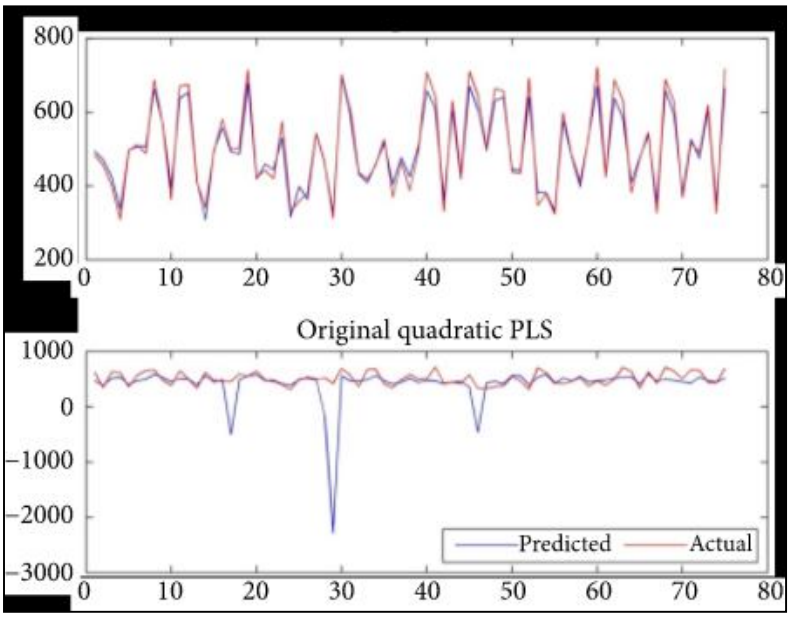
ISSN 2231-5063 Impact Factor : 4.6052(UIF) Volume - 6 | Issue - 4 | October - 2016

GRT Golden Research Thoughts

ROBUST NONLINEAR PARTIAL LEAST SQUARES REGRESSION USING THE BACON ALGORITHM

Dr.Ravindra S. Acharya
Professor in Mathematics,
Vishwakarma Institute of Information Technology, Kondhava, Pune.



ABSTRACT
Multidimensional square regression (PLS regression) is used as an alternative to normal minimum square regression in the presence of normal multicolinearity. This phenomenon is common in chemical engineering problems. In addition to the linear form of PLS, there are other versions based on the nonlinear method, such as quadrilateral PLS (QPLS2). The difference between QPLS2 and regular PLS algorithms is the use of quadratic regression instead of OLS regression in the calculation of latent variables. In this paper we propose a robust version of QPLS2 to overcome the sensitivity of outsiders using the Block Adaptive Computationally Efficient Outlier Nominator (BACON) algorithm. Our hybrid method is tested on both real and simulated data.

KEY WORDS: Multidimensional square, nonlinear method, quadratic regression.

INTRODUCTION:
Many statisticians showed interest in the mathematical properties of the method; De Jong proved that the PLS estimator is a regular version of the General Square Estimator. The same result was later demonstrated algebraically by Gautis et al. With the emergence of data showing nonlinear behavior in many areas, there needed to be a newer version of PLS regression that captures nonlinearity and provides a more transparent model. WLD developed the first nonlinear version of the PLS algorithm by quadratic regression to OLS to calculate PLS components. Vold suggested the spline PLS algorithm. Another nonlinear algorithm based on neural networks is proposed to counter the alignment of meteorological data.PLS is sensitive to regression outliers and leverage. Thus many robust versions have been proposed in the literature, but only for linear PLS. Hubert proposed two robust versions of the SIMPLS algorithm using a strong prediction for the variation-with-variation matrix. Kondylis and Hadi used the BACON algorithm to eliminate outsiders, resulting in a strong linear PLS.

Nonlinear PLS Regression:

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

$$min\|X\beta - Y\|_{\beta}$$

Equation – 1

Where,

$X \in \mathbb{R}^{nXm}$isa matrix representing the values of individual variables,

$Y \epsilon \mathbb{R}^{n}$Is the dependent variable, and β is the regression coefficient.

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

$$X = t_1P_1 + \cdots + t_mP_m + E$$

Equation – 2

Where the error is the, $p$ve is a set of vectors called loading, and length is the length of the length vector. As mentioned in the introduction, due to the visitation of data showing nonlinear behavior, several researchers proposed new PLS algorithms to capture the nonlinearity of this dataset. In this work we use quadrilateral nonlinear PLS as proposed by Wold.

Quadrilateral nonlinear PLS is a PLS algorithm that assumes the existence of a nonlinear relationship between two blocks of variables. Instead of the OLS regression presented in the linear PLS algorithm.

$$u = c_1 + c_2t$$

Equation – 3

Following quadratic regression has also used…

$$u = c_1 + c_2t + c_3t^2$$

Equation – 4

Each regression method performs poorly in the presence of outsiders. As a result of predictive instability, several approaches have been developed to overcome this problem, such as filtering out the datasets or giving them less weight to minimize their impact on the prediction process. The next section will focus on the BACON algorithm, as an approach that deletes external people to get a clean dataset.

Robust PLS Regression:

Outlier Detection and Robust Regression:

Strong regression is a way of dealing with outsiders, which is an observation made by different distributions. They can also be the result of error measurement and damage the quality of the estimate. Like OLS regression, PLS regression is also sensitive to outliers. Therefore, estimating them is a necessary process, so that stable predictions and accurate predictions can be made. Several researchers have suggested methods of dealing with external problems in PLS regression. Hubert used two strong predictions of variation-covariancemetrics in the SIMPLS algorithm, and Kandilis and Hadi used the Bacon algorithm for external investigations. Both approaches proved to be significant improvements compared to regular PLS. The BACON algorithm starts with a subgroup of shape observations supposedly free from outliers and then repeats it and adds observations consistent with the initial set. The first set is selected. The distance is then defined and used as a criterion to include observations in the initial subgroup. Following are the two gaps used in the literature…

$$d_i(S) = \sqrt{x_iS^{-1}x_i}$$

Equation – 5

and

$$d_i(x_i, m) = \|x_i - m\|$$

Equation – 6

S is the variance-consensus matrix of the entire data set, the $x_i$represents the *i*observation, the first distance is called the Mahanlobis distance, and the second is the distance of the observation from the median. Here are the detailed steps of the algorithm:

1. Select an initial set X_b
2. Calculate the distance ($\bar{x}_b$is the average of X_b of, andS_b is the matrix of the covariance X_b):

$$d_i(\bar{x}_b, S_b) = \sqrt{(x_i - \bar{x}_b)S_b^{-1}(x_i - \bar{x}_b)},\ i = 1$$

Equation – 7

3. Set new subset for all point which we have

$$d_i(\bar{x}_b, S_b) < C_{npr}\ X\ X_{p,\alpha/n}$$

Equation – 8

Where,

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

$C_{npr} = C_{np} + C_{hr}$

$$C_{np} = 1 + \frac{p+1}{n-p} + \frac{2}{n-1-3p}$$

$$C_{hr} = max\left[0, \frac{h-r}{h+r}\right], \qquad h = \frac{n+1+p}{n}$$

4. Repeat equation 2 and 3 until the subset does not change
5. $X_b$is the dataset free from outliers.

Robust Nonlinear PLS:

In order to obtain a stronger version of the algorithm, we merge the BACON algorithm into quadrilateral PLS:

1. Run the BACON algorithm on the dataset using the distance in equation (6) and keep the result X_b. Then delete the observations in the dependent variable associated with the outliers to get the outliers Y_b(free from outliers).
2. For each PLS parameter, repeat until the convergence of t (u is the first column of the Y_b)

i. Weight Calculation:

$$\omega = \frac{uX_b}{uu}$$

Equation – 9

ii. Scores Calculation:

$$t = \frac{X_b\omega}{\omega\omega}$$

Equation – 10

iii. Fit $u\ to\ c$ calculate using the quadrilateral function *r* Predict of using nonlinear estimates:

$$u = c_1 + c_2 t + c_3 t^2$$

Equation – 11

iv. Calculate:

$$q = \frac{Y_b r}{r r}$$

Equation – 12

v. Update u:

$$u = \frac{Y_b q}{q q}$$

Equation – 13

vi. Update ω as described in equation

vii. Calculate new vale for t:

viii.

$$t = \frac{X_b \omega}{\omega \omega}$$

Equation – 14

3. Calculate loading the final value using t:

$$p = \frac{t X_b}{t t}$$

4. Reduce X_b and Y_b

$$E = X_b - tp$$
$$F = Y_b - rp$$

Equation – 15

5. If additional dimensions are required, replace Xb and Y E with E and F and repeat steps from (2) to (4).

Application:

The goal of this application is to compare the performance of a strong quad PLS with the original quad PLS. Comparisons are made on both simulated and real data.

Actual Data:

We use the dataset presented in Equation [4], which consists of 8 different formulations of cosmetic products, predictive variables, and 11 dependent variables representing quality indicators collected in an experiment on 17 individuals. Since we cannot calculate the square error, we will compare the percentage of difference explained in both the strong and the original quadrant PLS:

$$var(Y, t_h) = \frac{1}{P^*} \sum_{i=1}^{p^*} cor(X_i, t_h)^2$$

Equation – 16

And

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

Equation – 17

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

Pretended Data:

In this section, the malicious study is used to evaluate the quality of the proposed robust method by following these steps:

1. The nonlinear function introduced in equation -10 which is used to create a dataset with 500 observations and 6 variables (where $X = (x_1, .., x_6)$ is generated by uniform distribution)

$$y = 10sin\,(\pi x_1 x_2) + 20(x_3 - 5)^2 + 10x_4 + 5x_5 + 0x_6$$

Equation – 18

2. Adding a small percentage of data (5%, 10% and 15%) from a multivariate normal distribution randomly corrupts the dataset.
3. We first apply quad PLS to the generated data and then we apply the strong quad PLS described earlier.
4. We have compared the apparent quadrant PLS with the proposed strong PLS, as well as using the approximate average square error and the approximate residual error (PRESS) of the square.

The dataset is replicated 1000 times. Explained variations, predictive square errors, and presses are the themes of all values calculated for each dataset. In the case of a 5% contamination rate, the original quad PLS gives a total explanatory variation of 73%, but when applying a strong quad PLS, this explained variance becomes 99% which is a significant improvement. The same can be said about the 10% and 15% contamination rates, where we see an improvement in the defined variation of the dependent variable. The dataset of 500 observations was then divided into two parts. The first used approximately 400 observations of the double model: one with the original quad PLS and one with the strong quad PLS. We then calculate the predictive residual mean square error (RMSEP) of the dependent variable based on the 100 left observations.

The results of the comparison of the three contamination rates show that the strong quadrant PLS gives a small average square prediction error in each case. The same table presents the values of PRESS for each rate, calculated leaving 10% observations. The same can be said about the sum of the predictive errors of the square as it is corrected in the case of strong quadrilateral PLS.

CONCLUSION:

PLS regression has developed significantly since its first introduction. The nonlinear format of the data found in the field of chemical engineering was the impetus behind the development of nonlinear PLS methods. In this paper we have proposed a robust version of the quadrilateral nonlinear PLS, to overcome the problems caused by outliers in the hybrid form of the quadrilateral PLS algorithm and BACON algorithmin order. Our method outperformed quad PLS for both real and simulated data.

REFERENCES:

1. Billor N., HadiA. S., and VellemanP. F., “BACON: blocked adaptive computationally efficient outlier nominators,” Computational Statistics and Data Analysis, vol. 34, no. 3, pp. 279–298, 2000.
2. Cherkassky V., GehringD., and MulierF., “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.
3. De Jong S., “PLS shrinks,” Journal of Chemometrics, vol. 9, no. 4, pp. 323–326, 1995.
4. Goutis C., “Partial least squares algorithm yields shrinkage estimators,” The Annals of Statistics, vol. 24, no. 2, pp. 816–824, 1996.
5. Hubert M. and BrandenK. V., “Robustmethods for partial least squares regression,” Journal of Chemometrics, vol. 17, no. 10, pp. 537–549, 2003.

6. KondylisA. and HadiA. S., "Derived components regression using the BACON algorithm," Computational Statistics & Data Analysis, vol. 51, no. 2, pp. 556–569, 2006.
7. Meng Z., ZhangS. and YanhY. et al., "Nonlinear partial least squares for consistency analysis of meteorological data," Mathematical Problems in Engineering, vol. 2015, Article ID 143965, 8 pages, 2015.
8. Wold S., Kettaneh-WoldN., and SkagerbergB., "Nonlinear PLS modeling," Chemometrics and Intelligent Laboratory Systems, vol. 7, no. 1-2, pp. 53–65, 1989.
9. Wold S., "Nonlinear partial least squares modelling. II. Spline inner relation," Chemometrics and Intelligent Laboratory Systems, vol. 14, no. 1–3, pp. 71–84, 1992.
10. Wold H., "Soft modelling 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.