	7.54	9.57	7.75
		ρ=0.9	9.85	10	9.73	9.89	9.45	9.69	9.55	7.83	9.51	7.86	9.22	8.43
	Prop.=0.1	ρ=0.5	9.98	10	10	10	9.87	9.87	9.38	8.01	9.23	7.51	8.77	7.42
		ρ=0.7	9.98	10	9.92	10	9.74	9.93	9.7	7.89	9.6	7.87	9.48	8.13
		ρ=0.9	9.85	10	9.6	9.97	9.29	9.8	9.72	8.11	9.5	8.46	9.42	8.78

V: PLS-VIP, B: PLS-BETA.

predictors is known as K (i.e., K=10), we select predictors having the first K largest VIP scores as well as those having the first K largest absolute coefficients, then compare the performance using average number of relevant predictors among K selected predictors over 100 runs. As seen in Table 4, when the structure of coefficients is equal, PLS-BETA method outperforms PLS-VIP method in most cases. When the structure of coefficients is unequal, the PLS-VIP method outperforms the PLS-BETA method.

4.3. Cutoff values of the PLS-VIP method

Although the performance of PLS-VIP method may depend on the cutoff value, 'greater than one rule' is generally used to select relevant predictors. The proper cutoff value of the PLS-VIP method according to 108 cases is provided by averaging v_i* which is defined by Eq. (17) where v varies from 0.01 to 3 with increments of 0.01. Here, G is a concave function of v and the subscript i is for a replication in a case. Table 5 shows the mean proper cutoff

Table 5
Mean proper cutoff values along the cases

			Equal coefficients			Unequal coefficients		
			k=0.33	k=0.74	k=1.22	k=0.33	k=0.74	k=1.22
Location of coefficients: middle	Prop.=0.5	ρ=0.5	0.80	0.81	0.80	0.64	0.66	0.67
		ρ=0.7	0.87	0.88	0.85	0.79	0.78	0.79
		ρ=0.9	0.97	0.94	0.94	0.93	0.94	0.93
	Prop.=0.25	ρ=0.5	0.93	0.93	0.95	0.71	0.72	0.71
		ρ=0.7	1.04	1.02	1.04	0.92	0.91	0.90
		ρ=0.9	1.08	1.09	1.07	1.05	1.04	1.03
	Prop.=0.1	ρ=0.5	1.11	1.15	1.16	0.84	0.87	0.88
		ρ=0.7	1.32	1.33	1.31	1.14	1.12	1.07
		ρ=0.9	1.39	1.40	1.38	1.35	1.33	1.30
Location of coefficients: extreme	Prop.=0.5	ρ=0.5	0.83	0.84	0.84	0.67	0.66	0.71
		ρ=0.7	0.93	0.93	0.90	0.83	0.84	0.81
		ρ=0.9	0.94	0.80	0.92	0.87	0.90	0.84
	Prop.=0.25	ρ=0.5	1.00	1.00	0.99	0.78	0.80	0.78
		ρ=0.7	1.12	1.08	1.07	0.94	0.92	0.94
		ρ=0.9	1.06	1.05	1.02	0.98	1.00	0.96
	Prop.=0.1	ρ=0.5	1.22	1.24	1.22	0.94	0.92	0.98
		ρ=0.7	1.44	1.41	1.36	1.18	1.16	1.15
		ρ=0.9	1.34	1.37	1.33	1.28	1.28	1.26

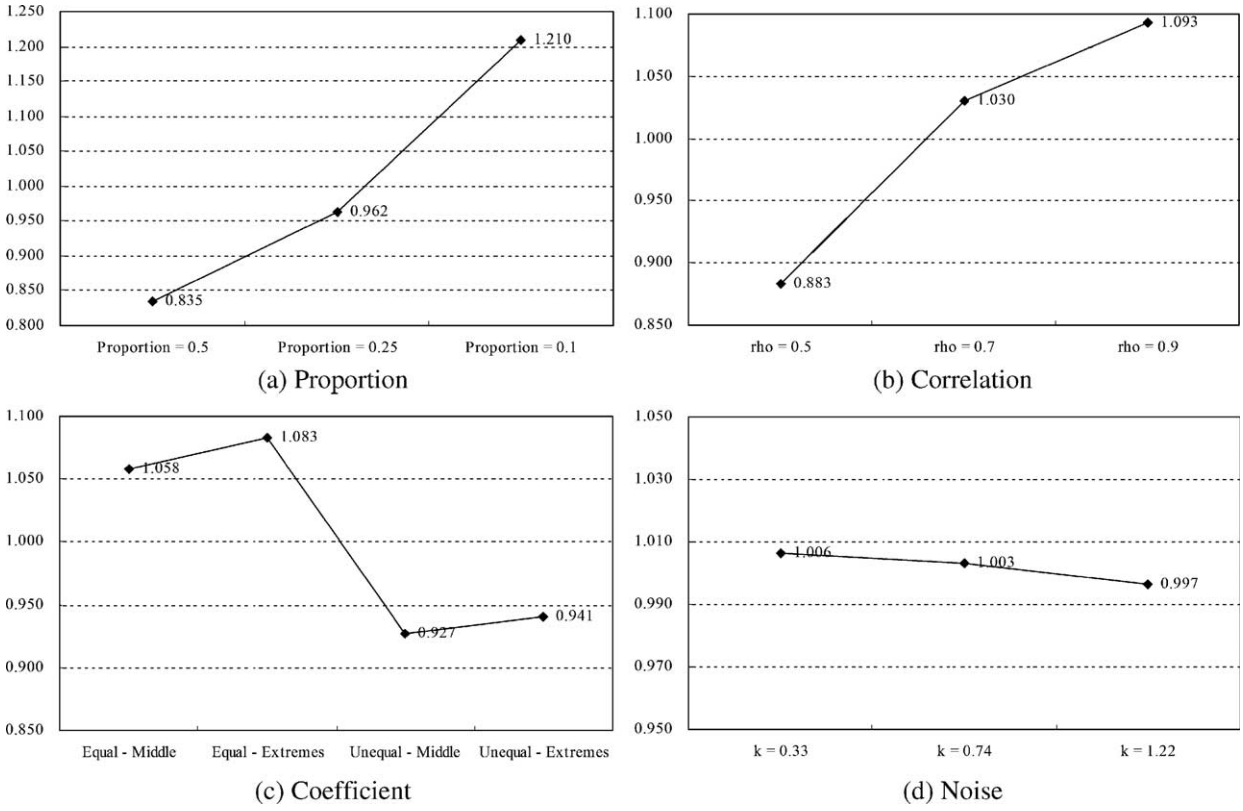


Fig. 3. Mean proper cutoff values according to factors.

values along the cases. As seen, some cases require higher or lower cutoff values than one to increase the variable selection performance of the PLS-VIP method.

$$v_i^* = \left\{ Min\left(\underset{v\in\{0.01,0.02,..,3\}}{\arg\max}\ G(v) \right) + Max\left(\underset{v\in\{0.01,0.02,..,3\}}{\arg\max}\ G(v) \right) \right\} / 2 \qquad (17)$$

Fig. 3 shows the effect plots of mean proper cutoff values according to factors. When proportion of relevant predictors is low, the magnitude of correlation is high, or the structure of coefficients is equal, the proper cutoff value is required to be greater than one.

5. Conclusions

In this paper, we conducted 10,800 experiments to explore the nature of the PLS-VIP method as compared with other variable selection methods. Experiments were designed by considering four factors including the proportion of the number of relevant predictors among total predictors, the magnitude of correlations between predictors, the structure of regression coefficients, and the magnitude of signal to noise.

First, the PLS-VIP method was compared with the Lasso and Stepwise method. The PLS-VIP method performed excellently in identifying relevant predictors and outperformed the other methods. It was also found that a model with good fitness performance may not guarantee good variable selection performance. Thus, for the purpose of selecting relevant process variables, process engineers must be careful when using model performance such as RMSE, R-squares, etc.

Second, the PLS-VIP method was compared with the PLS-BETA method. We found an interesting observation that PLS-VIP and PLS-BETA method might be complementary. So, if we use a strategy which combines these two methods for selecting relevant predictors, better variable selection performance could be achieved. Actually, Wold et al. [6] recommend a combination of PLS-VIP and PLS-Beta for variable selection, which states that both should be small for a variable to be excluded.

Finally, proper cutoff values of the PLS-VIP method were provided to judge whether using the 'greater than one rule' is adequate or not. The proper cutoff value may be higher than 1 under low proportion, high correlation, or an equal coefficients structure.

Acknowledgements

We would like to thank two anonymous referees for their valuable comments that have led to a substantial

improvement in the paper. This work was supported by the Brain Korea 21 project and by the Systems Bio-Dynamics Research Center at POSTECH.

References

[1] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[2] R. Tibshirani, J. R. Stat. Soc. 58 (1996) 267–288.
[3] D.C. Montgomery, E.A. Peck, G.G. Vining, Introduction to Linear Regression Analysis, 3rd ed., Wiley, New York, NY, 2001, pp. 131–154.
[4] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1–17.
[5] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, Multi-and Megavariate Data Analysis; Principles and Applications, Umetrics Academy, Umea, Sweden, 2001.
[6] S. Wold, E. Johansson, M. Cocchi, 3D QSAR in Drug Design; Theory, Methods, and Applications, ESCOM, Leiden, Holland, 1993, pp. 523–550.
[7] I.S. Helland, Communications in Statistics–Simulation and Computation 17 (1988) 581–607.
[8] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Ann. Stat. 32 (2) (2004) 407–499.
[9] C.L. Mallows, Technometrics 15 (1973) 661–675.
[10] U. Grenander, G. Szego, Toeplitz Forms and their Applications, University of California Press, Berkeley, 1958.
[11] M. Kubat, R.C. Holte, S. Matwin, Mach. Learn. 30 (1998) 195–215.
[12] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, New York, 2001, pp. 214–217.