Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/110140

The Performance of Re-Descending Weight Based Partial Robust M-Regression Methods

Mazni Mohamad[*], *Norazan Mohamed Ramli and Nor Azura Md Ghani*

Centre of Statistical and Decision Science Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia

Received: 2 Nov. 2016, Revised: 7 Dec. 2016, Accepted: 14 Dec. 2016
Published online: 1 Jan. 2017

Abstract: The presence of Partial Robust M-Regression (PRM) amongst other Partial Least Squares Regression (PLSR) techniques is mainly to offer a more robust and efficient method than the existing ones when data face outlier problem. PRM is conceptually different from other robust PLSR techniques because it proposed the usage of M-estimator instead of a more commonly used Least Squares (LS) estimator. Recently, there are several efforts among researchers to further enhance the PRM performance. Among those methods are Partial Robust M-Regression (based on Bisquare Weight Function) (PRMBS) and Partial Robust M-Regression (based on Hampel Weight Function) (PRMH). These two methods are re-descending weight based PRMs which differ from the original monotonous weight based PRM. This study compares the performance of PLS, PRM, PRMBS and PRMH under numerous outlying conditions for both low and high dimensional data sets. Some analysis of real data sets and simulation results in this study show the robustness and the effectiveness of the modified PRM methods.

Keywords: Outliers, PLS, PRM, Re-descending Weighting Functions

1 Introduction

Partial Least Squares (PLS) method was first introduced by Herman Wold way back in 1966 [1]. Partial Least Squares Regression (PLSR) arose with the intention to eliminate the problem of multicollinearity in a regression model. Multicollinearity normally exists when there are huge number of explanatory variables involved and they are highly dependent. The presence of multicollinearity will generally cause inaccuracy in terms of sign and magnitude of the parameter estimates of a model and this can lead to incorrect inferences and wrong interpretation.A straight forward solution to this problem is to reduce the dimension of the explanatory variables. This is normally done by obtaining latent variable, a new variable that has a linear combination with original variables. Principal component analysis (PCA) is among the most widely used techniques for dimensional reduction. In the context of regression, the application of principal component method is always referred as Principal Component Regression (PCR). But the problem with PCR is that there is no assurance that the principal components that explain the exploratory variables also relatable to the response variables. PLSR, on the other hand, estimates regression parameters by finding maximum covariance between latent and response variables such that the residuals of predictive model is at minimum [1].

There are several PLS algorithms offered. Among those common methods are Nonlinear Iterative Partial Least Squares (NIPALS) and Statistically Inspired Modification of the Partial Least Squares (SIMPLS). These methods, however, can easily be influenced by outliers [2]. Failure in identifying outliers will normally resulted in masking or swamping effects in the modeling processes [3].

Several robust PLS methods are therefore recommended to solve the problem concerning outliers. Among the first to introduce robust PLS are [4] who proposed the usage of an iterative reweighted formulation for inner PLS model. Then, [5] introduced another robust technique for inner PLS in 1995 where least median squares and repeated median are suggested to be used besides the iterative reweighted least squares. Alternatively, there are also robust methods meant for outer PLS model. [6] was the first to propose the method by using an Iterative Reweighted PLS (IRPLS).

[*] Corresponding author e-mail: mazni@tmsk.uitm.edu.my

[7]proposed a method that combine methods of [4] and[6], known as Iterative Predictors and Objects Weighting PLS (IPOW-PLS). There is also another approach to robust PLS which is covariance based of robust estimation. This approach is employed by [8] using the Stahel-Donoho estimator (SDE). This technique, however, cannot be applied to high dimensional data as claimed by [2]. Therefore, they proposed a robustified version of the SIMPLS algorithm known as RSIMPLS.

The above mentioned robust PLS techniques are all based on Least Squares (LS) estimator. In regression analysis, LS is the most efficient estimator if the distribution of error terms is normal. Unfortunately, it is not guaranteed to have errors with normal distributions all the time. Therefore, [9] suggested the use of partial M-estimator (PM) instead of LS for cases involving non-normal error distributions. Apart from that, M-estimator is also known to be robust against outliers. PRM, as claimed by [10] outperforms PLS and RSIMPLS in terms of computational cost and statistical properties. Even though PRM outperforms other methods, it is plausible to have some confines since it employs monotonous Fair weighting function which often does not weigh large outliers accordingly [11]. The main objective of this paper is to evaluate and compare the performance of the original PRM (based on Fair weight function) with the other two modified PRMs which are PRMBS and PRMH whose methods are based on re-descending weight functions. The two methods were discussed in [12] and [13] respectively.

2 Partial Robust M-Regression (PRM)

Introduction. In general, PRM offers similar technique as PLS in terms of dimensionality reduction. The major difference between the two is the type of estimators chosen. Principally, PRM uses M-estimator while PLS uses LS estimator. Since M-estimator only cater for vertical outliers, [9] considered Robust M-estimators (RM) in the formulation of PRM so that it is robust against both vertical outliers and leverage points.

Algorithm. Suppose an $n \times p$ data matrix $\mathbf{X}$ be the exploratory variables, and an $n \times 1$ data vector $\mathbf{y}$ be the response variable. The ith observation of $\mathbf{X}$ and $\mathbf{y}$ is denoted by x_i and y_i respectively. Consider the regression model

$$y_i = x_i'\beta + \varepsilon_i, \quad 1 \le i \le n \tag{1}$$

where β is a vector of unknown parameters of size $p \times 1$ and ε_i is the error terms.PRM does not solve the regression model in (1) directly, but instead it regresses the y-variables onto partial information of the x-variables using latent regression model.The h latent variables which are obtained after mean-centring the data can be written in matrix form $T_{n,h} = (t_1, \cdots, t_n)'$. The latent regression model is then obtained as follows:

$$y_i = t_i'\alpha + \phi_i, \quad 1 \le i \le a \tag{2}$$

with the new regression coefficients α and the error terms ϕ_i. Note that $a < n$. The coefficient vector can be estimated as usual except now RM estimator is used instead of LS estimator. In dealing with vertical outliers and leverage points, two types of weights, w_i^r and w_i^x are introduced. The weights w_i^r are computed from residuals $r_i = y_i - x_i\alpha$ where

$$w_i^r = f(\frac{r_i}{\sigma}, c). \tag{3}$$

with $\sigma = MAD(r_1, \cdots, r_n) = median_i \mid r_i - median_i r_i \mid$ be the estimated residual scale, and

$$f(z,c) = \frac{1}{(1+\mid \frac{z}{c} \mid)^2} \tag{4}$$

where $c = 4$ is the tuning constant, and f is the weight function known as Fair function. The weights w_i^x are computed from the scores $\mathbf{T}$ where

$$w_i^x = f(\frac{\parallel t_i - med_{L1}(T) \parallel}{median_i \parallel t_i - med_{L1}(T) \parallel}, c) \tag{5}$$

with $\parallel \dot{\parallel}$ is the Euclidean norm and $med_{L1}(T)$ denotes the L1-median calculated from score vectors. In order to reduce the negative impact of outliers on the regression model, PRM implemented iterative reweighted partial least squares (IRPLS) algorithm. Observations that are close to the centre of the data cloud in the predictor and response spaces will receive a weight close to or equal to one, while leverage and residual points will get a weight close to zero. Particularly, PRM algorithm comprises of the following steps:

Step 1: Determine the robust starting values for the weights $w_i = w_i^r w_i^x$.

Step 2: Execute classical PLS (SIMPLS) on weighted data, $w_i x_i$ and $w_i y_i$.

Step 3: Recalculate w_i^r from PLS residuals, w_i^x from PLS scores, and w_i .

Step 4: Iterate Step 2 and Step 3 until the estimated regression coefficients converge. (i.e., the difference between estimated regression coefficients is smaller than a certain onset value).

Step 5: Find estimated regression coefficients from the last step of weighted PLS.

3 Modified PRM

Re-descending Weight Functions. PRM uses Fair weighting function, a monotonic type estimator that comes from Huber family. Monotonic estimates often have computational advantage, but it may lose the robustness properties in the presence of bad leverage points in dataset [3]. Re-descending type of estimators is therefore recommended because they can produce better breakdown point and provide good efficiency under bounded influence function[14]. This study considers two

re-descending weight based modified PRMs, which are Partial Robust M-Regression (based on Bisquare Weight Function, PRMBS) [12] and Partial Robust M-Regression (based on Hampel Weight Function, PRMH) [13] in comparison to the original monotonous weight based PRM.

PRMBS and PRMH. Basically, methods in obtaining PRMBS and PRMH algorithms are similar to that of the original PRM. For each method, the robust starting values w_i need to be calculated. Since w_i comprises of w_i^r and w_i^x, equations (3) and (5) are once again referred to obtain those weights for both algorithms. The only different here is the weighting function considered in each algorithm. For PRMBS, a Tukey Bisquare weighting function is employed such that the f function in equation (4) is substituted by the following equation (6)

$$f(z,c) = \begin{cases} [(1-(\frac{z}{c})^2]^2 & ,z \le c \\ 0 & ,z > c \end{cases} \tag{6}$$

with the tuning constant c= 4.685.

$$f(z,c) = \begin{cases} 1 & ,\mid z \mid < a \\ \frac{a}{\mid z \mid} & ,a \le \mid z \mid < b \\ a\frac{\frac{c}{\mid z \mid}-1}{c-b} & ,b \le \mid z \mid \le c \\ 0 & ,otherwise \end{cases} \tag{7}$$

Similarly, for PRMH, another re-descending weight function which is Hampel weighting function is introduced to the algorithm. Now, equation (4) becomes the following f function as shown in equation (7) with tuning constants a=2, b=4 and c=8. Note that the tuning constants are generally chosen to give reasonably high eciency in the normal case; particularly, it can produce 95-percent eciency when the errors are normal, and still oer protection against outliers [15]. Once the the robust starting values were obtained, the remaining procedures (ie. Steps 2 to 5) in obtaining PRM algorithm are very much follows to complete the process of getting PRMBS and PRMH algorithms.

4 Simulation Study.

In this section, the performance of PLS, PRM, PRMBS and PRMH are compared by means of the statistical efficiency of each method through a simulation study. Each simulation design was set up to consist of low dimensional data ($n = 100, p = 50$) and high dimensional data ($n = 50, p = 100$) sets with various outlying conditions. For each design, 1000 data sets are generated and only univariate response ($q = 1$) is considered. The design for univariate response is however can always be extended to multivariate responses. The experiments for simulated data were set to be based on the following conditions:

$$\mathbf{T}_{(n,h)} \sim N(3,1) \tag{8}$$

$$\mathbf{B}_{(n,h)} \sim N(3,1) \tag{9}$$

with n be the number of observations, p is the number of parameters, and $h < p$. The data matrix $\mathbf{X}_0$ with perfect collinearity can then be obtained as follows:

$$\mathbf{X}_0 = \mathbf{T}\mathbf{B}^T \tag{10}$$

Finally, the output vector $\mathbf{y}_0$ can be calculated as

$$\mathbf{y}_0 = \mathbf{X}_0\beta_0 = \mathbf{T}\mathbf{B}^T\beta_0 \tag{11}$$

whereβ_0 is the true regression coefficient with $\beta_0 \sim N(3,1)$. The error terms were fixed to be normally distributed, and different types of outliers were introduced to data sets by randomly includes $n \times \alpha$ outliers ($\alpha = 0\%, 5\%, 10\%, 20\%$) to the original n observations. The $n \times \alpha$ observations were generated from a $N(10, 0.5)$ which constitutes a certain percentage of outliers in the samples. The Mean Squared Error (MSE) values of each simulation setup were calculated for all methods using formula written in equation (12). A particular method is considered the best if it produces the lowest MSE value which is defined as

$$MSE = \frac{1}{m_k}\sum_{i=1}^{k}(\hat{\beta}_i - \beta_0)^2 \tag{12}$$

5 Results and Analysis

Tables 1-4 display simulation results. In the absence of outliers, it can be seen that the performance of all methods are more or less the same. This is shown in Table 1. In Table 2, results of simulated MSE for data sets with

Table 1: MSE Values for Low and High Dimensional Data with No Outliers

	Low Dimensional Data	High Dimensional Data
SIMPLS	1.497169	1.408529
PRM	1.485838	1.409266
PRMBS	1.498268	1.408989
PRMH	1.409036	1.409036

different percentage of outliers in x are reported. As expected, at all levels of contamination in both low and

high dimensional data sets, the classical PLS which is SIMPLS, seems to be the least efficient method compared to robust methods. Despite the two newly proposed robust PRM methods, the original PRM outperforms those methods when low dimensional data sets are considered. Original PRM also performs better than the other two robust methods when high dimensional data sets are considered but only for low levels of contamination (ie: 5% and 10% of outliers). On the other hand, PRMBS and PRMH did better job when greater amount of outliers are present in such data sets. Results for simulated MSE for

Table 2: MSE Values for Low and High Dimensional Data with Different Percentage of Outliers in X

Low Dimensional Data					
% outliers	5	10	15	20	25
SIMPLS	81.22252	133.0187	129.9721	117.285	113.3093
PRM	1.411784	1.321365	4.462443	29.04703	39.88564
PRMBS	1.410461	1.351757	1.389804	75.29755	92.28901
PRMH	1.262792	1.374391	1.320785	69.37077	60.5677
High Dimensional Data					
% outliers	5	10	15	20	25
SIMPLS	77.7568	134.4023	136.3651	128.1005	112.3423
PRM	1.4630	1.9066	76.2601	69.6131	72.42998
PRMBS	1.4805	72.1007	71.13321	67.9063	122.6722
PRMH	46.4197	92.6246	87.4993	80.4938	71.61626

data sets with outliers in y are displayed in Table 3. SIMPLS once again cannot uphold its optimum efficiency level when outliers are involved at all levels and in all sets of data. In earlier discussion where outliers in x are considered, we have seen that original PRM outperforms PRMBS and PRMH for low dimensional data sets. Now, the results are no longer the same as outliers in y are considered. PRMBS and PRMH outperform original PRM at all contamination levels for both low and high dimensional data sets. Interestingly, PRMH performs better than PRMBS when amount of outliers are less than 15 percent in low dimensional data sets, whereas PRMBS outperforms PRMH when such data sets contain greater amount of outliers. In contrast, PRMH did better job than PRMBS when dealing with high dimensional data sets which consist of 20 percent outliers or more and vice versa. The following Table 4 shows results of simulated MSE values for low and high dimensional data in the presence of outliers in both x and y directions. It can be seen from the table that for cases where low dimensional data sets are contaminated with low percentage of outliers (ie: less than 20% outliers), the proposed PRMBS outperforms other methods. The original PRM is however performs better when greater amount of outliers are considered (ie: 20% or more). In the case of high dimensional data, PRMBS seems to consistently perform better than other methods.

Table 3: MSE Values for Low and High Dimensional Data with Different Percentage of Outliers in Y

Low Dimensional Data					
% outliers	5	10	15	20	25
SIMPLS	3.452491	7.872901	1.667845	2.061536	3.799479
PRM	1.462593	1.590083	1.710465	2.516873	1.412204
PRMBS	1.411243	1.533443	1.48748	1.465548	1.403227
PRMH	1.412539	1.46378	1.427978	1.480731	1.422577
High Dimensional Data					
% outliers	5	10	15	20	25
SIMPLS	3.420626	22.60026	40.58733	57.8072	21.85068
PRM	1.546828	2.090233	2.140008	3.524307	5.364179
PRMBS	1.518511	1.684201	1.815734	3.002523	3.835838
PRMH	1.527621	1.734916	1.819363	2.6178	3.359234

Table 4: MSE Values for Low and High Dimensional Data with Different Percentage of Outliers in Both X and Y Directions

Low Dimensional Data					
% outliers	5	10	15	20	25
SIMPLS	7.622335	43.03615	57.04251	127.9113	196.5656
PRM	1.536904	1.50333	1.423573	10.56944	1.159602
PRMBS	1.512869	1.500601	1.364451	60.05588	8.889315
PRMH	1.516554	6.811315	65.44217	85.57625	129.2603
High Dimensional Data					
% outliers	5	10	15	20	25
SIMPLS	129.3564	30.93684	50.48263	121.7722	173.504
PRM	1.40997	1.426957	7.432936	2.319515	62.1326
PRMBS	1.409766	1.429526	1.445844	1.119283	63.53992
PRMH	1.409585	68.22629	61.52605	78.72803	90.66833

6 Numerical Examples.

Finally, we apply all four methods to a couple of real data sets to further investigate the performance of each. We estimate the performance of each method based on the bias, standard error of prediction (SEP) and mean squared error (MSE) values. The three criteria are calculated as follows:

$$bias = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{13}$$

$$SEP = \sqrt{\frac{1}{n-1}\sum i = 1^{n}(y - \hat{y}_i - bias)^2} \tag{14}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{15}$$

where $\hat{y}_i$ is the estimated y_i .

PAC Data. The first data set that we used is a high dimensional PAC data. This data is accessible through R-package chemometrics. It consists of 209 observations with $467X$-variables and a response y-variable. It describes polycyclic aromatic compounds (y) in terms of GC-retention indices of which have been modelled by molecular descriptors (X)[16]. Reference [17] also used this data set in their study. As reported in [17], this data

Table 5: Bias, SEP and MSE Values for PAC Data

	BIAS	SEP	MSE
SIMPLS	-338.0862	1.131717	114570.0
PRM	0.8626399	1.275216	340.6148
PRMBS	-0.3621655	1.17488	288.6228
PRMH	3.028136	1.301496	363.1931

set is likely to contain outliers in the *y*-variables. Results for PAC data are shown in Table 5. PRMBS seems to be the best method since its calculated bias, SEP and MSE are the lowest. Subsequently, this also indicates that the outcome is consistent to that of simulation result concerning high dimensional data with outliers in *y*.

NIR Data. Another data set used in this study is NIR data. It consists of 166 samples of alcoholic mash that were obtained through fermentation processes that vary in accordance to different feedstock (corn, rye and wheat). The first derivatives of near infrared spectroscopy (NIR) absorbance values with wavelength range between 1115-2285 nm serves as 235 *X*-variables, and the concentration of glucose and ethanol (in g/L) be the *y*-variables[16]. Since the focus of this study is mainly on univariate response variable, we have chosen the variable ethanol concentration to be the response variable (*y*). The

Table 6: Bias, SEP and MSE Values for NIR Data

	BIAS	SEP	MSE
SIMPLS	-58.01548	0.2436876	3375.654
PRM	-2.024116	0.9073991	140.777
PRMBS	-0.08776794	0.4642604	35.78697
PRMH	0.3524564	0.4211974	29.57383

above Table 6 shows the NIR data results. Apparently, both re-descending weight based PRMs, which are PRMBS and PRMH produced better results than SIMPLS and original PRM.

7 Conclusion

On the whole, the classical PLS, with SIMPLS algorithm, loses its optimum efficiency criteria when data sets are contaminated with outliers at any directions. Simulation results show that original PRM did well in data sets which are contaminated with outliers in x. Conversely, when dealing with contaminated data sets with outliers in y, original PRM seems to be less efficient than the modified PRMs which are PRMBS and PRMH. When data sets being contaminated with outliers in both x and y, PRMBS produced the most consistent results and outperform other methods in most situations.

Acknowledgement

The authors gratefully acknowledge the financial support received from the Ministry of Higher Education of Malaysia and the Universiti Teknologi MARA for supporting this research under the Research Grant No. 600-RMI/DANA 5/3/CIFI (65/2013).

References

[1] G. M. Morales, Partial Least Squares (PLS) Methods: Origins, Evolutions and Application to Social Sciences, Commun. Stat - Theory Methods, vol. 40, pp. 2305-2317, 2011.
[2] M. Hubert and K. V. Branden, Robust Methods for Partial Least Squares Regression, Journal of Chemometrics, vol.17, pp, 537-549, 2003.
[3] P. Filzmoser, S. Serneels, R. Maronna and P.J. V. Espan, Robust Multivariate Methods in Chemometrics in Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Amsterdam: Elsevier,pp. 681-722, 2009.
[4] I. N. Wakelinc and H. J. H. Macfie, A Robust PLS Procedure, Journal of Chemometrics, vol. 6, pp. 189-198, 1992.
[5] M. I. Griep, I. N. Wakelinc, P. Vankeerberghen and D. L. Massart, Comparison of Semirobust and Robust partial least Squares procedures, Chemom.Intell. Lab. Syst, vol. 29, pp,37-50, Jul.1995.
[6] D. J. Cummins and C. W. Andrews, Iteratively Reweighted Partial Least Squares: A Performance Analysis by Monte Carlo Simulation, Journal of Chemometrics, vol. 9, pp. 489507, 1995.
[7] M. Forina, C. Casolino, and E. M. Almansa, The Refinement of PLS Models by Iterative Weighting of Predictor Variables and Objects, Chemom. Intell. Lab. Syst., vol. 68, pp. 2940, 2003.
[8] J. A. Gil and R. Romera, On Robust Partial Least Squares (PLS) Methods, Journal of Chemometrics, vol. 12, pp. 365378, 1998.
[9] S. Serneels, C. Croux, P. Filzmoser, and P. J. Van Espen, Partial Robust M-Regression, Chemom. Intell. Lab. Syst., vol. 79, pp. 5564, Oct. 2005.
[10] S. F. Mller, J. Von Frese, and R. Bro, Robust Methods for Multivariate Data Analysis,J. Chemom., vol. 19, pp. 549563, 2005.

[11] M. R. Norazan, Weighted Maximum Median Likelihood Estimation for Parameters in Multiple Linear Regression Model, Universiti Putra Malaysia, 2008.
[12] M. Mazni, N. Mohamed Ramli, N. A. Md Ghani @ Mamat, and S. Ahmad, Enhancement of Partial Robust M-Regression (PRM) Performance using Bisquare Weight Function, in AIP Conference Proceedings, 1613, pp. 122129, 2014.
[13] M. Mazni, N. A. Md Ghani @ Mamat, N. Mohamed Ramli, and S. Ahmad, The Refinement of Partial Robust M-Regression Model using Winsorized Mean and Hampel Weight Function, in AIP Conference Proceedings, 1643, pp.175-180 2015.
[14] C. H. Muller, Redescending M-Estimators In Regression Analysis, Cluster Analysis And Image Analysis, in Discussiones Mathematicae Probability and Statistics, vol. 24, pp. 5975, 2004.
[15] J. Fox, Robust Regression, in Appendix to An R and S-PLUS Companion to Applied Regression, pp.1-8, January, 2002.
[16] K. Varmuza and P. Filzmoser, Introduction to Multivariate Statistical Analysis. Boca Raton, FL: CRC Press, 2009.
[17] B. Liebmann, P. Filzmoser, and K. Varmuza, Robust and Classical PLS Regression Compared, Journal of Chemometrics. Spec. Issue Conf. Chemom. 2009, vol. 24, pp. 111120, 2010.

Mazni Mohamad had her first degree in Statistics from University of Tennessee, Knoxville USA. She got her master degree in Applied Statistics from Universiti Putra Malaysia and is now pursuing her doctoral degree in Statistics. She is currently a senior Statistics lecturer at the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA(UiTM), Malaysia. Her research interests are Robust Modeling, Outliers and Missing Values and Bootstrapping Techniques.

Norazan Mohamed Ramli has graduated from Polytechnic of East London, United Kingdom with BSc Mathematics and Statistics (First Class Honors). She then obtained her PostGraduate Diploma (Statistics and Operational Research), at the University of Essex, United Kingdom. She did her master degree in Applied Statistics and pursue her Doctoral degree in Statistics at the same university, which is Universiti Putra Malaysia. The areas of her research interest are Robust Modeling, Outliers and Missing Values, Bootstrapping Techniques and Quality Control Charts.

Nor Azura Md Ghani is an Associate Professor of Statistics at the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA(UiTM), Malaysia. She obtained all her degrees; bachelor, master and doctoral in Statistics from the same university, which is Universiti Kebangsaan Malaysia. Her research area of interests are Forensic Statistics, Pattern Recognition and Data Mining.