International Journal of Pure and Applied Mathematics
Volume 94 No. 3 2014, 307-322
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
doi: http://dx.doi.org/10.12732/ijpam.v94i3.2

ijpam.eu

CUTOFF THRESHOLD OF VARIABLE IMPORTANCE IN PROJECTION FOR VARIABLE SELECTION

Noppamas Akarachantachote[1], Seree Chadcham[2], Kidakan Saithanu[3 §]

[1,2]College of Research Methodology and Cognitive Science
Burapha University
THAILAND
[2]Department of Mathematics
Burapha University
169, Tambon Saensook, Amphur Muang, Chonburi, 20131, THAILAND

Abstract: At present, variable selection turns to prominence since it obviously alleviate a trouble of measuring multiple variables per sample. The partial least squares regression (PLS-R) and the score of Variable Importance in Projection (VIP) are combined together for variable selection. The value of VIP score which is greater than 1 is the typical rule for selecting relevant variables. Due to a constant cutoff threshold is not sometimes suitable for every data structure, a new cutoff threshold for VIP in classification task has been proposed and then compared to the classical one thru the interesting situation simulation. There were 180 situations generated based on four parameters: Percentage of the number of relevant variables, Magnitude of mean difference of relevant variables between two groups, Degree of correlation between relevant variables, and the sample size. The result of this study presents that the new cutoff threshold can improve in identifying relevant variables more than the previous threshold as seeing of good value of the average balanced accuracy in most of situations.

AMS Subject Classification: 62H30
Key Words: variable selection, VIP, PLS-R

Received: August 4, 2013

[§]Correspondence author

1. Introduction

Because of the progressive technology in the past decade, a large number of data can be accumulated. Data set with hundreds or thousands of attributes is called high dimensional data. For example, microarray data which is a lot of biological data of tissues is derived from DNA microarray experiments. The experiment allows simultaneous measurement of tens of thousands of gene expression levels per sample. However, the number of samples from the microarray experiment usually contains less than one hundred samples. The number of genes (variables) in data then far exceeds the number of samples. Such data set presents great challenges in data analysis because some existing methods of data analysis can not support it. Furthermore, each of gene does not hold for relevant information. There are only 5% of total genes containing relevant information about the grouping [1]. Therefore, selecting a subset of relevant genes and then using only some of them for the subsequent data analysis is essential.

Variable selection is the process of determining relevant variables from the original variable set. It offers several advantages such as avoiding overfitting, improving model performance, providing faster and more cost-effective models and gaining a deeper insight into the underlying processes. The methods of variable selection in the viewpoint of classification can be classified into three categories: filter, wrapper and embedded methods. Existing methods for variable selection reviewed in [2] was mentioned as a good review. For high dimensional data like microarray data, wrapper and embedded methods spend much of time in contrast to the filter method which considers only the intrinsic properties of the classification independence. Since it is independent and it performs only once for all classification algorithms, it can be computed fast and simply. Filter method is divided into two types corresponding to dependency of variable (univariate and multivariate). Univariate type considers each variable as independence from other variables while multivariate type includes variable dependency for selecting the relevant variable subset.

The VIP is a measurement including variable dependency which is considered as the benefit of multivariate filter method. In the situation of high dimensionality, it usually involves with correlation between variables and missing of observations or variables more than samples. Under this circumstance, nowadays the VIP score obtained by PLS-R has been paid an increasing attention as a significant measurement of each predictor variable [3], [4], [5]. Normally, the average of the squared values of the VIPs is equal to 1. The criterion of VIP value with greater than 1 is then often used as a cutoff point for variable selection [3], [5], [6], [7]. Predictor variables with the value of VIP score greater than 1 will be selected. However, data structures are generally diverse. The cutoff threshold then should not be the same in different type of data structure [3]. Determining the appropriate cutoff threshold is not simple. Too high value of cutoff threshold will lead to absent of some crucial variables. Oppositely, too low value of cutoff threshold will reach to more unrelated variables.

For this study, the new cutoff threshold of VIP is proposed for identification of relevant variables relying on the use of detection outlier with boxplot obtained from the added noise variables to estimate the cutoff threshold.

The rest of this paper is organized as follows. Section 2 presents background and related works. Section 3 describes the methodology. The results and discussions are given in Section 4. Final conclusions are concluded in Section 5.

2. Background and Related Works

2.1. The Approach of PLS-R

Partial least square (PLS) is the name of a set of algorithms developed in the 1960s and 1970s by Herman Wold to address problems in econometric path modeling. It was then subsequently adopted by his son Svante Wold and friends in the 1980s for regression problems in chemometric and spectrometric modeling [8] called partial least squares regression (PLS-R). The advantage of the PLS-R is handling data sets with many noisy, collinear variables and missing values. Additionally, the assumption of error distribution is not required in the PLS-R [9]. The number of PLS-R applications is steadily increasing in research fields such as bioinformatics, machine leaning and chemometrics [10].

The relationship between blocks of observed variables and means of latent variables of the PLS-R model is called components. These components are linear transformations of the original predictor variables which have high covariance with the response variables. In case of single response variable $\mathbf{y}$ and $p$predictor variables of $\mathbf{X}$basing on these components, $\mathbf{X}$ and $\mathbf{y}$ are decomposed as of Equation 1 and Equation 2, respectively.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{'} + \mathbf{E} \tag{1}$$

$$\mathbf{y} = \mathbf{T}\mathbf{q}^{'} + \mathbf{f} \tag{2}$$

where $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_h] \in \mathbf{R}^{n\times h}$ represents the sample sized n of the $h$components, $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_h] \in \mathbf{R}^{p\times h}$ and $\mathbf{q} = [q_1, \ldots, q_h] \in \mathbf{R}^{1\times h}$ denotes as loadings of $\mathbf{X}$

and $\mathbf{y}$, respectively. Generally, $\mathbf{P}$ and $\mathbf{q}$ are computed by ordinary least squares (OLS). $\mathbf{E}$ and $\mathbf{f}$are residuals of $\mathbf{X}$ and $\mathbf{y}$, respectively.

The construction of components is the major point of PLS-R. The components are the linear transformations of $\mathbf{X}$ which maximize covariance between response variable $\mathbf{y}$ and components. The approach of finding each of components is done sequentially. For the first component ($\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$), it is determined by maximizing the covariance between $\mathbf{y}$ and $\mathbf{t}_1$ under the constraint of$\|\mathbf{w}_1\| = 1$. To extract each other components, original matrix $\mathbf{X}$ and $\mathbf{y}$ has to be reconstructed by substituting of their residuals. This process is called deflation of matrices $\mathbf{X}$ and $\mathbf{y}$. The residuals of $\mathbf{X}$ and $\mathbf{y}$ for the first component are found out as of Equation 3 and Equation 4, respectively.

$$\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^{'} \tag{3}$$

$$\mathbf{f}_1 = \mathbf{y} - \mathbf{t}_1 q_1 \tag{4}$$

where $\mathbf{p}_1$ and q_1 are loadings defined by OLS fitting.

Also, the residual of a^{th}components $\mathbf{X}$and $\mathbf{y}$ are computed as of Equation 5 and Equation 6, respectively.

$$\mathbf{E}_a = \mathbf{E}_{a-1} - \mathbf{t}_a\mathbf{p}_a^{'} \tag{5}$$

$$\mathbf{f}_a = \mathbf{f}_{a-1} - \mathbf{t}_a q_a \tag{6}$$

where $\mathbf{E}_0 = \mathbf{X}$ and $\mathbf{f}_0 = \mathbf{y}$.

There are various approaches of PLS-R. The PLS-R above is called PLS1. More detailed variants of PLS can be found in [11]. The particular algorithm of PLS1 is given in Figure 1. $\mathbf{X}$ and$\mathbf{y}$ have been standardized to have mean 0 and unit variance before starting the procedure. The number of components (h) has to be determined at first time. There are many techniques to design the number of components. Some authors suggested to fixed the number of components from three to five [12], [13], [14] while as others recommended to identify the size of the space by classification performance of cross-validation [15].

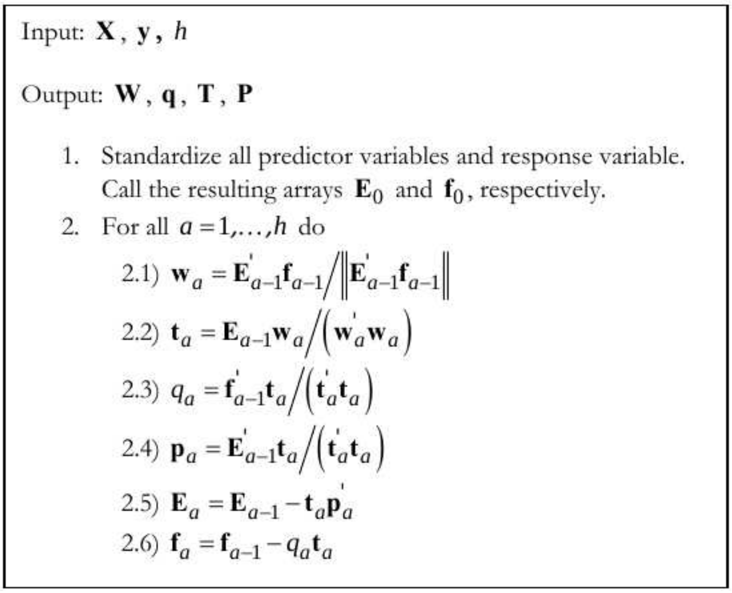
2.2. The VIP Score

The VIP score first published by Wold and others in 1993 [3] measures explicative power of predictor variables with respect to the response variable which basing on the PLS-R. The VIP score of variable$j$ is calculated as of Equation

7.

$$\mathrm{VIP}_j = \sqrt{\frac{\sum\limits_{a=1}^{h} \mathrm{R}^2(y,t_a)\,(\mathrm{w}_{aj}/\,\|w_a\|)^2}{(1/p)\sum\limits_{a=1}^{h} \mathrm{R}^2(y,t_a)}}, \tag{7}$$

where w_{aj} is weight of the j^{th}predictor variable in component a and $\mathrm{R}^2(y,t_a)$ is fraction of variance in $\mathbf{y}$ explained by the component a. The variable with higher value of VIP score shows that it is more relevant to predict the response variable.



```
Input: $\mathbf{X}$, $\mathbf{y}$, $h$

Output: $\mathbf{W}$, $\mathbf{q}$, $\mathbf{T}$, $\mathbf{P}$

    1. Standardize all predictor variables and response variable.
       Call the resulting arrays $\mathbf{E}_0$ and $\mathbf{f}_0$, respectively.
    2. For all $a = 1,\ldots,h$ do
        2.1) $\mathbf{w}_a = \mathbf{E}'_{a-1}\mathbf{f}_{a-1} / \|\mathbf{E}'_{a-1}\mathbf{f}_{a-1}\|$
        2.2) $\mathbf{t}_a = \mathbf{E}_{a-1}\mathbf{w}_a / (\mathbf{w}'_a\mathbf{w}_a)$
        2.3) $q_a = \mathbf{f}'_{a-1}\mathbf{t}_a / (\mathbf{t}'_a\mathbf{t}_a)$
        2.4) $\mathbf{p}_a = \mathbf{E}'_{a-1}\mathbf{t}_a / (\mathbf{t}'_a\mathbf{t}_a)$
        2.5) $\mathbf{E}_a = \mathbf{E}_{a-1} - \mathbf{t}_a\mathbf{p}'_a$
        2.6) $\mathbf{f}_a = \mathbf{f}_{a-1} - q_a\mathbf{t}_a$
```

Figure 1: Algorithm of PLS-R

2.3. Related Work

Two main problems encounter when high dimensional data are analyzed. Firstly, the number of predictors is larger than the sample size. Secondly, there is multicollinearity among predictor variables. Therefore, irrelevant variable should be eliminated from the data set before analyzing. The VIP has been used in microarray data to measure the importance of variables (genes) [16], [17], [18]. There are several techniques in the use of VIP. Most of works selected variables with the value of VIP score more than a constant value such as 1 [6], [16], or 2 [18]. Some studies like [17] used the VIP score to rank variables and choose the top k values. The other created new significant index based on the VIP [6].

The proposed method is compared to the works mentioned above as follows. Randomization of the order of the samples for generating noise variables applied from [19] is assessed to generating noise variable randomly. The use of VIP for ranking variable importance is evaluated to the classical of PLS-R coefficient [20], [21], [22], weight vector ($\mathbf{w}_1$) [19], and t-statistic [12]. Finally, consideration of cutoff threshold by use of boxplot is appraised to the using maximum value of importance index of noise variable [20], percentile of importance index of noise variable [19], [20], and range of importance index based on the t-Students distribution [22].

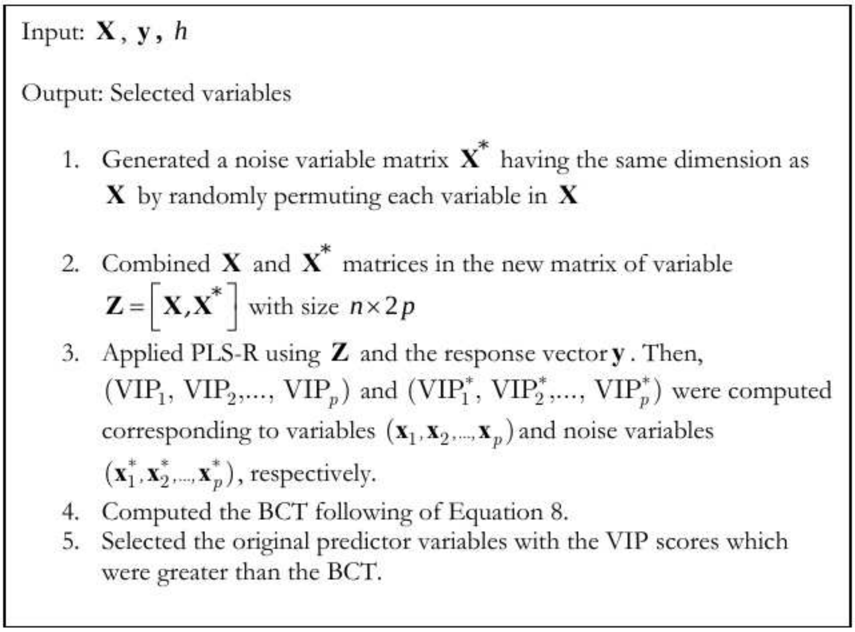
3. Methodology

The cutoff threshold presented here has many significant steps. Adding noise variables to the original data set is firstly and then computing the VIP scores of them. The VIP scores are always equal to or greater than 0 while only the VIP scores of noise variables ($\text{VIP}_{\text{noise}}$) should be closed to 0 because they are not relevant to the predict response variable. However, a chance of the $\text{VIP}_{\text{noise}}$ is far from 0 which will be probably identified as outlier. The outliers are observations inconsistent with other observations in the data set which is less likely to cause from the same population with other observations. Therefore, the outliers of $\text{VIP}_{\text{noise}}$ will be considered as scores of VIP of relevant variables. The cutoff threshold for detecting outlier is applied in selection of pertinent variables by estimating with boxplot. A boxplot demonstrated by Tukey [23] is a graphical display of data dispersion. It indicates which observations regarded as outliers. Without any of assumptions underlying statistical distribution, boxplot is suitable method for detecting outlier of $\text{VIP}_{\text{noise}}$. In addition, only the upper detection is required because the lower VIP represents that the variables are irrelevant. Boxplot Cutoff Threshold (BCT) is defined as of Equation 8.

$$BCT = Q_3 + 1.5 \times IQR \tag{8}$$

where Q_1 and Q_3 are lower and upper quartile of $\text{VIP}_{\text{noise}}$, respectively and the IQR is the difference between Q_3 and Q_1 called the interquartile range.

The algorithm of selecting variables via VIP with BCT (VIP-BCT) is shown as of Figure 2.



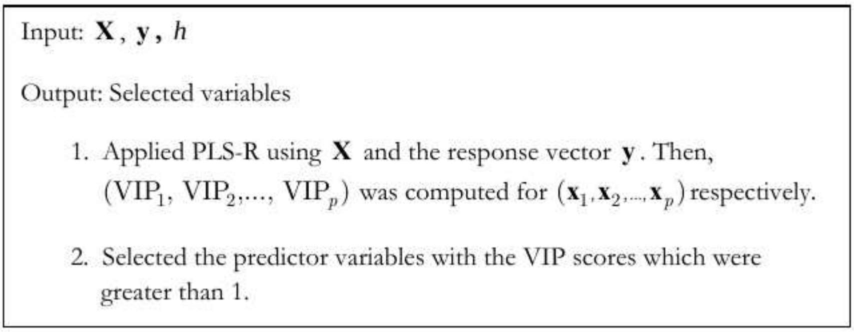
Input: $\mathbf{X}$, $\mathbf{y}$, h

Output: Selected variables

1. Generated a noise variable matrix $\mathbf{X}^*$ having the same dimension as $\mathbf{X}$ by randomly permuting each variable in $\mathbf{X}$
2. Combined $\mathbf{X}$ and $\mathbf{X}^*$ matrices in the new matrix of variable $\mathbf{Z} = \left[\mathbf{X}, \mathbf{X}^*\right]$ with size $n \times 2p$
3. Applied PLS-R using $\mathbf{Z}$ and the response vector $\mathbf{y}$. Then, $(\text{VIP}_1, \text{VIP}_2, \ldots, \text{VIP}_p)$ and $(\text{VIP}_1^*, \text{VIP}_2^*, \ldots, \text{VIP}_p^*)$ were computed corresponding to variables $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ and noise variables $(\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_p^*)$, respectively.
4. Computed the BCT following of Equation 8.
5. Selected the original predictor variables with the VIP scores which were greater than the BCT.

Figure 2: Algorithm of the VIP-BCT

The algorithm of the VIP-1 shown in Figure 3 was compared to the use of VIP-BCT. The step 1 and step 2 of the VIP-BCT were lost here because the cutoff threshold in VIP-1 was fixed to 1.



Input: $\mathbf{X}$, $\mathbf{y}$, h

Output: Selected variables

1. Applied PLS-R using $\mathbf{X}$ and the response vector $\mathbf{y}$. Then, $(\text{VIP}_1, \text{VIP}_2, \ldots, \text{VIP}_p)$ was computed for $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ respectively.
2. Selected the predictor variables with the VIP scores which were greater than 1.

Figure 3: Algorithm of the VIP-1

3.1. Design of Simulation

Comparison between the algorithms of VIP-BCT and VIP-1 was made thru a simulation program. In this experimental, it focused on a binary classification problem. Defined the vector of the binary response $\mathbf{y} = (-1, \dots, -1,\ 1, \dots, 1)'$ and the matrix of predictor variables $\mathbf{X} = (\mathbf{X}_1|\mathbf{X}_2)_{n\times p}$, where n was the sample size, p was the number of predictor variables (equal to 2,000), $\mathbf{X}_1$ was the $n\times d$matrix corresponding to d truly relevant variables and $\mathbf{X}_2$ was the matrix of the remaining $p-d$irrelevant variables. Since normal distribution has been widely utilized for gene expression data simulation [24], the irrelevant variables $\mathbf{X}_2$ are independently drawn from it and the relevant variables $\mathbf{X}_1$ are generated from different distribution or the same distribution with distinguishable parameters. Thus, the irrelevant variables were drawn from normal distribution with $\mu = 0, \sigma = 1$ and the relevant variables were generated from multivariate normal distribution with mean and variance-covariance as described below.

There were four parameters required to simulate as following.

1. Four levels of the percentage of the number of relevant variables (Prel): (1) 1% or $d = 20$, (2) 3% or $d = 60$, (3) 5% or $d = 100$ and (4) 10% or $d = 200$.

2. Three levels of the magnitude of mean difference of relevant variables between two groups (Mdif): (1) 1 unit, $\mu_{-1} = (-0.5\ -0.5\dots-0.5)'$ and $\mu_{+1} = (0.5\,0.5\dots0.5)'$

(2) 3 unit, $\mu_{-1} = (-1.5\ -1.5\dots-1.5)'$ and $\mu_{+1} = (1.5\,1.5\dots1.5)'$

(3) 5 unit, $\mu_{-1} = (-2.5\ -2.5\dots-2.5)'$ and $\mu_{+1} = (2.5\,2.5\dots2.5)'$

3. Five degrees of correlations between relevant variables($\mathbf{\Sigma}$): (1) $\mathbf{\Sigma}_1 = \mathbf{I}_{d\times d}$,

$$(2)\mathbf{\Sigma}_2 = \begin{bmatrix} 1 & 0.5 & \dots & 0.5 \\ 0.5 & 1 & \dots & 0.5 \\ \vdots & \ddots & & \vdots \\ 0.5 & \dots & & 1 \end{bmatrix}_{d\times d},\ (3)\mathbf{\Sigma}_3 = \begin{bmatrix} 1 & 0.9 & \dots & 0.9 \\ 0.9 & 1 & \dots & 0.9 \\ \vdots & \ddots & & \vdots \\ 0.9 & \dots & & 1 \end{bmatrix}_{d\times d},$$

$$(4)\mathbf{\Sigma}_4 = \begin{bmatrix} 1 & (0.5) & (0.5)^2 & (0.5)^3 & \dots & (0.5)^{d-1} \\ (0.5) & 1 & (0.5) & (0.5)^2 & \dots & (0.5)^{d-2} \\ \vdots & & \ddots & & \vdots & \\ (0.5)^{d-1} & & \dots & & 1 & \end{bmatrix}_{d\times d} \text{ and}$$

Actual Class	**Predicted Class**	
	Relevant variable	**Irrelevant variable**
Relevant variable	a_1	a_2
Irrelevant variable	a_3	a_4

Table 1: Confusion matrix and descriptions of its entry

$$(5)\ \mathbf{\Sigma}_5 = \begin{bmatrix} 1 & (0.9) & (0.9)^2 & (0.9)^3 & \ldots & (0.9)^{d-1} \\ (0.9) & 1 & (0.9) & (0.9)^2 & \ldots & (0.9)^{d-2} \\ \vdots & \ddots & & & \vdots & \\ (0.9)^{d-1} & & \ldots & & 1 & \end{bmatrix}_{d\times d}$$

4. Three sample sizes(n) : (1) $n = 40$, (2) $n = 70$ and (3) $n = 100$

3.2. Measure of Performance

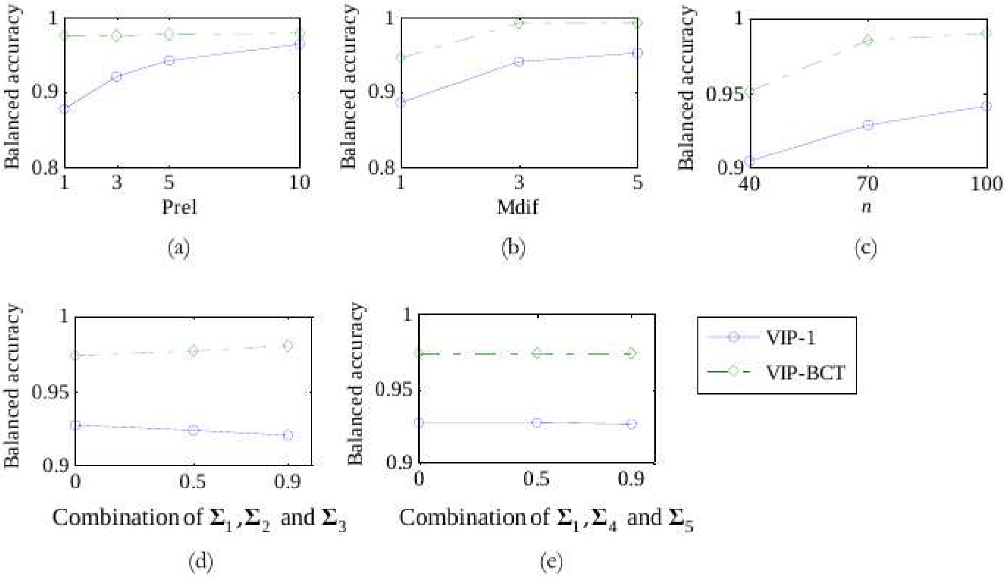
The balanced accuracy was applied and gauged to evaluate the both of performances between two different algorithms of cutoff threshold in variable selection. It is defined as the mean of sensitivity and specificity. Sensitivity is the ratio of the relevant variables classified correctly and the total number of variables while specificity is the ratio of irrelevant variables correctly classified and the total number of variables. Since relevant and irrelevant variable size here were not equal, the balanced accuracy was then chosen for evaluation instead of generally accuracy because of avoiding inflated performance estimates on unbalanced data sets.

Table 1 displayed the confusion matrix for balanced accuracy and descriptions of its entry.

From Table 1, a_1 is the number of relevant variables classified correctly, a_2 is the number of relevant variables classified incorrectly, a_3 is the number of irrelevant variables classified incorrectly and a_4 is the number of irrelevant variables classified correctly. Thus, sensitivity, specificity and balanced accuracy are respectively calculated as follows. $Sensitivity = \frac{a_1}{a_1+a_2}$, $Specificity = \frac{a_4}{a_3+a_4}$ and $Balanced\,accuracy = \frac{Sensitivity+Specificity}{2}$.

4. Results and Discussions

Three retaining components were fixed and 200 replications for each of 180 situations were made to evaluate performance between both of the two algorithms. The balanced accuracy of these two cutoff thresholds along the cases was exhibited as of Table 2. The bold figures denoted the best performance. In most of cases, the VIP-BCT outperforms the VIP-1. The superior magnitude of the VIP-BCT can be seen obviously when the Prel is low as of Figure 4 (a). Figure 4 (b) – (e) show the average balanced accuracy of the two cutoff thresholds according to the remaining parameters. All five figures confirm again that the VIP-BCT cutoff threshold can beat the VIP-1 cutoff threshold.



(a) (b) (c) (d) (e)

Figure 4: Average balanced accuracy of the VIP-BCT and the VIP-1 according to each of four parameters, (a) Prel, (b) Mdif, (c) n and (d) $\mathbf{\Sigma}$

Figure 5 (a) was a plot of predictor variables of the VIP-1. The variables which values of VIP were greater than 1 were selected. Figure 5 (b) and (c) were plots of the VIP-BCT for the original predictor variables and noise variables, respectively. Its cutoff which was calculated from the VIP of noise variables as shown with red dash line in Figure 5 (c) was higher than the VIP-1. As of this result, the VIP-BCT cutoff threshold is more selective. Note that the VIP

			$\boldsymbol{\Sigma}$									
Prel	**Mdif**	n	$\boldsymbol{\Sigma}_1$		$\boldsymbol{\Sigma}_2$		$\boldsymbol{\Sigma}_3$		$\boldsymbol{\Sigma}_4$		$\boldsymbol{\Sigma}_5$	
			VIP-1	**VIP-BCT**	**VIP-1**	**VIP-BCT**	**VIP-1**	**VIP- BCT**	**VIP-1**	**VIP- BCT**	**VIP-1**	**VIP- BCT**
		40	0.8396	**0.8519**	0.8396	**0.8498**	0.8424	**0.8564**	0.8374	**0.8528**	0.8413	**0.8557**
	1	70	0.8540	**0.9709**	0.8546	**0.9723**	0.8561	**0.9773**	0.8541	**0.9676**	0.8551	**0.9690**
		100	0.8615	**0.9890**	0.8617	**0.9880**	0.8637	**0.9904**	0.8612	**0.9888**	0.8624	**0.9904**
		40	0.8672	**0.9929**	0.8670	**0.9928**	0.8673	**0.9927**	0.8665	**0.9929**	0.8670	**0.9925**
1%	3	70	0.8862	**0.9919**	0.8858	**0.9915**	0.8859	**0.9911**	0.8864	**0.9918**	0.8864	**0.9913**
		100	0.9014	**0.9914**	0.9018	**0.9908**	0.9004	**0.9904**	0.9016	**0.9915**	0.9015	**0.9908**
		40	0.8730	**0.9927**	0.8726	**0.9929**	0.8728	**0.9930**	0.8727	**0.9930**	0.8731	**0.9927**
	5	70	0.8947	**0.9919**	0.8947	**0.9918**	0.8944	**0.9916**	0.8946	**0.9919**	0.8948	**0.9917**
		100	0.9117	**0.9913**	0.9121	**0.9910**	0.9123	**0.9911**	0.9125	**0.9916**	0.9123	**0.9910**
		40	0.8533	**0.8538**	**0.8594**	0.8577	0.8559	**0.8823**	**0.8531**	0.8526	**0.8571**	0.8558
	1	70	0.8801	**0.9683**	0.8807	**0.9770**	0.8844	**0.9837**	0.8802	**0.9696**	0.8803	**0.9701**
		100	0.8951	**0.9893**	0.8930	**0.9889**	0.8958	**0.9891**	0.8948	**0.9885**	0.8929	**0.9891**
		40	0.9079	**0.9930**	0.9078	**0.9918**	0.9067	**0.9916**	0.9083	**0.9930**	0.9078	**0.9925**
3%	3	70	0.9393	**0.9922**	0.9376	**0.9902**	0.9347	**0.9897**	0.9397	**0.9917**	0.9389	**0.9908**
		100	0.9591	**0.9914**	0.9555	**0.9893**	0.9494	**0.9897**	0.9598	**0.9915**	0.9585	**0.9901**
		40	0.9197	**0.9927**	0.9189	**0.9925**	0.9189	**0.9921**	0.9196	**0.9930**	0.9189	**0.9927**
	5	70	0.9517	**0.9918**	0.9513	**0.9907**	0.9506	**0.9904**	0.9512	**0.9921**	0.9515	**0.9914**
		100	0.9701	**0.9917**	0.9698	**0.9897**	0.9683	**0.9896**	0.9702	**0.9917**	0.9702	**0.9905**
		40	**0.8655**	0.8515	0.8776	**0.8896**	0.8735	**0.9170**	**0.8651**	0.8544	**0.8690**	0.8640
	1	70	0.8998	**0.9671**	0.8972	**0.9858**	0.8983	**0.9846**	0.9001	**0.9702**	0.8985	**0.9703**
		100	0.9196	**0.9892**	0.9072	**0.9896**	0.9050	**0.9895**	0.9191	**0.9888**	0.9145	**0.9879**
		40	0.9351	**0.9928**	0.9343	**0.9911**	0.9302	**0.9908**	0.9351	**0.9926**	0.9354	**0.9919**
5%	3	70	0.9668	**0.9920**	0.9633	**0.9901**	0.9540	**0.9900**	0.9669	**0.9917**	0.9663	**0.9905**
		100	0.9822	**0.9918**	0.9757	**0.9900**	0.9630	**0.9912**	0.9824	**0.9912**	0.9817	**0.9895**
		40	0.9477	**0.9930**	0.9473	**0.9920**	0.9466	**0.9913**	0.9478	**0.9928**	0.9477	**0.9926**
	5	70	0.9768	**0.9924**	0.9761	**0.9902**	0.9742	**0.9895**	0.9770	**0.9922**	0.9766	**0.9913**
		100	0.9894	**0.9918**	0.9886	**0.9892**	0.9861	**0.9898**	0.9894	**0.9918**	0.9893	**0.9906**
		40	**0.8882**	0.8529	0.8973	**0.9086**	0.8959	**0.9632**	**0.8877**	0.8526	**0.8907**	0.8630
	1	70	0.9347	**0.9683**	0.9088	**0.9896**	0.8990	**0.9881**	0.9347	**0.9699**	0.9322	**0.9716**
		100	0.9571	**0.9888**	0.9094	**0.9908**	0.8938	**0.9892**	0.9562	**0.9881**	0.9490	**0.9880**
10%		40	0.9726	**0.9930**	0.9662	**0.9904**	0.9541	**0.9913**	0.9721	**0.9927**	0.9721	**0.9918**
	3	70	0.9922	**0.9923**	0.9827	**0.9916**	0.9638	**0.9922**	**0.9921**	0.9919	**0.9920**	0.9901
		100	**0.9976**	0.9920	0.9867	**0.9926**	0.9669	**0.9924**	**0.9976**	0.9911	**0.9976**	0.9896
		40	0.9816	**0.9929**	0.9808	**0.9906**	0.9783	**0.9908**	0.9818	**0.9930**	0.9818	**0.9925**
	5	70	**0.9960**	0.9924	**0.9951**	0.9899	**0.9921**	0.9909	**0.9962**	0.9924	**0.9961**	0.9910
		100	**0.9991**	0.9918	**0.9985**	0.9909	**0.9953**	0.9923	**0.9991**	0.9917	**0.9991**	0.9902

Table 2: Balanced Accuracy of the VIP-1 and the VIP-BCT

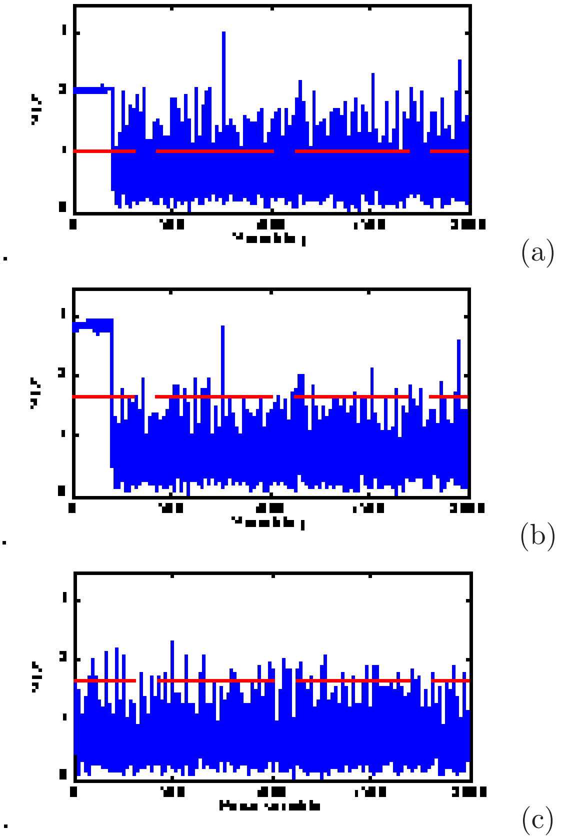


Figure 5: Plot of VIP from the case of Prel= 10%, Mdif= 1,n = 70 and $\mathbf{\Sigma} = \mathbf{\Sigma}_3$ for (a) only original predictors, (b) original predictors resulted from computing with noise variable and (c) noise variables.

scores of the original variables obtained from the VIP-BCT and the VIP-1 were not equal as seeing in Figure 5 (a) and (b) because the VIP included variable dependency. Therefore, the VIP of the original variables between analyzing of adding noise variables and without noise variable was different.

The cutoff threshold of the VIP-BCT was greater than 1 for all the cases but they tended to decrease when the Prel, Mdif and $n$were increasing. This result was corresponding to [3] in parameter of the Prel. That is, when the Prel was low the proper cutoff value was required to be greater than one. As of this reason, the VIP-BCT cutoff threshold certainly outperformed the other when the Prel was low. The average cutoff of the VIP-BCT cutoff threshold along the cases was displayed as of Table 3.

5. Conclusions

For this study, 180 situations were conducted and then compared the cutoff threshold of VIP between the new VIP-BCT and the traditional VIP-1. Experiment was designed by simulating four parameters: Prel, Mdif,$\mathbf{\Sigma}$ and n. The results demonstrate that in most of cases, the VIP-BCT delivered balanced accuracy better than the VIP-1 also it outstandingly performs in identifying relevant variables and outperforms the other. Appropriate cutoff values of VIP should be different depending on data structure. Their cutoff values of VIP need to be greater than 1 especially when the Prel, Mdif and n are low also they seem to be increasing when the three parameters decrease. There are various measurements for ranking the importance of variable. Thus, there are not usually explicit rule for estimating a suitable number of variables for those measurements. The BCT can be applied to be the cutoff threshold for any measurements and then the results of that application should be studied.

References

[1] W.J. Krzanowski, D.J. Hand, A simple method for screening variables before clustering microarray data, *Computational Statistics and Data Analysis*, **53** (2009), 2747-2753.

[2] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, *Bioinformatics*, **23(19)** (2007), 2507–2517.

[3] I.-G. Chong, C.-H. Jun, Performance of some variable selection methods when multicollinearity is present, *Chemometrics and Intelligent Laboratory Systems*, **78** (2005), 103-112.

[4] R. Gosselin, D. Rodrigue, C. Duchesne, A bootstrap-VIP approach for selecting wavelength intervals in spectral imaging application, *Chemometrics and Intelligent Laboratory Systems*, **100** (2010), 12-21.

[5] A. Lazraq, R. Cléroux, J.-P. Gauchi, Selecting both latent and explanatory variables in the PLS1 regression model, *Chemometrics and Intelligent Laboratory Systems*, **66** (2003), 117-126.

Prel	**Mdif**	$\boldsymbol{\Sigma}_1$			$\boldsymbol{\Sigma}_2$			$\boldsymbol{\Sigma}_3$			$\boldsymbol{\Sigma}_4$			$\boldsymbol{\Sigma}_5$		
		n = 40	n = 70	n = 100	n = 40	n = 70	n = 100	n = 40	n = 70	n = 100	n = 40	n = 70	n = 100	n = 40	n = 70	n = 100
	1	2.36	2.30	2.24	2.35	2.27	2.21	2.34	2.26	2.19	2.35	2.30	2.23	2.35	2.28	2.21
1%	3	2.26	2.13	2.04	2.25	2.12	2.00	2.24	2.10	1.98	2.26	2.13	2.03	2.25	2.11	2.01
	5	2.22	2.09	1.98	2.22	2.09	1.97	2.21	2.07	1.95	2.22	2.10	1.99	2.22	2.08	1.96
	1	2.27	2.17	2.07	2.23	2.06	1.95	2.20	2.02	1.90	2.28	2.15	2.05	2.25	2.10	1.99
3%	3	2.03	1.82	1.67	1.99	1.75	1.59	1.96	1.73	1.58	2.03	1.82	1.66	2.01	1.77	1.61
	5	1.97	1.74	1.58	1.95	1.70	1.52	1.93	1.68	1.51	1.97	1.74	1.58	1.96	1.72	1.55
	1	2.20	2.05	1.94	2.10	1.91	1.79	2.07	1.86	1.75	2.19	2.03	1.91	2.15	1.97	1.83
5%	3	1.87	1.62	1.45	1.80	1.53	1.39	1.76	1.54	1.43	1.86	1.61	1.44	1.83	1.57	1.39
	5	1.78	1.52	1.36	1.74	1.46	1.29	1.71	1.45	1.30	1.78	1.52	1.35	1.77	1.49	1.32
	1	2.05	1.84	1.68	1.88	1.71	1.64	1.85	1.70	1.66	2.04	1.82	1.66	1.98	1.74	1.60
10%	3	1.58	1.31	1.14	1.49	1.30	1.20	1.51	1.36	1.30	1.58	1.30	1.13	1.55	1.26	1.10
	5	1.48	1.21	1.05	1.41	1.16	1.04	1.41	1.19	1.09	1.48	1.20	1.04	1.46	1.19	1.02

Table 3: The average cutoff threshold of the VIP-BCT

[6] Y. Chen, *Statistical approaches for detection of relevant genes and pathway in analysis of gene expression data*, a Dissertation, University of California, USA, (2008).

[7] X.-M. Sun, X.-P. Yu, Y. Liu, L. Xu, D.-L. Di, Combining bootstrap and uninformative variable elimination: Chemometric identification of metabonomic biomarkers by nonparametric analysis of discriminant partial least squares, *Chemometrics and Intelligent Laboratory Systems*, **115** (2012), 37-43.

[8] A. Boulesteix, K. Strimmer, Partial least squares: a versatile tool for the analysis of high-dimensional genomic data, *Briefings in Bioinformatics*, **8(1)** (2006), 32-44.

[9] R. Manne, Analysis of two partial-least-squares algorithms for multivariate calibration, *Chemometrics and Intelligent Laboratory Systems*, **2** (1987), 187-197.

[10] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in partial least squares regression, *Chemometrics and Intelligent Laboratory Systems*, **118** (2012), 62-69.

[11] R. Rosipal, N. Krämer, Overview and recent advances in partial least squares, *Lecture Notes in Computer Science*, **3940** (2006), 34–51.

[12] G.-Z. Li, X.Q. Zeng, Feature selection for partial least square based dimension reduction, *Stud. Comput. Intell.*, **205** (2009), 3–37.

[13] D. Nguyen, D.M. Rocke, Tumor classi?cation by partial least squares using microarray gene expression data, *Bioinformatics*, **18(9)** (2002a), 39–50.

[14] D. Nguyen, D.M. Rocke, Multi-class cancer classification via partial least squares with gene expression profiles, *Bioinformatics*, **18(9)** (2002b), 1216–1226.

[15] J.J. Dai, L. Lieu, D. Rocke, Dimension reduction for classification with gene expression microarray data, *Statistical Applications in Genetics and Molecular Biology*, **5(1)** (2006), Article 6.

[16] J.-H. Cho, D. Lee, J.H. Park, K. Kim, I.-B. Lee, Optimal approach for classification of acute leukemia subtypes based on gene expression data, *Biotechnology Progress*, **18** (2002), 847-854.

[17] G. Ji, Z. Yang, W. You, PLS-based gene selection and identification of tumor-specific genes, *IEEE Transactions on Systems Man and Cybernetics-Part C: Applications and Reviews*, **41(6)** (2011), 830-841.

[18] M. Pérez-Enciso, M. Tenenhaus, Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach, *Human Genetics*, **112** (2003), 581–592.

[19] D. Johansson, P. Lindgren, A. Berglund, A multivariate approach applied to microarray data for identi?cation of genes with cell cycle-coupled transcription, *Bioinformatics*, **19(4)** (2003), 467-473.

[20] V. Centner, D.-L. Massart, Elimination of uninformative variables for multivariate calibration, *Analytical Chemistry*, **68(21)** (1996), 3851-3858.

[21] X.G. Shao, F. Wang, D. Chen, Q.D. Su, A method for near-infrared spectral calibration of complex plant samples with wavelet transform and elimination of uninformative variables, *Analytical Bioanalytical Chemistry*, **378(5)** (2004), 1382-1387.

[22] J. Moros, J. Kuligowski, G. Quintas, S. Garrigues, M. Guardia, New cut-off criterion for uninformative variable elimination in multivariate calibration of near-infrared spectra for the determination of heroin in illicit street drugs, *Analytica Chimica Acta*, **630** (2008), 150-160.

[23] E. Acuna, C. Rodriguez, A meta analysis study of outlier detection methods in classification, *Technical paper*, Department of Mathematics, University of Puerto Rico at Mayaguez, available at academic.uprm.edu/˜eacuna/paperout.pdf. In proceedings IPSI 2004, Venice (2004).

[24] L. Chen, *Ranking-based methods for gene selection in microarray data*, Thesis, Department of Computer Science and Engineering, University of South Florida, USA (2006).