An Efficient Estimation and Classification Methods for High Dimensional Data Using Robust Iteratively Reweighted SIMPLS Algorithm Based on *nu*-Support Vector Regression

ABDULLAH MOHAMMED RASHID¹, HABSHAH MIDI¹,³, WALIEED DHAN SLWABI²,⁴, AND JAYANTHI ARASAN³

¹Institute for Mathematical Research, Universiti Putra Malaysia (UPM), Serdang 43400, Malaysia
²Scientific Research Centre, Nawroz University (NZU), Duhok 42001, Iraq
³Department of Mathematics, Faculty of Science, Universiti Putra Malaysia (UPM), Serdang 43400, Malaysia
⁴Babylon Municipalities, Ministry of Construction, Housing, Municipalities and Public Works, Babylon 51001, Iraq

Corresponding author: Habshah Midi (habshah@upm.edu.my)

This work was supported in part by the Universiti Putra Malaysia Grant under Grant Geran Putra Berimpak (GPB) with project under Grant GPB/2018/9629700.

ABSTRACT The statistically inspired modification of the partial least squares (SIMPLS) is the most commonly used algorithm to solve a partial least squares regression problem when the number of explanatory variables (*p*) is larger than the sample size (*n*). Nonetheless, in the presence of irregular points (outliers), this method is no longer efficient. Therefore, the robust iteratively reweighted SIMPLS (RWSIMPLS), which is an improvement of the SIMPLS algorithm, is put forward to remedy this problem. However, the RWSIMPLS is still not very efficient with regard to its parameter estimations and outlier diagnostics. It also suffers from long computational times. This paper proposes a new robust SIMPLS that incorporates a new weight function constructed from *nu*-Support Vector Regression in its establishment. We call this method the robust iteratively reweighted SIMPLS based on *nu*-Support Vector Regression, denoted as SVR-RWSIMPLS. To avoid misclassification of observations, a new diagnostic plot is proposed to classify observations into regular observations, vertical outliers, good (GLPs) and bad leverage points (BLPs). The numerical results clearly indicate that the SVR-RWSIMPLS is more efficient, more robust and has less computational running times than the RWSIMPLS when multiple leverage points and vertical outliers exist. The proposed diagnostic plot is also very successful in classifying observations into correct groups.

INDEX TERMS Partial least square, robust regression, high dimensional data, support vector regression.

I. INTRODUCTION

As a solution to this problem, the Partial Least Squares Regression (PLSR) method is developed to build a regression model with multicollinear data where data are in high dimensional space. The main aim is to extract uncorrelated latent variables (components) using algorithms [1]. The Non-Linear Iterative Partial Least Squares (NIPALS) and the statistically inspired modification of the PLS (SIMPLS) [2] are very popular algorithms used to iteratively extract such components. When dealing with one or multivariate responses, the partial least squares regression is commonly referred to as PLS1 or PLS2, respectively. This article focuses only on PLS1 regression and SIMPLS.

It is now evident that the SIMPLS algorithm is very sensitive to outliers due to the use of the OLS estimation...
technique and non-robust covariance matrix in acquiring the components. Several robust PLSR algorithms have been proposed to deal with this problem. Wakeling and Macfie [3] were the first to introduce the algorithm by replacing univariate OLS regression steps with the robust alternatives; however, their approach received criticism as they employed non-robust initial weight. The iteratively reweighted algorithm by Cummins and Andrews [4] also received criticism because it was not able to handle outlying observations in the X-direction, popularly known as high leverage points (HLPs) or simply call leverage points. Then Hubert and Branden [5] proposed a robust covariance matrix into the SIMPLS algorithm, followed by Serneels et al. [6] who developed a partial robust M-regression estimator (PRM). Alin and Agostinelli [7] developed the Robust Iteratively Reweighted SIMPLS (RWSIMPLS) by employing a weight function introduced by Markatou et al. [8] that relies on the response variable, the chosen model distribution and the sample empirical distribution. They found that having longer computational times, the RWSIMPLS outperformed the SIMPLS and PRM in terms of having the smallest MSEs, the smallest empirical influence function values and highest empirical breakdown values. However, using this weight function has shortcomings as there is no specific rule to indicate which observations are the outliers. Moreover, their diagnostic plots to classify observations into four categories—regular observations, vertical outliers (outlying observations in Y-space), good leverage points (outlying observations in X space where they follow the pattern of the rest of the data) and bad leverage points (outlying observations in both X-space and Y-space) were constructed by plotting the robust standardized residuals versus the leverage values. This turned out to be not a good approach since it is now evident that leverage values are not very successful in identifying HLPs [9].

Motivated by the fact that employing the weighting function in the RWSIMPLS algorithm has shown to be more efficient than the SIMPLS and PRM, our primary aim is to perform some modifications to the existing RWSIMPLS algorithm by integrating a weight function based on our new proposed nu-Support Vector Regression for the detection of HLPs [10] to produce more efficient estimates with less computing times. Our secondary aim is to propose a better approach of classifying observations into the four categories mentioned above.

The paper is organized as follows. The Robust Iteratively Reweighted SIMPLS Algorithm (RWSIMPLS) is briefly discussed in Section II, followed by the discussion on our proposed method in Section III. Section IV presents the results of a simulation study and two real datasets, followed by the conclusion of the study in Section V.

II. THE ROBUST ITERATIVELY REWEIGHTED SIMPLS ALGORITHM (RWSIMPLS)

Several types of PLSR algorithms have been introduced recently, with the main aim to build a multivariate (multivariate) linear regression model,

$$Y = X\beta + \varepsilon_i$$  \hspace{0.5cm} (1)

where the response variables $Y = (y_1, y_2, \ldots, y_m)$ is $n \times m$ matrix of response variables; $X = (x_1, x_2, \ldots, x_p)$ is $n \times p$ matrix of $p$ predictors, $\beta = (\beta_1, \ldots, \beta_m)$ is $(p+1)\times m$ matrix; and $\varepsilon_i = (\varepsilon_{1i}, \ldots, \varepsilon_{mi})$ is $n \times m$ matrix of random errors. Hence, the general multivariate model for PLSR algorithm is as shown below,

$$X = TK^t + E$$ \hspace{0.5cm} (2)

$$Y = UQ^t + F,$$ \hspace{0.5cm} (3)

where the $n \times A$ component score matrices of $X$ and $Y$ are represented by $T = XW = (t_1, \ldots, t_A)$ and $U = YC = (u_1, \ldots, u_A)$, respectively whereby the number of components denoted as $A$ is smaller than or equal to $p$. The $p \times A$ and $m \times A$ loading matrices of $X$ and $Y$ are represented by $K = (k_1, \ldots, k_A)$ and $Q = (q_1, \ldots, q_A)$, respectively. The $p \times A$ matrix $W$ and the $m \times A$ matrix $C$ are the weight matrices. The aim of PLSR is to build a relation between $X$ and $Y$ via uncorrelated latent variables that can be extracted such that the components of response and predictor variables will have maximum covariance. Hence, adding a constraint which is to have score vectors with length $1$ will produce a unique maximization. The initial score vectors $t_1$ and $u_1$ are the solutions to the maximization problem (4) under the constraints $||t|| = ||u|| = 1$.

$$\max Cov(t, u) = \max Cov(X_w, Y_c)$$ $$t, u \in \mathbb{W}_c$$

$$= \max t'w = \max (X_w)'Y_c$$ $$w \in \mathbb{W}_c$$

$$= \max W'X'IY_c$$ $$w \in \mathbb{W}_c$$ \hspace{0.5cm} (4)

The singular value decomposition can be used to obtain the weight vectors $w$ and $c$. Hence, the maximum covariance matrix of Equation (4) can be reached through vectors $w1$ and $c1$ corresponding to the largest singular value of $X$ and $Y$ [11]. Additional constraints need to be added to obtain score vectors. The orthogonal of score vectors is usually considered as

$$t_\alpha f_j = 0$$ and $$u_\alpha u_j = 0$$

for $1 \leq \alpha < j \leq A$.

The empirical covariance matrix of the mean centered data that is highly sensitive to the outer points is deflated at the outset of SIMPLS.

The regression coefficient matrices of $X$ and $Y$ on their corresponding score matrices are represented by the loadings of $K$ and $Q$, respectively. All $A$ components are iteratively extracted by utilizing the OLS estimation technique and non-robust covariance matrix which make this algorithm very sensitive to outliers and less effective with non-normal errors.

To rectify this problem, Alin and Agostinelli [7] established the Robust Iteratively Reweighted SIMPLS (RWSIMPLS) by utilizing weight function constructed by Markatou et al. [8]. The weight is integrated in the algorithm of RWSIMPLS with the main aim of reducing influential observations. The weight function relies on the response variable, the chosen
model distribution, and the sample empirical distribution and is defined as follows:

\[ w_{r_i} = w \left( r_{j} (\hat{\beta}) ; m (; \delta) , \hat{P}_w \right) \]

\[ = \min \left\{ 1 , \frac{[ A(\delta (r_{j} (\hat{\beta}) ) + 1]^+}{\delta (r_{j} (\hat{\beta}) ) + 1} \right\} \] (5)

where \( r_{j} (\hat{\beta}) \) is the ith residual and \( A(\delta) \) is a residual adjustment function. The weight takes a value between 0 and 1 in [0,1]. [8], [12] developed a weighted likelihood estimation based on those weights. They noted that weight value equal to 1 corresponds to the maximum likelihood estimates, represented by \( A(\delta) = \delta \). Moreover, \( A(\delta) = 2(\delta + 1)1/2 - 1 \) corresponds to the Hellinger distance estimate that is more constant than the maximum likelihood estimates if certain assumptions fail. Alin and Agostinelli [7] utilized the Hellinger residual adjustment function to calculate the weights. In the first step of the RWSIMPLS algorithm, the median is employed to center the variables \( X \) and \( Y \). The RWSIMPLS algorithm is based on deflating the weighted covariance matrix \( S^w = (X^w)^T Y^w \), where \( X^w \) and \( Y^w \) are the weighted matrices obtained by multiply every row by the square root of

\[ w_{r_i} = \text{median} \{ w_{r_i} \} \]

for \( s = 1, 2, \ldots, m; \ i = 1, 2, \ldots, n \)

(6)

where the \( w_{r_i} \) is the weight of the ith residual corresponding to the sth response. Please refer to Alin and Agostinelli [7] for details on the RWSIMPLS algorithm.

III. THE PROPOSED SVR-RWSIMPLS ESTIMATOR BASED ON nu-SVR

A. CHOICE OF WEIGHT FOR THE PROPOSED METHOD

In regression, outliers can be classified into vertical outliers, residual outliers, and leverage points [9]. Vertical outliers are observations that are outlying in the \( Y \)-space while residual outliers are observations that have large residuals. Outlying observations in the \( X \)-space is referred to as leverage points and it can be classified further into good and bad leverage points. Among the three types of outliers, leverage point has the most detrimental effect on the computed values of various estimates which leads to misleading conclusion about the fitted regression model [13]. Hence robust methods that are able to reduce/eliminate the effect of leverage points should be used as an alternative to least squares method. Coakley and Hettmansperger [14] proposed a new class of Generalized-M estimator (GM6) to reduce the effect of multiple leverage points in regression model. The shortcoming of this method is that its efficiency tends to decrease as the number of ‘good’ leverage points increases. Dhhan et al. [15] improved the efficiency of GM6 estimator by developing a GM estimator based on fixed parameters \( \varepsilon \)-tube support vector regression (FP-\( \varepsilon \)-SVR) that is highly efficient and has good resistance against multiple leverage points. Other robust methods that able to eliminate the effect of multiple leverage points with high robustness and computing efficiency when applied to energy system data, can be found in [16], [17].

The RWSIMPLS algorithm employed weight with the main aim of reducing influential observations (IOs). According to Belsley et al. [18], IOs are those observations which are either done alone or together with several other observations, and have an adverse effect on the computed values of various estimates. In this connection, vertical outliers and bad leverage points can be considered as IOs. As noted [9], [19], [20], good leverage points are not considered as IOs because they follow the pattern of the majority of the data. They may contribute to precise estimates since they have lesser or no impact on OLS estimates. As such, to get efficient estimates, only BLPs and VOs, not GLPs, are down weighted [15]. Hence, an efficient weight function should be formulated so that only genuine IOs will be assigned with smaller weight and GLPs is assigned with weight 1. The RWSIMPLS algorithm which employs the weight function has improved its performance compared to the SIMPLS and PRM algorithm. Nonetheless, the use of weighting function [8] in Equation (6) has its shortcomings due to the absence of specific rules to indicate which observations are the IOs. In view of this, we propose the weight function based on an efficient diagnostic method of identifying bad leverage points and vertical outliers in high dimensional data. A good weight function is one that follows on the efficient diagnostic method that is able to correctly identify VOs and BLPs.

Firstly, we developed a diagnostic robust method to identify outliers (vertical and bad leverage points) based on nu-SVR. Then we minimized their effects by assigning lower weights and incorporating this weighting function in our new proposed RWSIMPLS algorithm which is denoted as SVR-RWSIMPLS. The SVR is non-parametric approach that follows a statistical learning theory and the concept of structural risk reduction. Dhhan et al. [21] developed fixed parameters \( \varepsilon \)-tube SVR (FP-\( \varepsilon \)-SVR) to identify outliers (vertical and bad leverage points). However, this method was not very successful in detecting outliers. Motivated by their work, Abdullah et al. [10] developed nu-SVR which proved to be highly effective in identifying multiple outliers (VOs and BLPs) under various settings. The proposed method has managed to identify the actual number of outliers in a data set by taking control of free parameters \( C = 200, \ nu = 0.22 \) and the kernel parameter \( h = 0.05 \).

In the \( \nu \)-tube-SVR, the non-sparseness property of the \( \varepsilon \)-tube-loss function is used to detect abnormal points in data. The \( \nu \)-tube loss function is defined as follows:

\[ L_{\nu} = \begin{cases} 0 & \text{if } |y_i - f(x)| \leq \nu \\ |y_i - f(x)| - \nu & \text{otherwise}. \end{cases} \]

As noted by [22], [23], the convex optimization problem can be rephrased as:

\[ \minimize \frac{1}{2} ||w||^2 + C(\nu + 1/2 \sum_{i=1}^{l} (\xi_i + \xi_i^*) \]

\[ \text{subject to:} \quad y_i - (x^T w + b) + \xi_i - \xi_i^* \leq \nu, \quad \xi_i, \xi_i^* \geq 0 \]

\[ \sum_{i=1}^{l} \xi_i = \nu \]

\[ \sum_{i=1}^{l} \xi_i^* = \nu \]

\[ w, b, \xi_i, \xi_i^* \in \mathbb{R} \]

\[ \text{for } i = 1, 2, \ldots, l \]

\[ \text{where } l \text{ is the number of training samples.} \]
Equation (7) can be written as the next convex quadratic problem

\[
\begin{aligned}
\text{subject to } & f(x) - y_i; \leq nu + \xi_i \\
& y_i - f(x); \leq nu + \xi^* \\
& \xi_i, \xi^*_i \geq 0, \quad nu \geq 0.
\end{aligned}
\] (7)

The weight vector and final regression function are represented as follows:

\[
w = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) \Phi(x_i)
\]

\[
Z_i = f(x) = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) k(x_i, x) + b.
\]

\[
0 \leq \alpha_i^*, \alpha_i \leq C, i = 1, 2, \ldots, l,
\] (8)

where \(\alpha_i^*\) and \(\alpha_i\) are the Lagrange multipliers; \(K(x_i, x)\) is the kernel function; and \(b\) is the bias term of the SVR model. Controlling the values of the parameters of the nu-SVR approach will reduce the effect of outliers in the solution as mentioned by Rojo-Alvarez et al. [24]. Moreover, the highest Lagrange multipliers (\(\alpha_i^*\) and \(\alpha_i\)) belong to the outlying point in the data set which is considered an outlier. As a result, the values of Lagrange multipliers based on the cost parameter value \(C\) will be affected by any change in the cost value. The non-parametric approach is advantageous due to its flexibility which allows for a significant change in the estimation value corresponding to the abnormal point with a small change to the estimation values corresponding to the rest of the data.

To be precise, the parameter \(nu\) works automatically to minimize the value of the \(\epsilon\) parameter and controls the number of the support vector that is allowed to use all observations. In return, it assists in diagnosing the outliers from the first iteration. The cost parameter \(C\) produces a high variance between Lagrange multipliers (\(\alpha_i^*\) and \(\alpha_i\)), implying a significant value of \(Z\) belonging to outliers which can be identified using a specific cut-off point criterion. Since the distribution of \(Z_i\) is intractable, following Habshah et al. [13], we propose the following cut-off point (CP);

\[
CP = \text{median } (Z_i) + 3 \times \text{MAD } (Z_i)
\] (9)

where

\[
\text{MAD} = \text{median } \{|Z_i - \text{median}(Z_i)|\} / d
\]

The weighted standard deviations can be considered as the robust alternative to the ordinary standard deviations because the effect of vertical outliers and bad leverage points have

\[
w_i = \min \left[ 1, \frac{CP}{Z_i} \right]
\] (10)

where weight equals 1 is given to regular and good leverage points, and weight equals to \(CP / Z_i\) is given to vertical outliers and bad leverage points.

**B. ALGORITHM OF THE PROPOSED SVR-RWSIMPLS ESTIMATOR**

This section describes how the new version of the RWSIMPLS which is based on nu-SVR, denoted as SVR-RWSIMPLS, is developed. The establishment of the SVR-RWSIMPLS is done by integrating the weighting function obtained from nu-SVR which requires less computing times than the RWSIMPLS. Moreover, at the outset, the matrices of \(X\) and \(Y\) are centered based on Decile Mean (DM) instead of the median which has low efficiency at normal distribution. Rana et al. [26] showed that the accuracy of decile mean is better than the median as it has a smaller bias and smaller standard errors in a variety of situations. The centered variables are scaled as follows:

\[
\tilde{x}_{ij} = \frac{x_{ij} - \text{DM}(x)}{\text{wsdx}}, \quad j = 1, \ldots, p
\]

\[
\tilde{y}_{is} = \frac{y_{is} - \text{DM}(y)}{\text{wsdy}}, \quad s = 1, \ldots, n
\] (11)

where \(\text{DM}\) is the decile mean where it is defined as the sum of all deciles (\(D_i\)) divided by the number of deciles,

\[
\text{DM} = \frac{D_1 + D_2 + \ldots + D_9}{9}
\]

and \(\text{wsdx}\) and \(\text{wsdy}\) are weighted standard deviations for variables \(X\) and \(Y\) based on weight of nu-SVR as given in Equation (10).
been reduced by assigning smaller weights to them. The steps for our proposed algorithm can be summarized as follows:

**Step 1:** Calculate the initial weight based on nu-SVR as in Equation (10).

**Step 2:** Both matrices of the predictor and response variables are scaled as in Equation (11).

**Step 3:** Using the weight in Step (2), compute the weighted predictors variables, $X^w$ and response variable $Y^w$ by multiplying each row by the squared root of weight as in the following equations:

$$x^w_{ij} = \tilde{x}_{ij}\sqrt{w_{ij}}, \quad j = 1, \ldots, p$$

$$y^w_{is} = \tilde{y}_{is}\sqrt{w_{is}}, \quad s = 1, \ldots, n$$

**Step 4:** Choose a small sample of size 5 without replacement, for example, $X^s_w$ and $Y^s_w$ from $X^w$ and $Y^w$, then obtain the starting coefficient estimates $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_m)$ using the ordinary steps of SIMPLS on $X^s_w$ and $Y^s_w$.

**Step 5:** Compute $r_{is} = y_{is} - x^w_{ij}\hat{\beta}_i$, where $\hat{\beta}_i = (\hat{\beta}_{i0}, \ldots, \hat{\beta}_{ik})$ and $x^w_{ij} = (x^w_{11}, \ldots, x^w_{nn})$ for $s = 1, \ldots, m$ and $i = 1, \ldots, n$.

**Step 6:** Utilize the robust scale for residuals as follows:

$$r_{is} = \frac{r_{is} - DM(r)}{wsdr}$$

where $DM$ and $wsdr$ are the decile mean and weighted standard deviations of residuals, respectively, calculated using Equation (11).

**Step 7:** Compute the weighted covariance matrix $S^w = X^w Y^w$.

**Step 8:** Employ the SIMPLS method on $S^w$ to estimate the parameters.

**Step 9:** Repeat Steps 5-8 until convergence, i.e., when the maximum of the absolute difference for 2 consecutive $\hat{\beta}_i$ is less than 0.00001.

**Step 10:** Steps 3-9 are repeated several times, for example, for $c = 5$ times. Then the final parameter estimates $\hat{\beta}$, which corresponds to the minimum value 5 of the coefficients set.

In the next section, the performance of the proposed SVR-RWSIMPLS is assessed using a simulation study and real data set.

### IV. RESULTS

This section includes the Monte Carlo simulation study and real data to evaluate the performance of the SVR-SIMPLS method compared to the SIMPLS and RWSIMPLS methods. All analyses were done using R program [27].

### A. MONTE CARLO SIMULATION STUDY

#### 1) STANDARD NORMAL AND SKEWED ERRORS DISTRIBUTION

To investigate the efficiency of the estimators, we compared the proposed method with the SIMPLS and the recently proposed RWSIMPLS method [7]. We followed the simulation design of [6], [7] by considering the linear regression model as in Equation (1). The true value of the regression parameters is randomly drawn from a normal distribution with $\mu = 0$ and $\sigma = 0.001$. The explanatory variable matrix $X$ is computed by $X = TK$ so that the rank of $X$ equals the number of components $A$, which is the number of columns of $T$. Columns $T$ and $K$ are generated from the standardized normal distribution, $k_a \sim N(0, 1)$ and $t_a \sim N(0, 1)$, where $a$ is the number of components, $a = 1, \ldots, A$. Moreover, this simulation study contains different distributions of errors $\epsilon_i$: the standard normal, Student-t distribution with 2 degrees of freedom ($t_2$) and 5 degrees of freedom ($t_5$), and Cauchy and Slash distributions (heavy-tailed distributions). The performance of the proposed method was evaluated using MSE. The mean square errors for the $\hat{\beta}$ estimates were computed as shown in Equation (14):

$$MSE = \frac{1}{S_{rep}} \sum_{j=1}^{S_{rep}} \left\| \hat{\beta}^j - \beta_0 \right\|^2$$

where $S_{rep}$ is the number of replications used in the simulation studies. This study considered different size of $p$ and $n$, and under each setting, we repeated the simulation 1000 times, i.e., $S_{rep} = 1000$ [7]. The results are presented in Table 1. It is interesting to observe that at normal errors, the SIMPLS, which is supposedly to be the most efficient estimator, performed less efficiently than the SVR-RWSIMPLS. In fact, it was found that the SVR-RWSIMPLS showed a better performance compared to the SIMPLS and RWSIMPLS as it indicated the least MSE values except for $n = 24$, $p = 6$ and $A = 2$. However, in this case, the MSE of SVR-RWSIMPLS is only slightly larger than the RWSIMPLS. The scenarios changed dramatically for SIMPLS when the errors were skewed, particularly its performance became worst under Cauchy and slash errors distributions where its MSE values increased significantly. The skewed error distributions reveal that the SVR-RWSIMPLS provides the best result as it shows the least values of MSE compared to the other two methods.

#### 2) CONTAMINATED DATA WITH HIGH LEVERAGE POINTS

The performance of our SVR-RWSIMPLS is further assessed by investigating its performance in the presence of outliers in both $X$ and $Y$ spaces (bad leverage points). The simulation design is similar to the previous design, except that we only considered standard normal errors distribution, $N(0, 1)$ with $p$ explanatory variables ($p = 10, p = 200$ and $p = 400$), various sample sizes ($n = 30, 50, 100, 150$, and $200$) and different percentages of contamination $\theta = (10\%, 15\%$, and $20\%)$. Following Dhhan et al. [15], at each step, a certain percentage of good $X$ and $Y$ values were deleted and replaced with large values equal 50. The SVR-RWSIMPLS, RWSIMPLS, and SIMPLS were then applied to the data. The results are exhibited in (Tables 2–4) and (Figures 1–5). As can be expected, the SIMPLS produced poor results in all scenarios. Let us first focus our attention to Table 2, for low dimensional data, $p = 10$. At 10% contamination, $n = 30k50$, the SVR-RWSIMPLS method has the smallest values of MSE, but its
TABLE 1. Mean square errors for regression parameter estimates.

| Methods       | Nominal model | Student-t | heavy-tailed |
|---------------|---------------|-----------|--------------|
|               | N (0,1)       | $t_3$     | $t_2$        | Cauchy       | Slash        |
| $n=24$        | SIMPLS        | 0.0336296 | 0.05335375   | 0.4402057    | 7826.661     | 302.091      |
| $p=6$         | RWSIMPLS      | 0.02389684| 0.03750914   | 0.03266952   | 0.265320     | 0.1778451    |
| $A=2$         | SVR-RWSIMPLS  | 0.02777849| 0.02822138   | 0.02627901   | 0.02732833   | 0.02854402   |
| $n=100$       | SIMPLS        | 0.00305636| 0.004998174  | 0.03632237   | 227.2379     | 264.437      |
| $p=10$        | RWSIMPLS      | 0.00434233| 0.00434268   | 0.00485336   | 0.024789     | 0.01329      |
| $A=2$         | SVR-RWSIMPLS  | 0.00282935| 0.002975546  | 0.003255414  | 0.003245335  | 0.002689746  |
| $n=15$        | SIMPLS        | 0.00331445| 0.005343418  | 0.03818306   | 1632.9658    | 370.4845     |
| $p=60$        | RWSIMPLS      | 0.00322959| 0.002330171  | 0.004129336  | 0.027896     | 0.01623299   |
| $A=2$         | SVR-RWSIMPLS  | 0.00278756| 0.003171543  | 0.003142835  | 0.003410749  | 0.004217177  |
| $n=10$        | SIMPLS        | 0.00354941| 0.005880699  | 0.1256945    | 40.00439     | 10.431394    |
| $p=100$       | RWSIMPLS      | 0.00225841| 0.003170725  | 0.0050803    | 0.035869     | 0.01811016   |
| $A=2$         | SVR-RWSIMPLS  | 0.00166381| 0.002990409  | 0.00305629   | 0.003662611  | 0.003178425  |

TABLE 2. Mean square errors for estimation when the number of explanatory variables $= 10$.

| Cont. | $n$ | SIMPLS        | RWSIMPLS      | SVR-RWSIMPLS |
|-------|-----|---------------|---------------|--------------|
| 10%   | 30  | 0.7639098     | 0.01533377    | 0.005365298  |
|       | 50  | 0.7718271     | 0.009040246   | 0.006771647  |
|       | 100 | 0.7690149     | 0.004813517   | 0.007143712  |
|       | 150 | 0.7564968     | 0.004212268   | 0.006809811  |
|       | 200 | 0.7664702     | 0.003137886   | 0.007816038  |
| 15%   | 30  | 0.7727444     | 0.02011477    | 0.005770184  |
|       | 50  | 0.8274196     | 0.01407867    | 0.005511995  |
|       | 100 | 0.6871687     | 0.008313192   | 0.006239471  |
|       | 150 | 0.7613761     | 0.007542461   | 0.006968241  |
|       | 200 | 0.6290731     | 0.006730203   | 0.004965005  |
| 20%   | 30  | 0.6273778     | 0.02941466    | 0.004181227  |
|       | 50  | 0.8274958     | 0.02273204    | 0.004082914  |
|       | 100 | 0.7481607     | 0.02450419    | 0.005003077  |
|       | 150 | 0.6309797     | 0.02782497    | 0.005855168  |
|       | 200 | 0.6241432     | 0.03146631    | 0.006428079  |

MSE values are slightly larger than the MSE for RWSIMPLS for $n = 100, 150, 200$. However, at 15% and 20% contaminations, the RWSIMPLS outperformed the SIMPLS and RWSIMPLS, irrespective of sample sizes, $n$. The performances can clearly be seen in Figure 1.

(Tables 3 and 4) and (Figures 2 and 4) illustrate the MSE of the three methods, along with the computation running times for high dimensional data ($p = 200$ and $p = 400$). It is interesting to observe that the SVR-RWSIMPLS consistently provides the smallest MSE and the least computational times, irrespective of the sample size and percentage of contaminations. The computation time for the SIMPLS was not compared because this method performed poorly as it produced the largest values of MSE. For a quick interpretation of the results, refer to (Figures 3 and 5).
TABLE 3. Mean square errors for estimation when the number of explanatory variables = 200.

| MSE         | Cont. | n   | SIMPLS       | RWSIMPLS     | Time by seconds | SVR-RWSIMPLS | Time by seconds |
|-------------|-------|-----|--------------|--------------|-----------------|--------------|-----------------|
|             | 10%   | 30  | 0.03290503   | 0.0008911415 | 1107.6          | 0.0002855855 | 513             |
|             |       | 50  | 0.03192774   | 0.000699333  | 2241.6          | 0.0002591454 | 540.6           |
|             |       | 100 | 0.03167063   | 0.0004637943 | 2740.2          | 0.0003371861 | 645.6           |
|             |       | 150 | 0.04132779   | 0.000423488  | 2977.2          | 0.0002975806 | 742.2           |
|             |       | 200 | 0.04179056   | 0.0003853905 | 4176            | 0.0003162723 | 726.6           |
|             | 15%   | 30  | 0.03179764   | 0.001225104  | 1059.6          | 0.0002176592 | 496.8           |
|             |       | 50  | 0.03259856   | 0.0009575926 | 1488.6          | 0.0002046799 | 529.8           |
|             |       | 100 | 0.03262583   | 0.0007924541 | 2317.8          | 0.0002390018 | 639.6           |
|             |       | 150 | 0.03211558   | 0.0008935601 | 3028.2          | 0.0002292943 | 759.6           |
|             |       | 200 | 0.0314189    | 0.0008845971 | 4536            | 0.0002324568 | 1031.4          |
|             | 20%   | 30  | 0.03208444   | 0.001756421  | 1008.6          | 0.0002120732 | 660.6           |
|             |       | 50  | 0.03167009   | 0.001583091  | 1228.8          | 0.00024465  | 703.2           |
|             |       | 100 | 0.03218679   | 0.001712099  | 2172.6          | 0.0002115406 | 729             |
|             |       | 150 | 0.03205535   | 0.002373762  | 3255.6          | 0.000240343 | 851.4           |
|             |       | 200 | 0.3301046    | 0.002596269  | 4356            | 0.0002439595 | 1062.6          |

TABLE 4. Mean square errors for estimation when the number of explanatory variables = 400.

| MSE         | Cont. | n   | SIMPLS       | RWSIMPLS     | Time by seconds | SVR-RWSIMPLS | Time by seconds |
|-------------|-------|-----|--------------|--------------|-----------------|--------------|-----------------|
|             | 10%   | 30  | 0.01629676   | 0.0007119017 | 3313.8          | 0.000391306 | 762             |
|             |       | 50  | 0.01592448   | 0.0005815551 | 4464            | 0.0004694295 | 904.2           |
|             |       | 100 | 0.01636355   | 0.0005719626 | 6192            | 0.0003785648 | 1095            |
|             |       | 150 | 0.01659864   | 0.0004985534 | 9288            | 0.0004685707 | 1354.8          |
|             |       | 200 | 0.01597498   | 0.0005139229 | 12852           | 0.0004186294 | 1815            |
|             | 15%   | 30  | 0.01620027   | 0.0009427852 | 3045.6          | 0.0003340352 | 941.4           |
|             |       | 50  | 0.01627427   | 0.0007958495 | 4068            | 0.0004418993 | 838.8           |
|             |       | 100 | 0.01625759   | 0.0006629426 | 6444            | 0.0004280874 | 1032            |
|             |       | 150 | 0.01617784   | 0.0007730875 | 9144            | 0.0004729997 | 1240.2          |
|             |       | 200 | 0.0162865    | 0.0007841005 | 12888           | 0.0003960245 | 1573.2          |
|             | 20%   | 30  | 0.01599191   | 0.001150069  | 3288.6          | 0.0004160888 | 766.2           |
|             |       | 50  | 0.01593849   | 0.00106874   | 4032            | 0.0003601853 | 775.8           |
|             |       | 100 | 0.01584584   | 0.001384029  | 6480            | 0.0004476544 | 1114.2          |
|             |       | 150 | 0.01633287   | 0.001488029  | 9684            | 0.0003646024 | 1370.4          |
|             |       | 200 | 0.01614497   | 0.001563681  | 13572           | 0.0004283372 | 1616.4          |

B. REAL EXAMPLES
1) OCTANE DATA SET
The octane data set is the first real example to illustrate the performance of our proposed SVR-RWSIMPLS compared to the commonly used SIMPLS and RWSIMPLS methods. This data set consists of near-infrared absorbance spectra with p = 226 wavelengths collected on n = 39 gasoline samples. Esbensen et al. [28] had also used this data set in their study. [29], [30] noted that this data set contains outliers in samples 25, 26, 36, 37, 38, and 39 which contain added ethanol. The SVR-RWSIMPLS, SIMPLS, and RWSIMPLS were then applied to the data set by considering number of
components $A = 2$, following the methods by [31], [7]. Based on Alin and Agostinelli [7], the empirical influence function and empirical breakdown values of the three methods were constructed. To compute the empirical influence function, the $Y$ variable was first scaled before being contaminated by replacing the first value of the $X$ variables ($x_{11} = -0.060193$) as well as the first value of $Y$ variable ($y_1 = -1.2271$) with the values between $-40$ and $40$. The empirical influence values were then computed using Equation (15)

$$
\left\| \hat{\beta}_i^c - \hat{\beta} \right\| \frac{1}{n} \tag{15}
$$

where $\hat{\beta}_i^c$ is the coefficient estimate for the contaminated data; $\hat{\beta}$ is the coefficient estimate for the clean data; and
A. M. Rashid et al.: Efficient Estimation and Classification Methods for High Dimensional Data

FIGURE 4. Mean square errors for estimation when the number of explanatory variables \( p = 400 \).

FIGURE 5. Computation times in seconds for SVR-RWSIMPLS and RWSIMPLS for simulated data sets with an increasing number of observations \( n \) and with \( p = 400 \).

\( n \) is the sample size. The empirical influence function versus the contamination values \( [y_1 = i, x_{11} = i], i = -40, -38, \ldots, 38, 40 \) is presented in Figure 6(a). Simultaneously changing \( y_1 \) and \( x_{11} \) in the data resulted in more impact on the SIMPLS coefficient estimates. However, all these three methods had the lowest empirical influence values when contaminated at \( (y_1 = 0, x_{11} = 0) \). It can be observed that there is large spike when the first value of \( y \) is contaminated with \(-4\), and this has a negative effect on the SIMPLS estimates. Even though the empirical influence values for RWSIMPLS are smaller than those of the SIMPLS, RWSIMPLS cannot outperform the SVR-RWSIMPLS. Besides, it can be observed that there are moderate fluctuations for RWSIMPLS estimates. On the other hand, the proposed SVR-RWSIMPLS was impressive as it has the lowest empirical influence values without showing any fluctuations in its estimates. The robustness of the three methods were investigated further by studying the empirical breakdown values.

To construct an empirical breakdown plot, the response variables of the octane data were contaminated up to 50\% by deleting the original \( y_i \) and replacing it with 20 for \( i = 1, 2, \ldots, 20 \) whereby the first value of \( y \) is first contaminated \( (\frac{2}{n} \ast 100\% \ Cont.) \), followed by the first 2 values of \( y \ (\frac{4}{n} \ast 100\% \ Cont.), \) etc. Then the difference between the contaminated coefficient estimates vector \( \hat{\beta}_c \) and the coefficient estimates vector was computed at each amount of contamination for a clean data \( \hat{\beta} \). The results are displayed in Figure 6(b). It can be seen that that the SIMPLS performs poorly as it cannot cope well even when there is only one outlier in the data. The results also signify that the RWSIMPLS can tolerate slightly over 33\% of contamination. The proposed SVR-RWSIMPLS method has the smallest empirical breakdown points, and it can stand up to 48\% of contamination.

The performance of the RWSIMPLS method was further assessed with regard to its detection of vertical outliers and leverage points. Alin and Agostinelli [7] constructed a diagnostic plot for identifying vertical outliers and good and bad leverage points by plotting the RWSIMPLS standardized residual against leverage values. They employed leverage values i.e. the diagonal elements of the hat matrix \( H = X(X'X)^{-1}X' \) to identify HLPs. Any observation which corresponds to a leverage value greater than the cut-off point, \( 2p/n \), is considered as HLPs. As pointed out by [13], leverage values are not successful in detecting multiple HLPs. As such, using leverage values for diagnosing HLPs will produce
erroneous results. Hence, a new diagnostic plot is proposed by plotting the standardized residuals of SVR-RWSIMPLS against the cut-off point \( Z_i \) as shown in Equation (9). Figure 7 presents the diagnostic plots of the Octane data set constructed based on the RWSIMPLS and SVR-RWSIMPLS. As seen in Figure 7(a), the SVR-RWSIMPLS method successfully diagnoses the actual number of outliers in the octane data set without showing any masking and swamping effects. Nevertheless, the RWSIMPLS method badly suffers from masking and swamping effects. Additionally, the computational running time for octane data is 3.51 seconds for SVR-RWSIMPLS method and 4.09 seconds for RWSIMPLS method. This indicates that the computational running times for SVR-RWSIMPLS is faster than the RWSIMPLS which is in agreement with the results of the simulation study.

2) GAS DATA SET

The second real example in this paper is the gas data set [32]. The gas data set has been studied by several researchers such as [31], [7]. According to Branden and Hubert [31], it is a clean data set which contains near-infrared spectra (NIR) of gasoline with sample size \( n = 60 \) with the number of explanatory variables, \( p = 409 \), which are the NIR spectra values measured in 2-nm intervals from 900 nm to 1700. The response variable was represented by the octane number with the values of between 83.4 and 89.6.

We followed the same process as described for the octane data in order to further assess the performance of our proposed method by constructing the empirical influence function, empirical breakdown values, and diagnostic plot. To compute the empirical influence function, the response variable \( Y \) needs to be scaled at the outset in order to obtain a...
similar range with the predictor variables. After that, the data were contaminated by concurrently substituting the first value of the first predictor variable ($x_{11} = -0.05019$) and the first value of the response variable ($y_1 = -1.2271$) with values between $-40$ and $40$. The empirical influence values were computed using Equation (13) and displayed in Figure 8(a).

FIGURE 8. Section (a) provides comparison for SIMPLS, RWSIMPLS, and SVR-RWSIMPLS for empirical influence function and Section (b) provides comparison for SIMPLS, RWSIMPLS, and SVR-RWSIMPLS for empirical breakdown for Gas data set.

FIGURE 9. Diagnostics for SVR-RWSIMPLS and RWSIMPLS methods on the gas data set Section (a-b) in case of clean data and Section (c-d) for contaminated data.
Similar to the octane data, it can be observed from Figure 8(a) that the SIMPLS method has higher values of empirical influence function as a result of the contamination, and it also has a high impact on the SIMPLS algorithm. When the data were contaminated with 0 \((x_{11} = 0, y_i = 0)\), all the three methods prove to be equally good with smallest empirical influence values. However, when the first y value was contaminated with a value close to -4, the SIMPLS produced very bad results with a very huge spike. The RWSIMPLS has smaller empirical values but indicates moderate instability in the parameter estimates as evident by the existence of small fluctuations in the empirical influence values. On the other hand, the proposed SVR-RWSIMPLS has the smallest empirical influence values with no significant fluctuations which shows the robustness of the SVR-RWSIMPLS estimates.

Next, the empirical breakdown values for each of the three methods were constructed for the gas data set. The response variables were contaminated by up to 50% by substituting the \(y_i\) with 20 for \(i = 1, 2, \ldots, 30\) that is, the first value of \(y_{i}^{(1)} = 100\%\ Cont.) is the first to be contaminated, followed by the contamination of the first 2 \((y_{i}^{(2)} = 100\%\ Cont.)\) and so on. For each amount of contamination, the difference between the contaminated coefficient estimate vector \(\beta_i\) and the coefficient estimate vector \(\hat{\beta}\) for clean data was computed, i.e. \(\|\hat{\beta} - \beta_i\|\). The results are displayed in Figure 8(b) which clearly show that the SIMPLS method is extremely sensitive even in the presence of one outlier in the response variable. The empirical breakdown values of SVR-RWSIMPLS are smaller than those of the RWSIMPLS; however, both can withstand the outliers by up to 48%.

Finally, the diagnostic plots based on the RWSIMPLS and SVR-RWSIMPLS were constructed for the clean gas data. The plots are shown in Figure 9(a-b). As seen in Figure 9(a), the plot based on the SVR-RWSIMPLS method does not show any outlier in the data set as highlighted by Branden and Hubert [31]. On the contrary, the plot based on the RWSIMPLS method presented in Figure 9(b) has shown many good observations that are declared as outliers (swamping effect) which will lead to a misleading conclusion. We then purposely contaminated the data by replacing the observations \(\text{(cases 1, 2, 3, 4, and 5)}\) in both \(X\) and \(Y\) directions (bad leverage points) with large values \((150, 151, 152, 153,\) and \(154)\). Furthermore, to create vertical outliers, two observations in the \(Y\) direction (Cases 6 and 7) were substituted with large values \((100\) and 105). The results are shown in Figures 9(c and d). A good method is the one that can successfully detect observations 1, 2, 3, 4, and 5 as bad leverage points and observations 6 and 7 as vertical outliers. It is interesting to note from Figure 9(c) that the diagnostic plot based on SVR-RWSIMPLS correctly diagnosed the abnormal points in \(X\) and \(Y\) spaces as bad leverage points as well as in \(Y\) space as vertical outliers. Nonetheless, the plot based on the RWSIMPLS method failed to correctly diagnose those observations. This diagnostic plot suffered from masking and swamping problems as shown in Figure 9(d). Furthermore, the computational running time for gas data for RWSIMPLS is 3.48 seconds and for SVR-RWSIMPLS is 3.36 seconds.

**V. CONCLUSION**

The aim of this study was to develop a robust iteratively reweighted of the SIMPLS algorithm (RWSIMPLS) based on weight obtained from our newly proposed nu-SVR (SVR-RWSIMPLS) in detecting outliers. Decile mean was employed instead of a median to center the response as well as the predictor variables. We also established a new diagnostic plot for classifying observations into regular observations, vertical outliers, and good and bad leverage points. The numerical results show that the SVR-RWSIMPLS is more efficient and more robust than the RWSIMPLS and SIMPLS when multiple leverage points and vertical outliers exist as it has the smallest MSE, the smallest empirical influence function and the smallest empirical breakdown values, and it can withstand up to 48% contamination. The computational running times for SVR-RWSIMPLS is also shown to be faster. The proposed diagnostic plot based on SVR-RWSIMPLS is also very successful in classifying observations into the correct categories. Nevertheless, the plot based on RWSIMPLS fails to correctly identify the actual outliers (multiple leverage points and vertical outliers). Hence, the newly developed SVR-RWSIMPLS method can be a suitable alternative to model multicollinear data.

**REFERENCES**

[1] S. Wold, H. Martens, and H. Wold, "The multivariate calibration problem in chemistry solved by the PLS method," in *Matrix Pencils*. Cham, Switzerland: Springer, 1983, pp. 286–293.

[2] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics Intell. Lab. Syst.*, vol. 18, no. 3, pp. 251–263, Mar. 1993.

[3] I. N. Wakelin and H. J. H. Macfarie, "A robust PLS procedure," *J. Chemometrics*, vol. 6, no. 4, pp. 189–198, Jul. 1992.

[4] D. J. Cummins and C. W. Andrews, "Iteratively reweighted partial least squares: A performance analysis by Monte Carlo simulation," *J. Chemometrics*, vol. 9, no. 6, pp. 489–507, Nov. 1995.

[5] M. Hubert and K. V. Branden, "Robust methods for partial least squares regression," *J. Chemometrics*, vol. 17, no. 10, pp. 537–549, 2003.

[6] S. Serneels, C. Croux, P. Filzmoser, and P. J. Van Espen, "Partial robust M-regression," *Chemometrics Intell. Lab. Syst.*, vol. 79, nos. 1–2, pp. 55–64, Oct. 2005.

[7] A. Alin and C. Agostinelli, "Robust iteratively reweighted SIMPLS," *J. Chemometrics*, vol. 31, no. 3, p. e2881, Mar. 2017.

[8] M. Markatou, A. Basu, and B. G. Lindsay, "Weighted likelihood equations with bootstrap root search," *J. Amer. Stat. Assoc.*, vol. 93, no. 442, pp. 740–750, Jun. 1998.

[9] M. Alguraibawi, H. Midi, and A. H. M. R. Imon, "A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model," *Math. Problems Eng.*, vol. 2015, pp. 1–12, Jan. 2015.

[10] M. R. Abdulrah, M. Habshah, W. Dihan, and J. Arasan, "Detection of outliers in high dimensional data using nu-support vector regression," *J. Appl. Statist.*, paper CIAS-2020-0085-R2.

[11] K. Varmuz and P. Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*. Boca Raton, FL, USA: CRC Press, 2016.

[12] M. Markatou, "Robust statistical inference: Weighted likelihoods or usual m-estimation?" *Commun. Statist. Theory Methods*, vol. 25, no. 11, pp. 2597–2613, Jan. 1996.

[13] M. Habshah, M. R. Norazan, and A. H. M. Rahmatullah Imon, "The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression," *J. Appl. Statist.*, vol. 36, no. 5, pp. 507–520, May 2009.
A. M. Rashid et al.: Efficient Estimation and Classification Methods for High Dimensional Data

HABSHAH MIDI received the B.A. degree in mathematics from Drew University, USA, in 1979, the Master of Applied Statistics degree from The Ohio State University, USA, in 1981, and the Ph.D. degree from Universiti Kebangsaan Malaysia, in 1995. She is currently a Professor with the Department of Mathematics, Universiti Putra Malaysia (UPM). She also supervises Ph.D. and Master students with more than 30 have already graduated. She is also actively involved with the Malaysia Institute of Statistics and has been appointed as the Vice President, since 2014. In 2018, she has published a book entitled Robust Nonlinear Regression with Application Using R with John Wiley & Sons. She has published more than 150 articles in citation-indexed journals. Her research interests include regression diagnostics, robust statistics, quality control technique, support vector regression, machine learning, pattern recognition, computational statistics, and modelling for high dimensional data. She was awarded the Vice Chancellor Fellowship Award (Excellence in Teaching), in 2007, and was nominated by UPM for the National Academic Award (Excellence in Teaching), in 2016.

WALEED DHHAN SLWABI received the B.S. and M.S. degrees in statistics from Al-Mustansiryiah University, Baghdad, Iraq, in 2001 and 2010, respectively, and the Ph.D. degree in statistics from Universiti Putra Malaysia (UPM), Serdang, Selangor, Malaysia, in 2016. From 2010 to 2011, he was a Lecturer with the University of Babylon, Babylon. From 2011 to 2013, he was a Lecturer with Al Mustaqbal University College, Babylon. From 2017 to 2018, he was a Lecturer with Al-Mussaib Technical Institute, Al-Furat Al-Awsat Technical University. Since 2017, he has been a Research Fellow with the Scientific Research Center, Nawroz University (NZU), Duhok, Iraq. Since 2018, he has been an Assistant Manager at Babylon Municipalities, Babylon Government, Babylon. He has published more than ten articles in citation-indexed journals. He was awarded the bronze medal at the Invention Research Exhibition, UPM, in 2016.

JAYANTHI ARASAN was born in Selangor, Malaysia, in 1971. She received the bachelor’s degree in statistics and operational research from the University of Science, Malaysia, in 1995, the master’s degree in applied statistics from Universiti Putra Malaysia, in 2000, and the D.Phil. degree in statistics from University of Oxford, U.K., in 2006. She is currently an Associate Professor in statistics with the Department of Mathematics, University Putra Malaysia. She has graduated three Ph.D. students and is also supervising five Ph.D. students as a Main Supervisor and 13 as a Co Supervisor. She is also actively involved in statistical research that focusses on survival analysis and computational statistics. She has published 54 articles in journals, 40 papers in conference proceedings, one teaching module, six book chapters, and a book in press.