Hindawi
Journal of Applied Mathematics
Volume 2018, Article ID 7696302, 5 pages
http://dx.doi.org/10.1155/2018/7696302

Research Article

Robust Nonlinear Partial Least Squares Regression Using the BACON Algorithm

Abdelmounaim Kerkri,[1] **Jelloul Allal,**[1] **and Zoubir Zarrouk**[2]

[1] *Department of Mathematics, University Mohamed the First, Oujda 60000, Morocco*
[2] *Department of Management, Faculty of Social Sciences, Oujda 60000, Morocco*

Correspondence should be addressed to Abdelmounaim Kerkri; krkabdelmounaim@gmail.com

Received 11 June 2018; Revised 18 August 2018; Accepted 18 September 2018; Published 2 October 2018

Academic Editor: Lucas Jodar

Partial least squares regression (PLS regression) is used as an alternative for ordinary least squares regression in the presence of multicollinearity. This occurrence is common in chemical engineering problems. In addition to the linear form of PLS, there are other versions that are based on a nonlinear approach, such as the quadratic PLS (QPLS2). The difference between QPLS2 and the regular PLS algorithm is the use of quadratic regression instead of OLS regression in the calculations of latent variables. In this paper we propose a robust version of QPLS2 to overcome sensitivity to outliers using the Blocked Adaptive Computationally Efficient Outlier Nominators (BACON) algorithm. Our hybrid method is tested on both real and simulated data.

1. Introduction

After it was developed by Wold [1], PLS regression became a classic way to overcome correlation in regression analysis; this method is popular in many fields such as genomics and chemometrics. Many statisticians showed interest in the mathematical properties of the method; De Jong [2] proved that the PLS estimator is a regularized version of the ordinary least squares estimator. The same result was later demonstrated algebraically by Goutis et al. [3]. With the arising of data that show nonlinear behavior in many fields, it was necessary to have a new version of PLS regression that captures the nonlinearity and provides more parsimonious models. Wold [4] developed the first nonlinear version of the PLS algorithm by substituting OLS with a quadratic regression to calculate the PLS components. Wold [5] also proposed the spline PLS algorithm. Another nonlinear algorithm based on neural networks to deal with the nonlinearity of meteorological data was proposed [6].

PLS regression is sensitive to outliers and leverages. Thus several robust versions have been proposed in the literature, but only for linear PLS. Hubert [7] proposed two robust versions of the SIMPLS algorithm by using a robust estimation for the variance-covariance matrix. Kondylis and Hadi [8] used the BACON algorithm to eliminate outliers, resulting in a robust linear PLS.

In this work we attempt to obtain a robust version of the quadratic PLS algorithm QPLS2, by using the BACON algorithm. An application on real and simulated data is used to validate the method.

2. Nonlinear PLS Regression

Every linear regression method is based on the following optimization problem:

$$\min \left\| X\beta - Y \right\|_{\beta}, \tag{1}$$

where $X \in \mathbb{R}^{n\times m}$ is a matrix presenting the values of the independent variables, $Y \in \mathbb{R}^{n}$ is the dependent variable, and β is the coefficient of the regression.

Instead of regular predictors, PLS regression uses a set of latent variables called scores: $t_k = X_k \omega_k$ (with X_k the deflated version of the initial matrix X). The latent variables (also called the PLS components) are iteratively calculated, based on the decomposition:

$$X = t_1 {p'}_1 + \cdots + t_m {p'}_m + E, \tag{2}$$

where E is the error, p_k is a set of vectors called the loadings, and ω_k a weight vector of length k. As mentioned in the introduction, owing to the encounter of data that showed nonlinear behavior, many researchers proposed new PLS algorithms to capture the nonlinearity of these datasets. In this work we use the quadratic nonlinear PLS as proposed by Wold [4].

The quadratic nonlinear PLS is a PLS algorithm that supposes the existence of nonlinear relations between the two blocks of variables. Instead of the OLS regression presented in the linear PLS algorithm

$$u = c_1 + c_2 t, \tag{3}$$

Wold et al. [4] used a quadratic regression:

$$u = c_1 + c_2 t + c_3 t^2. \tag{4}$$

Every regression method performs poorly in the presence of outliers. As a result of the instability of the estimations, many approaches have been developed to overcome this problem, such as filtering the outliers from the dataset, or giving them lower weights to minimize their effect on the estimation process. The next section will focus on the BACON algorithm, as an approach that deletes the outliers to obtain a clean dataset.

3. Robust PLS Regression

3.1. Outliers Detection and Robust Regression. Robust regression is a way of dealing with outliers, which are observations that come from a different distribution. They can also be the result of error measurements, and can harm the quality of the estimation. Just like OLS regression, PLS regression is also sensitive to outliers [8]. Hence their detection is a necessary procedure, in order to have stable estimations, and accurate predictions.

Many researchers proposed methods of dealing with the outlier problem in PLS regression. Hubert [7] used two robust estimations of the variance-covariance matrix in the SIMPLS algorithm, and Kondylis and Hadi [8] used the BACON algorithm for outlier detection. Both approaches proved to be a significant improvement over the regular PLS.

The BACON algorithm [9] starts with a subset of observations of size m^* that is supposedly free of outliers, and then it iteratively adds the observations that are consistent with the initial set. The observations left out are the outliers.

The first set is chosen. Then the distance is defined and used as a criterion for including the observation in the initial subset. Here are two distances used in the literature

$$d_i(S) = \sqrt{x'_i S^{-1} x_i} \tag{5}$$

and

$$d_i(x_i, m) = \|x_i - m\| \tag{6}$$

S is the variance-covariance matrix of the entire data set, x_i represents the i observation, the first distance is called the Mahanalobis distance, and the second is simply the distance of the observation from the median m. Here are the detailed steps of the algorithm:

(1) Select an initial set X_b

(2) Compute the distances ($\overline{x}_b$ is the mean of X_b, and S_b is the matrix of covariance of X_b):

$$d_i(\overline{x}_b, S_b) = \sqrt{(x_i - \overline{x}_b)' S_b^{-1} (x_i - \overline{x}_b)}, \quad i = 1, \ldots, n \tag{7}$$

(3) Set the new subset with all the points that have

$$d_i(\overline{x}_b, S_b) < C_{npr} \times \chi_{p,\alpha/n} \tag{8}$$

where $\chi_{p,\alpha/n}$ is the $(1-\alpha)$ Chi-square percentile and

$$\begin{aligned} C_{npr} &= C_{np} + C_{hr} \\ C_{np} &= 1 + \frac{p+1}{n-p} + \frac{2}{n-1-3p} \\ C_{hr} &= \max\left[0, \frac{h-r}{h+r}\right], \\ h &= \frac{n+1+p}{2} \end{aligned} \tag{9}$$

(4) Repeat (2) and (3) until the subset does not change.

(5) X_b is the dataset free from outliers.

3.2. Robust Nonlinear PLS. We merge the BACON algorithm with the quadratic PLS, with the goal of obtaining a robust version of the algorithm:

(1) Run the BACON algorithm on the dataset using distance (6), and keep the outcome X_b. Then delete the observations in the dependent variable related to the outliers to obtain Y_b (free from outliers).

(2) For every PLS dimension, repeat until convergence of t (u is a the first column of Y_b)

(i) Calculate the weights:

$$w = \frac{u' X_b}{u' u}. \tag{10}$$

(ii) Calculate the scores:

$$t = \frac{X_b w}{w' w}. \tag{11}$$

(iii) Fit u to c using the quadratic function and calculate r the prediction of u using the nonlinear estimates:

$$u = c_1 + c_2 t + c_3 t^2. \tag{12}$$

(iv) Calculate

$$q = \frac{Y'_b r}{r' r}. \tag{13}$$

Table 1: Comparison between explained variance of proposed robust quadratic PLS and original quadratic PLS in cosmetic dataset.

	t_1	t_1	t_1	t_1	t_1	t_1	t_1	t_1	Cumulated variance
X^a	0.286	0.196	0.129	0.139	0.05	0.11	0.08	0.003	0.99
X^b	0.277	0.239	0.155	0.177	0.051	0.093	0.004	0	0.99
Y^a	0.180	0.077	0.137	0.134	0.042	0.04	0.065	0.03	0.68
Y^b	0.33	0.181	0.103	0.117	0.06	0.037	0.0.32	0.05	0.91

(v) Update u :

$$u = \frac{Y_b q}{q' q}. \tag{14}$$

(vi) Update w as described in (i).

(vii) Calculate the new value of t:

$$t = \frac{X_b w}{w' w}. \tag{15}$$

(3) Calculate the loadings using the final value of t:

$$p' = \frac{t' X_b}{t' t}. \tag{16}$$

(4) Deflate X_b and Y_b:

$$\begin{aligned} E &= X_b - t p' \\ F &= Y_b - r q'. \end{aligned} \tag{17}$$

(5) If an additional dimension is required, replace X_b and Y_b with E and F and repeat the steps from (2) to (4).

4. Application

The goal of this application is to compare the performance of the robust quadratic PLS with the original quadratic PLS. The comparison is conducted on both simulated and real data.

4.1. Real Data. We use the dataset presented in [4], which contains 8 different formulations of cosmetic products, as predictive variables, and 11 dependent variables presenting quality indicators collected in an experiment on 17 individuals.

Since we cannot calculate the mean squared error, we will compare the percentage of explained variance in both the robust and original quadratic PLS:

$$\operatorname{var}(Y, t_h) = \frac{1}{p*} \sum_{i=1}^{p*} cor(Y_i, t_h)^2 \tag{18}$$

and

$$\operatorname{var}(X, t_h) = \frac{1}{p} \sum_{i=1}^{p} cor(X_i, t_h)^2 \tag{19}$$

t_h is the latent component of the h_{th} PLS iteration, $p*$ is the number of dependent variables, and p is the number of predictive variables.

In Table 1, a comparison of the original and robust quadratic PLS shows that the latter improves the explained variance in the dependent variables from 68% to 91%, which is a considerable amount. This is an indicator that the dataset contained outliers that affected the estimation in the case of the original quadratic PLS.

4.2. Simulated Data. In this section, a contamination study is used to assess the quality of the proposed robust method, by following these steps:

(1) The nonlinear function presented in [10] which is used to generate a dataset with 500 observations and 6 variables (where $X = (x_1, \ldots, x_6)$ is generated by a uniform distribution):

$$\begin{aligned} y = {} & 10 \sin(\pi x_1 x_2) + 20 (x_3 - 5)^2 + 10 x_4 + 5 x_5 \\ & + 0 x_6. \end{aligned} \tag{20}$$

(2) The dataset is randomly contaminated by adding a small percentage of data (5%, 10%, and 15%) from a multivariate normal distribution.

(3) We first apply the quadratic PLS to the generated data, and then we apply the robust quadratic PLS described previously.

(4) We compare the original quadratic PLS with the proposed robust PLS using the explained variance, as well as the predictive mean squared error and the predicted residual error sum of squares (PRESS).

The dataset is simulated 1000 times. The explained variance, predictive mean squared error, and PRESS are the mean of all values calculated for each dataset.

In case of a 5% contamination rate (Table 2), the original quadratic PLS yields a total explained variance of 73%, but when applying the robust quadratic PLS, this explained variance becomes 99% which is a considerable improvement. The same can be said about the 10% and 15% contamination rates, where we see an improvement in the explained variance of the dependent variable.

The dataset of 500 observations was then split in two parts. The first contained 400 observations used in the estimation of two models: one with the original quadratic PLS and one with the robust quadratic PLS. Then we calculate the predictive residual mean squared error (RMSEP) of the dependent variable on the 100 left out observations.

The results of a comparison (Table 3) of the three contamination rates show that the robust quadratic PLS yields a smaller mean squared prediction error in every case. The

Table 2: Comparison between explained variance of proposed quadratic algorithm and original one in simulated dataset for the three contamination rates (5%, 10%, and 15%).

Contamination rate	Explained variance by original quadratic PLS		Explained variance by robust quadratic PLS	
	X	Y	X	Y
5%	0.99	0.73	1	0.99
10%	1	0.68	1	0.99
15%	1	0.67	1	0.99

Table 3: Comparison between optimal mean squared prediction error and predictive error sum of squares of proposed quadratic algorithm and original one for simulated dataset with three contamination rates (5%, 10%, and 15%).

Contamination rate	Mean squared prediction error (MSPE)		Predictive error sum of squares (PRESS)	
	Quadratic PLS	Robust quadratic PLS	Quadratic PLS	Robust quadratic PLS
5%	103.07	12.83	100.9	89
10%	110.42	60.9	108.75	17.15
15%	119.08	4.32	117.37	17.15

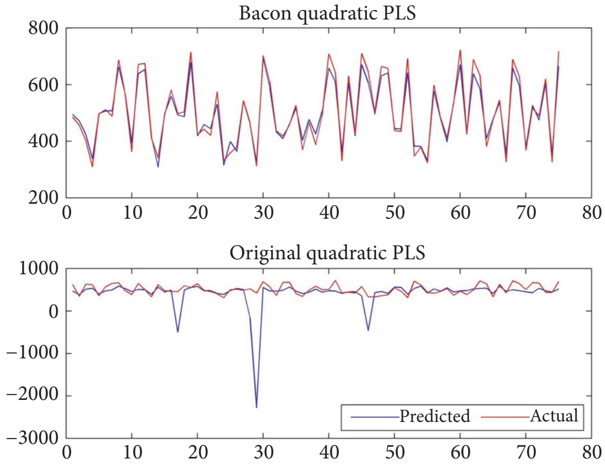

Figure 1: Comparison of predicted and actual values of test dataset in case of quadratic and robust quadratic PLS regression on 5% contaminated data.

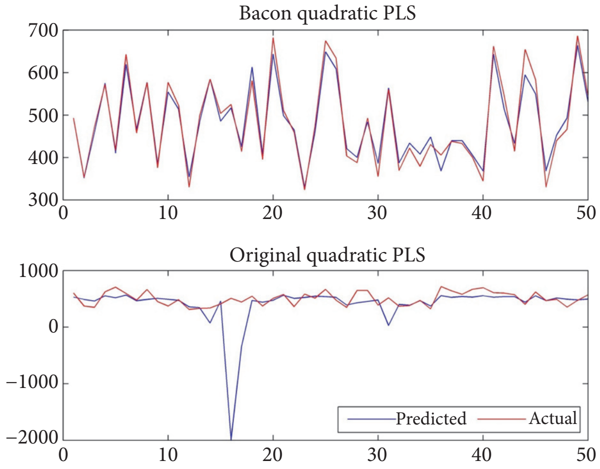

Figure 2: Comparison of predicted and actual values of test dataset in case of quadratic and robust quadratic PLS regression on 10% contaminated data.

same table presents the values of the PRESS for each rate, calculated by leaving 10% of the observations. The same can be said about the predictive error sum of squares as it is improved in the case of the robust quadratic PLS.

Figures 1, 2, and 3 show a comparison of the predicted values and the actual values of the simulated dataset, for both quadratic and robust quadratic PLS regression. For all contamination rates the prediction is improved significantly in the case of the proposed robust quadratic PLS, as it gives better predictions than the original one.

5. Conclusion

PLS regression has developed considerably since it was first introduced. The nonlinear nature of data encountered in the field of chemical engineering was the motivation behind developing nonlinear PLS methods. In this paper we proposed a robust version of the quadratic nonlinear PLS,

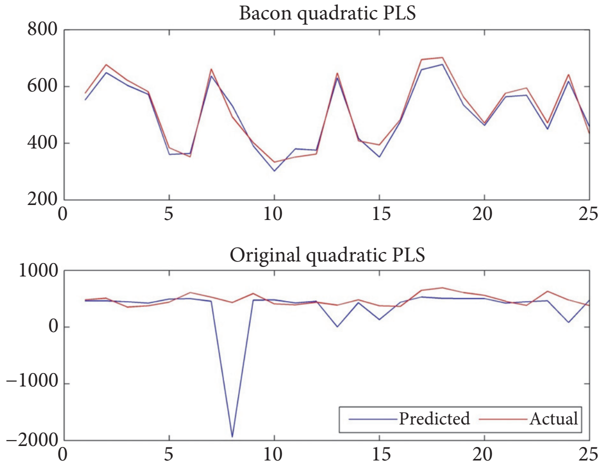

Figure 3: Comparison of predicted and actual values on test dataset in case of quadratic and robust quadratic PLS regression on 15% contaminated data.

in a hybrid form between the quadratic PLS algorithm and the BACON algorithm in order to overcome problems caused by outliers. Our method outperformed the quadratic PLS for both real and simulated data.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] H. Wold, “Soft modeling by latent variables: the nonlinear iterative partial least squares approach,” *In Perspectives in Probability and Statistics, Papers in Honour of MS Bartlett*, pp. 520–540, 1975.

[2] S. De Jong, “PLS shrinks,” *Journal of Chemometrics*, vol. 9, no. 4, pp. 323–326, 1995.

[3] C. Goutis, “Partial least squares algorithm yields shrinkage estimators,” *The Annals of Statistics*, vol. 24, no. 2, pp. 816–824, 1996.

[4] S. Wold, N. Kettaneh-Wold, and B. Skagerberg, “Nonlinear PLS modeling,” *Chemometrics and Intelligent Laboratory Systems*, vol. 7, no. 1-2, pp. 53–65, 1989.

[5] S. Wold, “Nonlinear partial least squares modelling. II. Spline inner relation,” *Chemometrics and Intelligent Laboratory Systems*, vol. 14, no. 1–3, pp. 71–84, 1992.

[6] Z. Meng, S. Zhang, Y. Yanh et al., “Nonlinear partial least squares for consistency analysis of meteorological data,” *Mathematical Problems in Engineering*, vol. 2015, Article ID 143965, 8 pages, 2015.

[7] M. Hubert and K. V. Branden, “Robust methods for partial least squares regression,” *Journal of Chemometrics*, vol. 17, no. 10, pp. 537–549, 2003.

[8] A. Kondylis and A. S. Hadi, “Derived components regression using the BACON algorithm,” *Computational Statistics & Data Analysis*, vol. 51, no. 2, pp. 556–569, 2006.

[9] N. Billor, A. S. Hadi, and P. F. Velleman, “BACON: blocked adaptive computationally efficient outlier nominators,” *Computational Statistics and Data Analysis*, vol. 34, no. 3, pp. 279–298, 2000.

[10] V. Cherkassky, D. Gehring, and F. Mulier, “Comparison of adaptive methods for function estimation from samples,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 7, no. 4, pp. 969–984, 1996.