مجلة الغري للعلوم الاقتصادية والادارية

مجلد (21) عدد (2) 2025

Improving the Performance of Robust Partial Least Squares Regression Using an Iterative Approach

Mahammad Mahmoud Bazid
Salahaddin University, College of Administration and Economics,
Mahammad.bazid@su.edu.krd

Prof. Dr. Taha Hussein Ali
Salahaddin University, College of Administration and Economics
taha.ali@su.edu.krd

Abstract

The robust partial least squares regression method provides a solution to outliers and noise in the estimated models by maximizing the explanation ratios of the independent and dependent variables (determination coefficient). In this article, three proposed methods were presented to deal with outliers or noise data and coefficient estimation accuracy of the partial least squares regression model. The first depends on the iterative method, which determines outliers and estimates them using the initial estimated values and the mean square error, as well as determining the optimal value that gives the least mean squares error for the partial least squares regression model. The second (Robust-Iteration) and third (Iteration-Robust) proposed methods rely on hybrid estimators of the iteration and robust approaches, that maximize the explanation ratios in the independent and dependent variables while minimizing the mean squared error. Simulation results and real data from chemical experiments (the quality of a chemical product based on various physicochemical properties) demonstrated the efficiency and accuracy of the proposed methods in handling outliers and noise in the data compared with the partial least squares regression method.

Keywords*:* Partial Least Squares Regression, Robust Partial Least Squares Regression, Outliers, noise data, and Residuals*.*

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

تحسين أداء الانحدار الجزئي للمربعات الصغرى الحصين باستخدام أسلوب تكراري

محمد محمود بازيد
جامعة صلاح الدين، كلية الادارة والاقتصاد
Mahammad.bazid@su.edu.krd

أ. د. طه حسين علي
جامعة صلاح الدين، كلية الادارة والاقتصاد
taha.ali@su.edu.krd

المستخلص

توفر طريقة الانحدار الجزئي للمربعات الصغرى القوية حلاً للضوضاء والقيم الشاذة في النماذج المقدرة من خلال تعظيم نسب التفسير للمتغيرات المستقلة والتابعة (معامل التحديد). وقد تم تقديم ثلاث طرق مقترحة لمعالجة مشكلة القيم الشاذة أو ضوضائية البيانات ودقة المعاملات المقدرة لنموذج الانحدار الجزئي للمربعات الصغرى. تعتمد الطريقة الأولى على الطريقة التكرارية، والتي تحدد القيم المتطرفة وتقديرها باستخدام القيم المقدرة الأولية ومتوسط خطأ المربعات، بالإضافة إلى تحديد القيمة المثلى التي تعطي أقل خطأ متوسط مربعات لنموذج الانحدار الجزئي للمربعات الصغرى. تعتمد الطريقتان المقترحتان الثانية (الحصين-التكراري) والثالثة (التكراري-الحصين) على مقدرين هجينين للتكرارية والأساليب الحصينة، والتي تعظم نسب التفسير في المتغيرات المستقلة والتابعة مع تقليل متوسط خطأ المربعات. وقد أظهرت نتائج المحاكاة والبيانات الحقيقية من التجارب الكيميائية (جودة المنتج الكيميائي بناءً على خصائص فيزيائية كيميائية مختلفة) كفاءة ودقة الطرق المقترحة في التعامل مع القيم الشاذة والضوضاء في البيانات مقارنة بطريقة الانحدار الجزئي للمربعات الصغرى الحصين.

الكلمات المفتاحية: الانحدار الجزئي للمربعات الصغرى، الانحدار الجزئي للمربعات الصغرى الحصين، القيم الشاذة، الضوضائية والبواقي.

1. Introduction

A multivariate statistical method called partial least squares analysis makes comparing many explanatory and response variables possible. One of the covariance-based statistical techniques known as structural equation modelling, or SEM, is partial least squares (Ali et al 2023). It was created to handle multiple regression when there is multicollinearity, missing

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

values, or a short sample size. Simulations and actual data have been used to illustrate partial least squares regression. It has gained much popularity in hard sciences, particularly chemistry and chemometrics, where many correlated variables and few observations provide a significant challenge. Even though data has comparable issues, its use in marketing has been more restricted. PLSR is especially beneficial when several variables are closely related since it may minimize the dimensionality of the data while retaining the critical information required for prediction. Cross-validation or the use of an independent test set is often used to establish the number of components to use in PLSR. Over the years, numerous improvements and processes have been developed, resulting in the technique's widespread use in chemometrics and significant documentation in literature. Studies show that PLSR often requires fewer components than PCR to achieve high prediction accuracy. It is conceptually based on maximum likelihood estimates and likelihood ratio tests, and it uses orthogonal scores and weights to determine the number of relevant components (GELADI & KOWALSKI, 1986). The flexibility of PLSR allows it to handle complicated data structures, including multi-way arrays. All things considered, PLSR is a valuable approach for regression analysis, particularly when multicollinearity is a concern. In this field, research and application are currently underway (Ali & Saleh, 2022). Robust PLSR, the traditional PLSR, is designed to be more resilient to outliers in data. Researchers commonly use PLSR in cases with numerous predictor variables and potential multicollinearity because it generates new components (latent variables) that capture the maximum covariance between predictors and responses. Outliers, however, have the potential to significantly skew results

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

and reduce model reliability, rendering standard PLSR techniques vulnerable to their effects. To get around this, robust PLSR employs strategies that reduce the impact of anomalous findings. Using robust estimators, such as S-estimators, to create a robust covariance matrix is a crucial tactic that allows the model to prioritize normal data points above outliers. (González et al., 2009). Select the right number of components to construct a partial least squares regression model that optimizes the relation between the independent and dependent variables in the covariance matrix. This technique produces predictions for the beginning values and residuals of the dependent variables.

In this article, three proposed methods were presented to deal with outliers or noise data and coefficient estimation accuracy of the partial least squares regression model. The first depends on the iterative method, which determines outliers and estimates them using the initial estimated values and the mean square error, as well as determining the optimal value that gives the least mean squares error for the partial least squares regression model. The second (Robust-Iteration) and third (Iteration-Robust) proposed methods rely on hybrid estimators of the iteration and robust approaches, that maximize the explanation ratios in the independent and dependent variables while minimizing the mean squared error.

2. Partial Least Squares Regression

Partial Least Squares (PLS) is a comprehensive family of approaches for modelling links between sets of observable data using latent variables. It includes regression and classification tasks as well as dimension reduction methods and modelling tools. The basic premise of all PLS approaches is that the observed data is created by a system or process which is driven by

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

a limited number of latent (not directly observable or measured) variables. Projections of the data seen to its latent structure by way of PLS were created by Herman Wold (Wold, H. (1975). PLS has garnered a tremendous amount of interest in chemometrics. This method has become a typical instrument for handling a broad range of chemical data challenges. The success of PLS in chemometrics led to a variety of applications in various scientific domains like bioinformatics, food research, medicine, pharmacology, social sciences, and physiology–to mention just a few (Ali, 2018). The core ideas of PLS offer an overview of its application to diverse data analysis situations. Our purpose is to give a brief introduction, that is, a beneficial guide for everyone who is concerned with data analysis. In its general version, PLS builds orthogonal score vectors (sometimes called latent vectors or components) by optimizing the covariance between multiple sets of variables. PLS dealing with two blocks of variables is examined, however, the PLS extensions to describe relations among a greater number of sets exist. PLS is analogous to Canonical Correlation Analysis (CCA) where latent vectors with the greatest correlation are retrieved. There are many PLS strategies to extract latent vectors, and each of them gives birth to a variety of PLS. Furthermore, it works well in a range of fields, such as genomics and chemometrics, when there are more predictors than data ($p > n$). Handling noisy or poor datasets is another major benefit. By focusing on the most significant factors, PLSR minimizes the effect of noise and gives accurate forecasts even in demanding conditions. A further essential component of PLS is the capacity to display high-dimensional data using the collection of extracted latent variables. The diagnostic function of PLS tools focuses on

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

score and loading plots allowing us to better grasp data structure and assess existing links across data sets but also to discover outliers in the measured data. Successful application of PLS on regression difficulties connected with numerous (Rosipal and Krämer, 2005).

2.1. Model Construction

The nonlinear iterative partial least squares (NIPALS) algorithm's characteristics provide the foundation of the PLS model. The data matrix may be represented by the score matrix, as was shown in the PCR section. A regression between the scores for the X and Y blocks would make up a basic model. The outside relations (X and Y blocks separately) and the inside relation (connecting both blocks) make up the PLS model. According to the PCA section, the **X** block's outer relation is (Pirouz, 2006):

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \quad (1)$$

$$\mathbf{Y} = \mathbf{UQ}' + \mathbf{F} \quad (2)$$

- **X** is a $n \times m$ predictor matrix.
- **Y** is a $n \times p$ response matrix.
- **T** and **U** are $n \times 1$ matrices that are, as well, projectors of **X** (the **X** score, component or factor matrix) and projectors of **Y** (the **Y** scores).
- **P** and **Q** are, accordingly, $\boldsymbol{m \times 1}$ and $\boldsymbol{p \times 1}$ loading matrices
- matrices **E** and **F** are the error terms, supposed to be independent and symmetrically distributed random normal variables (Geladi and Kowalski, 1986).

3. Robust Partial Least Squares Regression

The use of a robust approach should be taken into consideration if outliers are likely to appear in the data. The primary benefit of the robust PLSR approach that is being described, which includes recurrent double cross-validation, is that it eliminates the need for outlier discovery before the model is created and provides a realistic estimate of the model's future

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

performance (Beyaztas and Shang, 2022). The robust approach is almost as effective as the traditional approach when there are no outliers in the data. It is shown that RPLSR accurately estimates the genuine underlying model parameters for fabricated data; in particular, when aberrant observations are included in the calibration data, the resilient techniques perform noticeably better than traditional PLS (Ali & Awaz, 2017). Therefore, when it comes to non-outliers, robust models outperform the classical models. However, identifying outliers in fresh data is still a challenge. One simple method is to compute the robust weights Wi X after performing robust autoscaling in X for both the new data and the data used to create the model. Other robust outlier identification techniques are more advanced (Gil and Romera, 1998).

3.1. The primary benefits of using robust Partial Least Squares Regression

Robust Partial least squares regression techniques are explicitly formulated to reduce the impact of outliers in the dataset. Classical PLSR is susceptible to outliers, which may substantially skew the findings and result in suboptimal model performance. Robust PLSR techniques, such as the suggested PLS-Smult, have superior prediction capability in the presence of outliers in the data (Hubert et al. 2008). Simulation experiments demonstrate that robust techniques surpass standard PLSR for efficiency, goodness-of-fit, and predictive capability, particularly in polluted datasets. Robust PLSR techniques provide superior fitting to standard data points notwithstanding the presence of outliers. This is essential for preserving the model's integrity when the data is not entirely pristine (Pensia et al., 2024).

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

3.2 methodology

1. In PLSR, choosing the weights W maximizes the covariance:

$$MAX_{wie} = covaiance(X * W, Y) \quad (3)$$

2. Robust PLSR has a regularization penalty to prevent overfitting. The optimization becomes (Hubert & Branden, 2003):

$$\mathbf{MAX_W Cov(XW, Y) - \lambda \|W\|^2} \quad (4)$$

In this case, $\boldsymbol{\lambda}$ regulates the penalty's strength; simpler models result from greater $\boldsymbol{\lambda}$.

3. The ultimate related may be written as follows:

$$\mathbf{Y = TQ + E} \quad (5)$$

$\mathbf{T = XW}$ are the components (latent variables), **Q** relates **T** to **Y**, **E** residual error.

Latent variables may pick up unimportant patterns (noise) in **X** if regularization is not used. The approach gives priority to strong, broadly applicable patterns via regularization. Using the retrieved latent variables, the trained RPLSR model forecasts new **Y** values from the provided **X** (Serneels et al., 2005).

3.3. Different Approaches to Robust PLS Iteratively Reweighted Least Squares

Assume that we want to determine y's multiple regression on X. The following is the IRLS approach.

1. Determine the regression coefficient's starting value.

$$\mathbf{\hat{B} = (X'X)^{-1}X'Y} \quad (6)$$

2. Determine the regression's residuals.

$$\mathbf{r = y - X\hat{B}} \quad (7)$$

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

3. select a new regression coefficient for use weights (Gil and Romera, 1998).

$$\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{\Phi}'\mathbf{\Phi}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Phi}'\mathbf{\Phi}\mathbf{Y} \quad (8)$$

4. Outliers and noise:

The outliers may significantly affect statistical studies, producing exaggerated variances, skewed parameter values, and false conclusions (Phillips and Eyring, 1983). For example, conventional techniques such as Least Squares (LS) regression are very susceptible to outliers. The fitted model may be significantly changed by even one outlier, making it untrustworthy. This sensitivity results from LS's excessive weighting of extreme values via minimizing the sum of squared errors. Finding outliers in multiple variables dealing with multivariate data makes outlier detection more difficult (Ali et al., 2024). There has been extensive discussion of the typical diagnostics for outlier detection during the calibration stage of a multivariate calibration experiment. These include diagnostics for "outside" the model space, like as an F-test on the spectral residuals, and those for "inside" the model space, such as sample leverage or Mahalanobis distance proportional to sample leverage. The difference between known and anticipated concentrations may be determined during the calibration phase and utilized as an additional outlier detection diagnostic. Both studentized and leverage-corrected concentration residuals may be constructed by scaling the concentration residuals by their standard deviation and a function of the sample leverage (Aggarwal & Aggarwal, 2017). These diagnostics may perform badly when there are numerous outliers, but they may work well when there is only one outlier, especially if the outlier is removed during calculation as in cross-validation (Pell, 2000). In multidimensional space,

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

outliers could be difficult to recognize. Simple two-dimensional plots of the dependent and independent variables may be generated with just one independent variable, and outliers are clear to recognize. Plotting the least squares residuals against variables like the fitted response or in serial order to find outliers is a routine practice when the model comprises multiple independent variables (Cummins and Andrews, 1995).

5. Proposed Methods

We summarize the three proposed methods for treating outliers and noise in the following:

First Proposed:

- Estimate a partial least squares regression model that maximizes the covariance matrix between the independent and dependent variables after choosing many suitable components to obtain predictions of the initial values of the dependent variable and the residual.
- Identifying outliers y(o) from the standard residuals of a partial least squares regression model that are outside an interval (∓ 2.5) or the largest residual value.
- Calculate the initial average of mean Squares Error **(AMSE)** of the model from the following formula:

$$\mathbf{AMSE} = \sum_{k=1}^{2} \sum_{j=1}^{p+1} (\mathbf{MSE}(k,j)/2(p+1) \qquad (9)$$

The number of principal components is p. MSE includes two parts, the mean square error of X (MSEx) which measures how the model explained the variation in the independent variables, and the mean square error of Y (MSEy) which measures the accuracy of the model:

$$\mathbf{MSEx} = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(-\hat{x}_{ij}\right)^2 \qquad (10)$$

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

MSEx quantifies the error between the original x and the reconstructed x from the model.

$$\mathbf{MSEy} = \frac{\mathbf{1}}{\boldsymbol{np}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \left(-\hat{\boldsymbol{y}}_{\boldsymbol{ij}}\right)^{\mathbf{2}} \qquad (\mathbf{11})$$

MSEy quantifies the error between the actual value y and the predicted by the PLSR model.

- Estimate outliers using the following equation:

$$\mathbf{Y(o)} = \hat{\mathbf{y}}\mathbf{(o)} - \mathbf{Residual(o)} \qquad (\mathbf{12})$$

Residual (o) is the outlier residual, using **Y(o)** instead of $\hat{\mathbf{y}}\mathbf{(o)}$ with Y to estimate a PLS model and compute AMSE.

- If the AMSE value is greater than (0.001), then the outlier in equation (3) will be re-estimated and get AMSE for a new PLS model and so on until the AMSE is less than (0.001).
- Finally, the estimated values of the outliers with the least AMSE are used to create the PLS model.

Second Proposed:

The second proposed method is based on the hybrid method (Robust-Iteration) which uses a robust estimator (Savitsky-Golay filter using iterative reweighing in combination) to handle outliers and noise in data based on maximizing the explanation ratio of the independent and dependent variables as inputs to the iterative method that minimizes the AMSE as in the first proposal(Menon and Seelamantula, 2014).

Third Proposed:

The third proposed method is based on the hybrid method (Iteration-Robust) which uses the iterative process that minimizes the AMSE as in the first proposal as inputs a robust estimator to handle outliers and noise in data

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

based on maximizing the explanation ratio of the independent and dependent variables.

6. Simulation Study

To demonstrate the efficiency of the proposed methods and compare them with the robust method in handling noise and outliers in the PLSR model, a simulation was conducted by generating random data for the independent and dependent variables, and the addition of two outliers to the dependent variable.

6.1. First Experiment Simulation

The estimated and residual values of the PLSR model for the first simulation of the methods (PLSR, Robust PLSR, Iteration, Robust-Iteration, and Iteration-Robust) using 5 factors (n = 25 and m = 30) are shown in Figures 1 and 2. The values of the dependent variable for the generated data show two outliers (marked in red points). The outliers affect the PLSR method and produce unacceptably large residuals.

The Robust-PLSR method was robust against outliers and provided an increase in the explanation proportions for the independent (from 35.9589 to 61.9948) and dependent (from 85.4535 to 87.7472) variables while reducing the value of AMSE (from 1.5178 to 0.6084) as in Table 1.

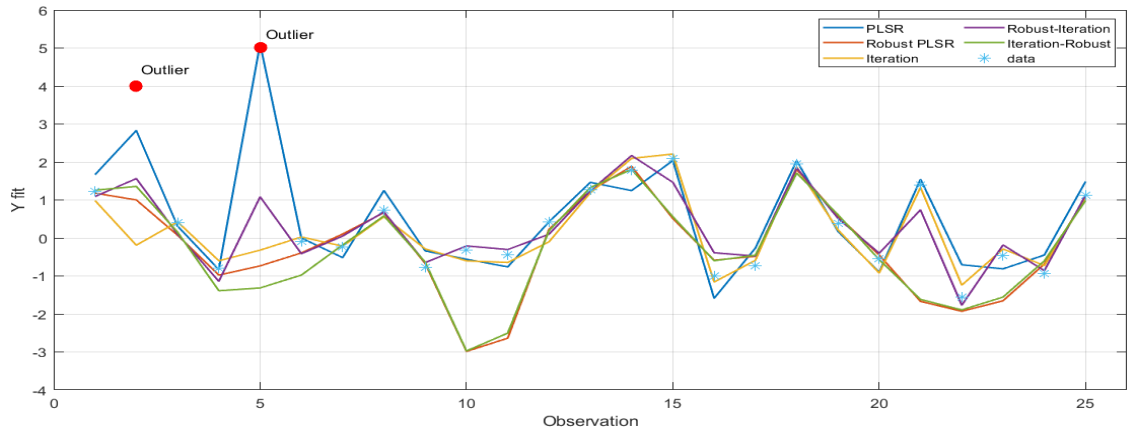
The first proposed method (Iteration-PLSR) is also robust against outliers and provided an increase in the explanation proportions for the independent (from 35.9589 to 38.4520) and dependent (from 85.4535 to 92.5245) variables while reducing the value of AMSE (from 1.5178 to 0.1091). The increase in the proportion of explanation of the independent variables was limited. Still, the decrease was large in AMSE, and this is logical in the mechanism of the iterative method in minimizing AMSE and does not focus

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

on maximizing the proportion of explanation, especially the independent variables, which is less important than the proportion of explanation for the dependent variables in the analysis of the PLSR model.

The second proposed method (Robust-Iteration) is also robust against outliers and provided an increase in the explanation proportions for the independent (from 35.9589 to 61.9809) and dependent (from 85.4535 to 86.1474) variables while reducing the value of AMSE (from 1.5178 to 0.1080), noting the big difference in reducing the value of AMSE compared to the robust method (from 0.6084 to 0.1080) and at the same time a significant increase in the explanation proportions.

The third proposed method (Iteration-Robust) is also robust against outliers and provided an increase in the explanation proportions for the independent (from 35.9589 to 63.2518) and dependent (from 85.4535 to 94.3926) variables while reducing the value of AMSE (from 0.6084 to 0.1080), noting the big difference in reducing the value of AMSE compared to the robust method (from 0.6084 to 0.5587) and at the same time a significant increase in the explanation proportions noting the small difference in reducing the value of AMSE compared to the robust method (from 0.6084 to 0.5587) and at the same time a significant increase in the explanation proportions.



مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

Figure 1. Estimated values of the dependent variable for the first experiment simulation

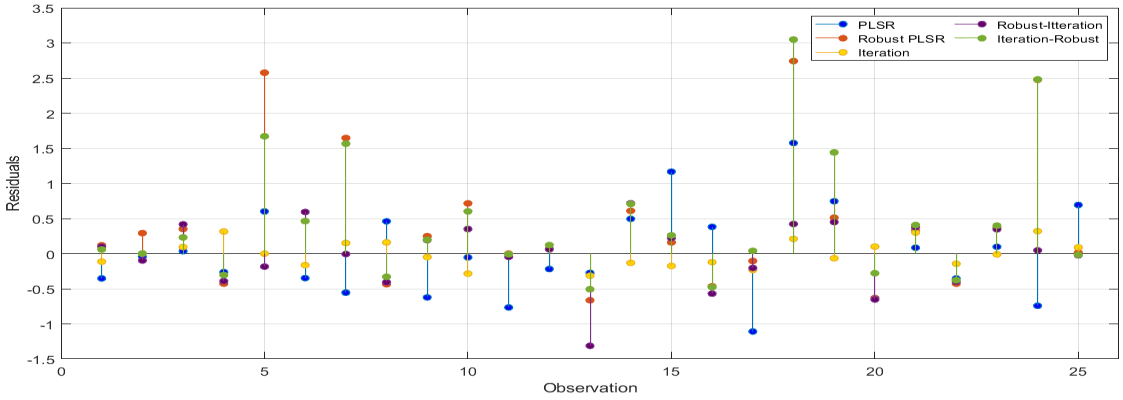


Figure 2. Residual values for the first experiment simulation

Table 1. The Results First Experiment Simulation

Method	R^2X	R^2Y	AMSE
Without Filter	35.9589	85.4535	1.5178
Robust	61.9948	87.7472	0.6084
Iteration	38.4520	92.5245	0.1091
Robust-Iteration	61.9809	86.1474	0.1080
Iteration-Robust	63.2518	94.3926	0.5587

6.2. Repeat Simulation Experiments

To compare the efficiency between proposed and traditional methods and the generalization of simulation results, the data generation experiments are repeated (1000) times. The simulation included several different sample sizes (25, 50, 75, and 100), and numbers of independent variables (30, 60, 90, and 120) using different numbers of principal components (5 and 10). The average simulation results are summarized in Tables 2-5:

Table 2. The Average Simulation Results (n = 25 and m = 30)

Method	Number of principal components	R^2X	R^2Y	MSE
Without Filter	5	35.5580	92.2240	1.3742
Robust		71.1660	89.3795	0.4936
Iteration		35.6919	96.6607	0.0610
Robust-Iteration		71.5301	86.1316	0.0630
Iteration-Robust		68.1663	89.5655	0.4618
Without Filter	10	61.9367	98.4562	1.0294
Robust		94.0308	97.6272	0.3945
Iteration		61.8635	99.6902	0.0993

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

Robust-Iteration	94.1640	92.3708	0.1033
Iteration-Robust	93.6213	98.2045	0.3479

Table 3. The Average Simulation Results (n = 50 and m = 60)

Method	Number of principal components	R^2X	R^2Y	MSE
Without Filter		18.7640	91.2817	2.5090
Robust		54.2390	86.5358	0.8495
Iteration	5	18.8749	94.1093	0.0240
Robust-Iteration		54.7919	82.1705	0.0261
Iteration-Robust		51.5203	86.4152	0.8343
Without Filter		34.5158	97.7965	2.1824
Robust		76.7634	96.8834	0.6529
Iteration	10	34.4380	99.0342	0.0322
Robust-Iteration		77.3408	90.8289	0.0346
Iteration-Robust		75.4982	97.1472	0.6414

Table 4. The Average Simulation Results (n = 75 and m = 90)

Method	Number of principal components	R^2X	R^2Y	MSE
Without Filter		12.7219	90.7463	3.7149
Robust		46.9228	85.1105	1.2118
Iteration	5	12.7744	92.8238	0.0146
Robust-Iteration		47.6250	80.7667	0.0164
Iteration-Robust		44.3053	84.5679	1.2081
Without Filter		23.8167	97.4078	3.3964
Robust		67.8119	96.6309	0.9771
Iteration	10	23.7860	98.6086	0.0184
Robust-Iteration		68.6285	90.6234	0.0204
Iteration-Robust		66.3667	96.6490	0.9770

Table 5. The Average Simulation Results (n = 100 and m = 120)

Method	Number of principal components	R^2X	R^2Y	MSE
Without Filter		9.6325	90.6681	4.9445
Robust		42.6572	83.8610	1.5823
Iteration	5	9.6718	92.2896	0.0105
Robust-Iteration		43.3303	79.3518	0.0119
Iteration-Robust		40.5091	83.5027	1.5806
Without Filter		18.1599	97.3582	4.6285
Robust		62.2057	96.1906	1.3185
Iteration	10	18.1597	98.3606	0.0127
Robust-Iteration		63.1931	90.1629	0.0145
Iteration-Robust		61.0509	96.2849	1.3181

6.3. Simulation Result Discussion

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

Simulation results show that the first proposed method (Iteration PLSR) has the lowest average of AMSE for all simulation cases, followed by the second proposed method (Robust-Iteration). The third proposed method (Iteration-Robust) has a lower average of AMSE than the robust method for all simulation cases. The robust method and the proposed methods address the problem of data noise and outliers and provide highly efficient estimators sorted by order of least AMSE average (Iteration PLSR, Robust-Iteration, Iteration-Robust, and Robust PLSR) as shown in Figure 3 for 100 iterations.

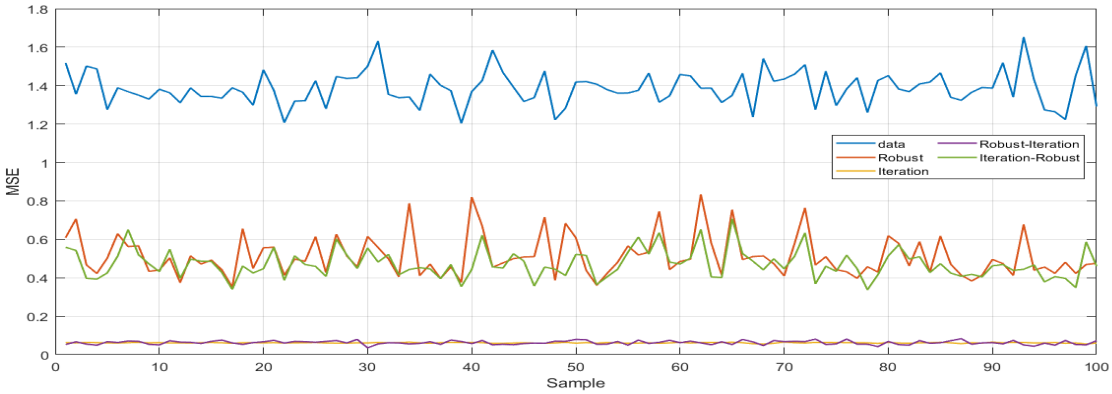


Figure 3. Average of AMSE for 100 iterations (when n = 25 and m = 30)

The explanation proportion R^2X of the independent variables increased for the robust and hybrid (Robust-Iteration and Iteration-Robust) methods because their techniques depend on increasing the explanation proportion, the results of the iterative PLSR method were close to the traditional method for all simulation cases. The robust and hybrid methods provide highly efficient estimators sorted by order of great R^2X average (Robust-Iteration, Robust PLSR, and Iteration-Robust PLSR) as shown in Figure 4 for 100 iterations.

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

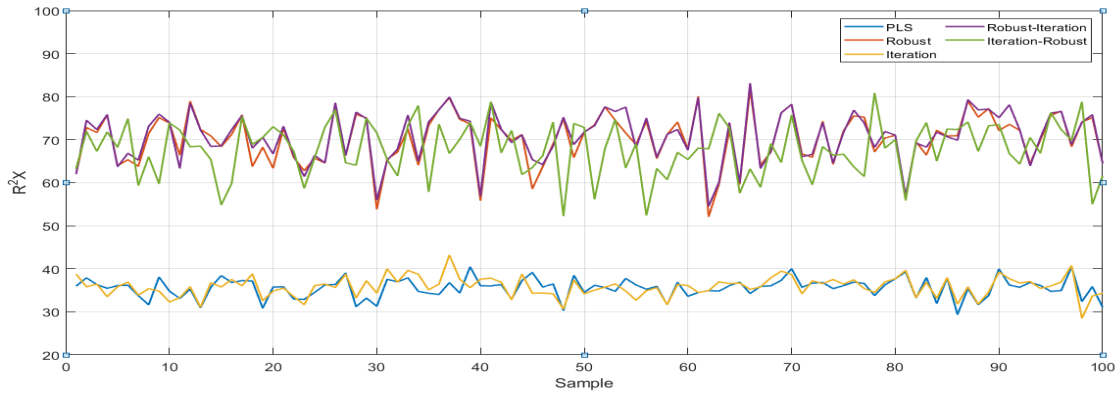


Figure 4. Average of R^2X for 100 iterations (when n = 25 and m = 30)

The explanation proportion R^2Y of the dependent variables increased for the iteration proposed method compared with other methods for all simulation cases. R^2Y has an importance greater than R^2X in the analysis of PLSR models, which confirms the efficiency of the proposed iterative method in processing data noise and outliers and providing a greater explanation proportion than other methods (as shown in Figure 5 for 100 iterations) in addition to reducing the AMSE average.

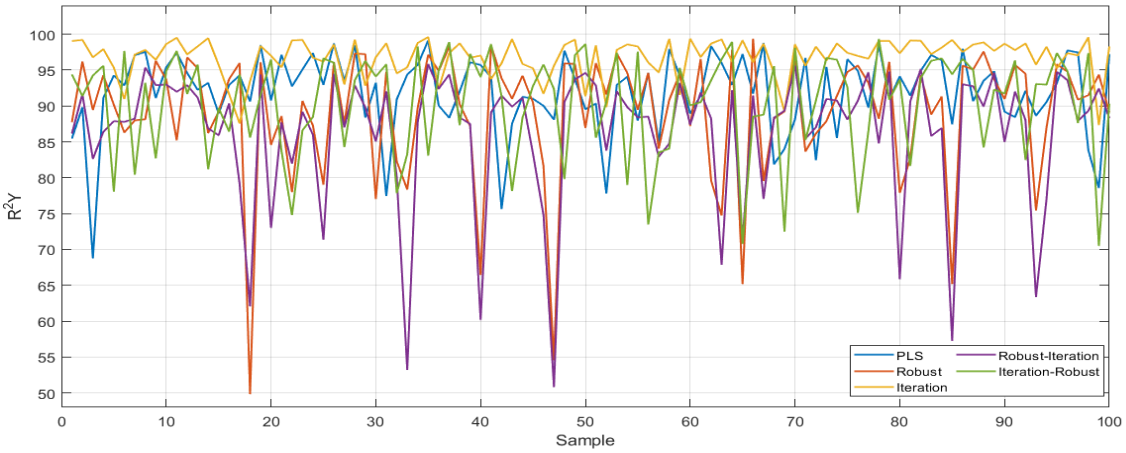


Figure 5. Average of R^2Y for 100 iterations (when n = 25 and m = 30)

Increasing the number of principal components resulted in lower values of the AMSE average and increases for R^2X and R^2Y of all methods used and for all simulation cases. Increasing the number of observations and

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

independent variables resulted in greater values of the AMSE average and decreasing for R^2X and R^2Y of all methods used and for all simulation cases

7. Real Data

Real data represents the quality of a chemical product (dependent variable) based on various physicochemical properties as independent variables (Cortes et al. 2009). The application includes 10 observations, and 11 Independent variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free SO2, total SO2, density, pH, sulphates, and alcohol. The dependent variable is a quality score from 0 to 10.

To detect outliers, the PLSR model was estimated, and the residuals were calculated. All values were within the interval (± 2.5), indicating there are no outliers in the data (as in Figure 6), but there may be noise that can be distinguished through analysis.

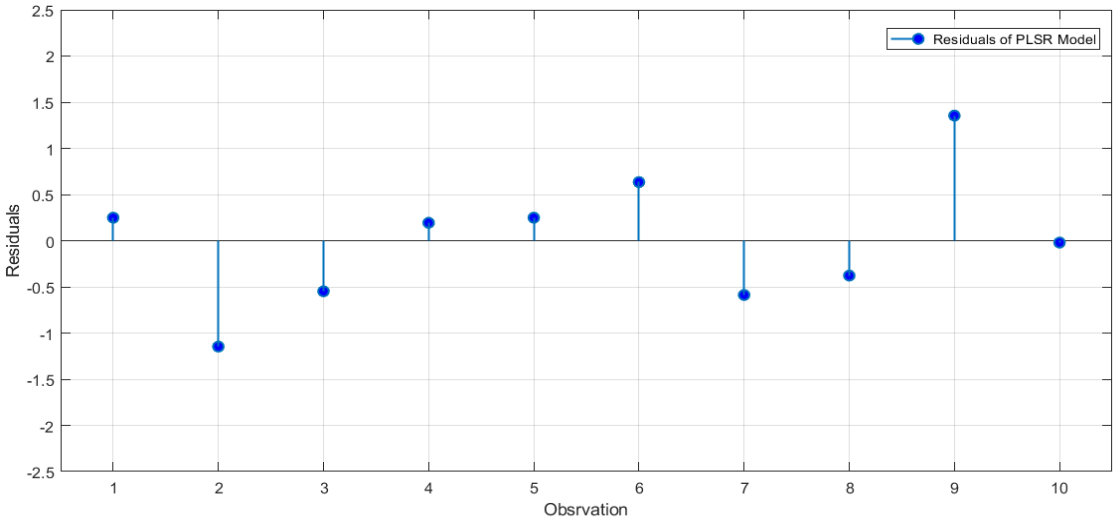


Figure 6. Residuals of the PLSR Model for the quality of a chemical product

Figure 7 shows the actual and estimated values for the quality of a chemical product (without outliers) from the five models and shows the large variation

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

in estimated values depending on the method used to estimate the PLSR model parameters (using four principal components).

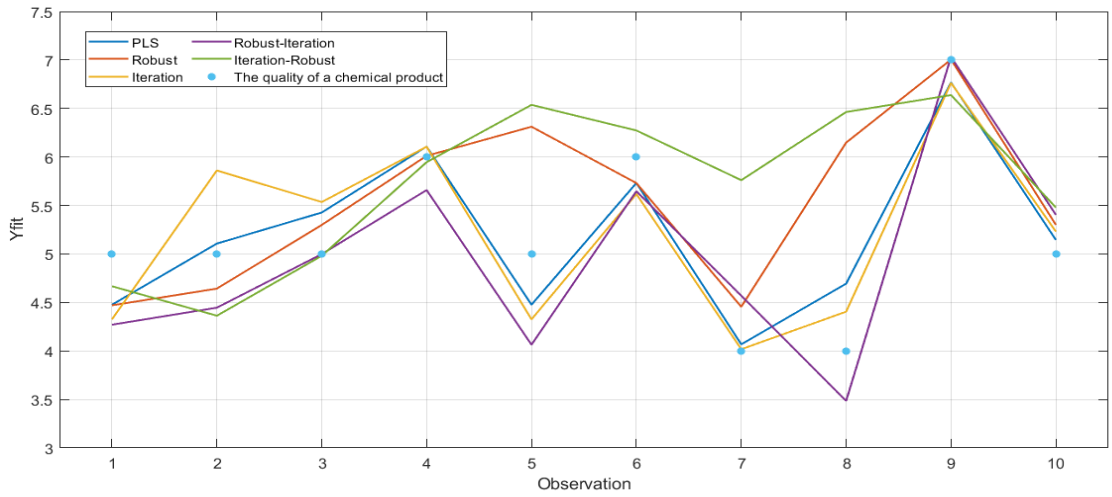


Figure 7. Estimated values for the quality of a chemical product

Principal components (1-8) were used, and the comparison criteria were calculated as in Table 6. Four principal components were identified that were appropriate for this data and had an explanation proportion R^2Y greater than 80% for all methods used (the residuals shown in Figure 8. The Robust-PLSR method was robust against noise and provided an increase in the explanation proportions for the independent (from 99.9578 to 99.9806) and dependent (from 81.7622 to 95.3366) variables while decreasing the value of AMSE (from 21.6199 to 20.5099). The result is logical because the robust PLSR method focuses on increasing the explanation ratio and does not reduce the AMSE. The first proposed method (Iteration-PLSR) is also strong against noise and provided an increase in the explanation proportions for the independent (from 99.9578 to 99.9637) and dependent (from 81.7622 to 86.5515) variables while reducing the value of AMSE (from 21.6199 to 0.2550). The increase in the proportion of explanation of the independent variables was limited. Still, the decrease was large in AMSE, and this is

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

logical in the mechanism of the iterative method in minimizing AMSE and does not focus on maximizing the proportion of explanation, especially the independent variables. The second proposed method (Robust-Iteration) is also robust against noise and provided an increase in the explanation proportions for the independent (from 99.9578 to 99.9826) and dependent (from 81.7622 to 82.5689) variables while reducing the value of AMSE (from 21.6199 to 0.2661), noting the big difference in reducing the value of AMSE compared to the robust method (from 20.5099 to 0.2661). The third proposed method (Iteration-Robust) is also strong against noise and provided an increase in the explanation proportions for the independent (from 99.9578 to 99.9940) and dependent (from 81.7622 to 98.3917) variables while increasing the value of AMSE (from 21.6199 to 29.8144), noting a significant increase in the explanation proportion R^2Y.

Table 6. PLSR Model Results

Method	Number of principal components	R^2X	R^2Y	MSE
Without Filter		97.2387	40.7418	53.4096
Robust		97.9777	81.3572	49.5859
Iteration	1	97.2411	79.2258	0.1653
Robust-Iteration		97.9512	27.8298	0.1605
Iteration-Robust		90.7149	2.0022	29.4958
Without Filter		99.3699	48.8577	35.8883
Robust		99.0461	62.7041	30.5487
Iteration	2	99.5181	85.4491	0.1792
Robust-Iteration		98.9247	81.5484	0.1760
Iteration-Robust		98.9349	70.9867	31.0834
Without Filter		99.8083	67.4772	26.9966
Robust		99.9440	84.9304	20.5015
Iteration	3	99.7700	83.4951	0.2190
Robust-Iteration		99.9474	82.3319	0.2167
Iteration-Robust		99.9312	96.8229	46.3394
Without Filter		**99.9578**	**81.7622**	**21.6199**
Robust		**99.9806**	**95.3366**	**20.5099**
Iteration	4	**99.9637**	**86.5515**	**0.2550**
Robust-Iteration		**99.9826**	**82.5689**	**0.2661**
Iteration-Robust		**99.9940**	**98.3917**	**29.8144**

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

Without Filter		99.9971	92.5782	18.0218
Robust		99.9984	99.7944	20.1219
Iteration	5	99.9971	92.7806	0.3062
Robust-Iteration		99.9984	77.6060	0.2925
Iteration-Robust		99.9990	99.8129	30.6559
Without Filter		99.9987	95.8079	15.4497
Robust		99.9998	99.9748	24.7527
Iteration	6	99.9987	96.6076	0.3744
Robust-Iteration		99.9997	65.8603	0.3377
Iteration-Robust		99.9999	99.8869	28.2983
Without Filter		99.9999	98.0650	13.5194
Robust		100.0000	100.0000	27.9707
Iteration	7	99.9999	100.0000	0.4882
Robust-Iteration		100.0000	67.1727	0.4168
Iteration-Robust		100.0000	83.3125	5.9165
Without Filter		100.0000	100.0000	12.0173
Robust		100.0000	100.0000	31.2977
Iteration	8	100.0000	100.0000	0.6667
Robust-Iteration		100.0000	56.9519	0.6666
Iteration-Robust		100.0000	100.0000	27.1532

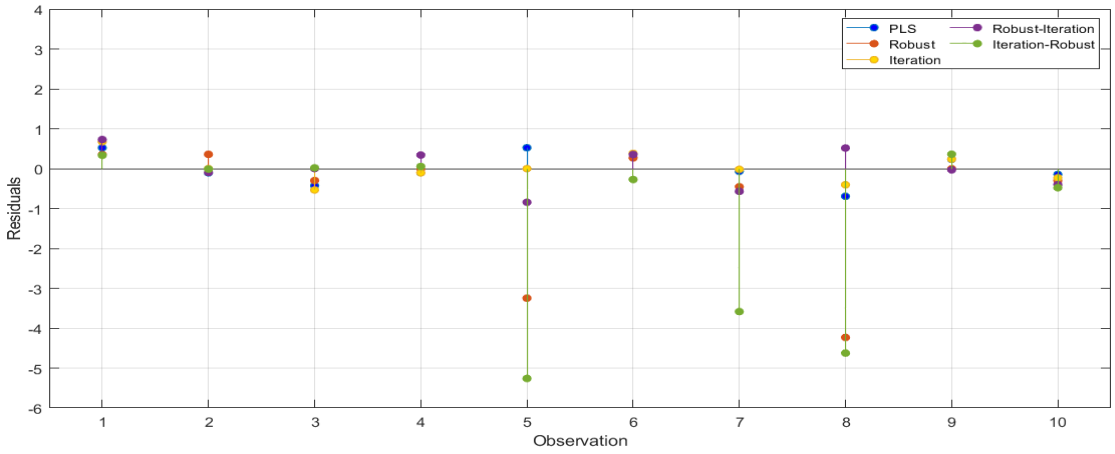


Figure 8. Residuals of models for the quality of a chemical product

Also, from the results of Table 6, it is noted that there is a large difference between the results of the five methods according to the number of principal components used in the analysis. Therefore, the number of principal components used can be determined according to the best results provided by that method (minimum AMSE and R^2Y greater than 80%).

Table 7. Best PLSR Models Results

Method	Number of principal components	R^2X	R^2Y	MSE

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

Without Filter	8	100.0000	100.0000	12.0173
Robust	5	99.9984	99.7944	20.1219
Iteration	2	99.5181	85.4491	0.1792
Robust-Iteration	2	98.9247	81.5484	0.1760
Iteration-Robust	7	100.0000	83.3125	5.9165

The best results summarized in Table 7 show that the traditional method was better than the robust method when using 8 principal components with noise in the data and no outliers. Based on the AMSE criterion, the proposed methods were the best in handling the data noise and the accuracy of the estimated parameters of the PLSR model with fewer principal components used in the analysis and according to the order of preference (Robust-Iteration, Iteration, and Iteration-Robust).

To illustrate the effect of outliers on the PLSR model analysis, two outliers were substituted for the original values of the quality of a chemical product, as shown in Figure 9.

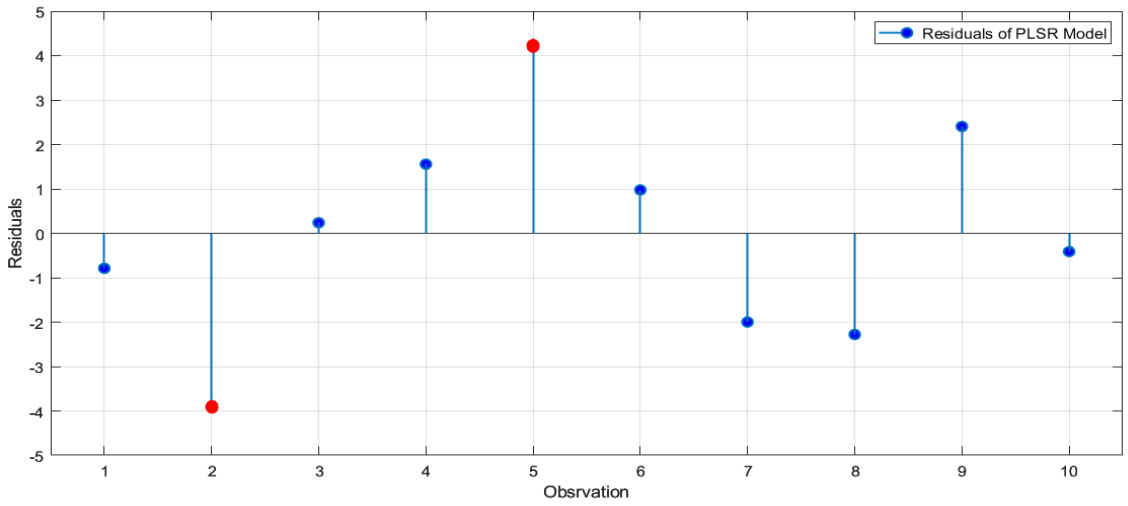


Figure 9. Residuals of models for the quality of a chemical product (with outliers)

The PLSR model was estimated (using one principal component), and the residuals were calculated. The second and fifth residual values were outside

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

the interval (±2.5) so they are considered outliers. Figure 10 shows the actual and estimated values for the quality of a chemical product (with outliers) from the five models and the large variation in estimated values (using six principal components).

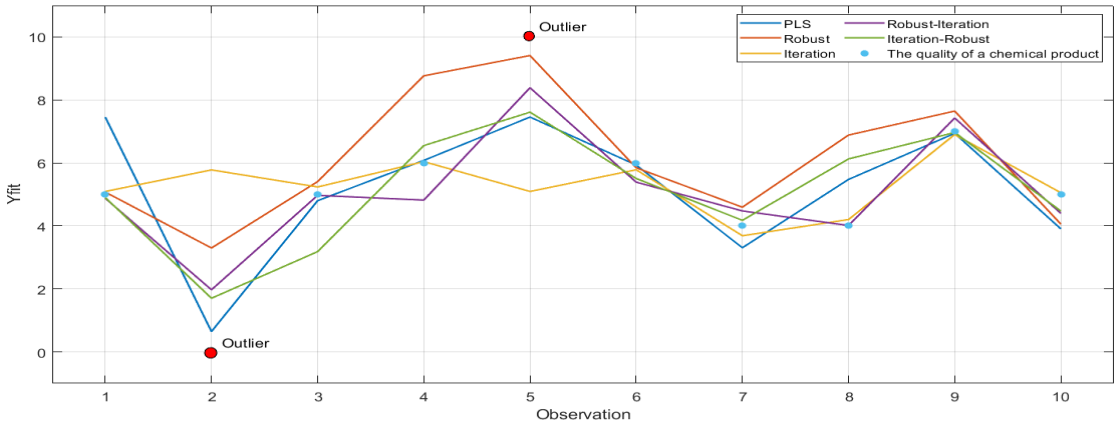


Figure 10. Estimated values for the quality of a chemical product (with outliers)

Principal components (1-8) were used, and the comparison criteria were calculated as in Table 8.

Table 8. PLSR Model Results (with outliers)

Method	Number of Factors	R^2X	R^2Y	MSE
Without Filter	1	97.1631	9.3118	55.8920
Robust		97.5697	26.0653	47.8410
Iteration		97.2411	79.2258	0.0982
Robust-Iteration		97.5741	1.7856	0.0987
Iteration-Robust		97.5967	10.2053	35.6451
Without Filter	2	99.4895	25.1263	38.1560
Robust		98.5966	53.1498	25.1009
Iteration		99.5181	85.4491	0.1315
Robust-Iteration		99.0771	91.2017	0.1244
Iteration-Robust		98.6907	86.5735	23.2478
Without Filter	3	99.8459	38.2313	29.1015
Robust		99.9495	64.5901	22.9897
Iteration		99.7700	83.4951	0.2015
Robust-Iteration		99.9321	98.2032	0.1429
Iteration-Robust		99.9473	90.2696	23.0041
Without Filter	4	99.9157	62.3202	23.5157
Robust		99.9623	80.3598	12.4473

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

Iteration		99.9637	86.5515	0.2553
Robust-Iteration		99.9637	94.7132	0.1975
Iteration-Robust		99.9849	98.3837	23.8985
Without Filter		99.9971	68.4267	19.7485
Robust		99.9964	83.6249	10.2156
Iteration	5	99.9971	92.7806	0.3111
Robust-Iteration		99.9964	98.6058	0.2511
Iteration-Robust		99.9991	99.8666	30.8004
Without Filter		**99.9990**	**70.7737**	**17.0476**
Robust		**99.9997**	**96.5006**	**10.4503**
Iteration	6	**99.9997**	**96.6075**	**0.3723**
Robust-Iteration		**99.9991**	**87.6478**	**0.3193**
Iteration-Robust		**99.9993**	**88.9315**	**9.4471**
Without Filter		99.9996	77.8482	14.9964
Robust		100.000	99.9998	35.2513
Iteration	7	99.9997	97.5985	0.4807
Robust-Iteration		100.000	40.0492	0.4976
Iteration-Robust		100.000	82.7793	5.9307
Without Filter		100.000	78.2986	13.3996
Robust		100.000	99.9999	27.1440
Iteration	8	100.000	100.000	0.6565
Robust-Iteration		100.000	85.5634	0.6459
Iteration-Robust		100.000	100.000	27.1550

Six principal components were identified that were appropriate for this data and had an explanation proportion R^2Y greater than 80% for all methods used (the residuals shown in Figure 11. The Robust-PLSR method was robust against outliers and provided an increase in the explanation proportions for the independent (from 99.9990 to 99.99997) and dependent (from 70.7737 to 96.5006) variables while decreasing the value of AMSE (from 17.0476 to 10.4503). The first proposed method (Iteration-PLSR) is also strong against outliers and provided an increase in the explanation proportions for the independent (from 99.9990 to 99.99997) and dependent (from 70.7737 to 96.6075) variables while reducing the value of AMSE (from 17.0476 to 0.3723). Still, the decrease was large in AMSE, and this is logical in the mechanism of the iterative method in minimizing AMSE and does not focus on maximizing the proportion of explanation, especially the dependent

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

variables. The second proposed method (Robust-Iteration) is also robust against outliers and provided an increase in the explanation proportions for the independent (from 99.9990 to 99.9991) and dependent (from 70.7737 to 87.6478) variables while reducing the value of AMSE (from 17.0476 to 0.3193), noting the big difference in reducing the value of AMSE compared to the robust method (from 10.4503 to 0.3193). The third proposed method (Iteration-Robust) is also strong against outliers and provided an increase in the explanation proportions for the independent (from 99.9990 to 99.9993) and dependent (from 70.7737 to 88.9315) variables while decreasing the value of AMSE (from 17.04503 to 9.4471).

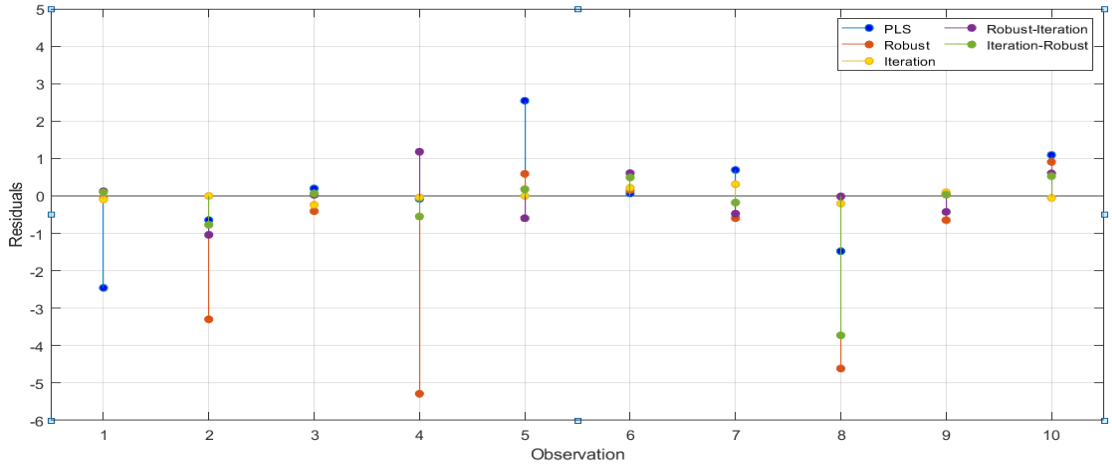


Figure 11. Residuals of models for the quality of a chemical product (with outliers)

The five methods with and without outliers, and with noise, provided results of different efficiency depending on the number of principal components used in the analysis. Depending on the 6 principal components, the robust method and the proposed methods address the problem of data noise and outliers and provide highly efficient estimators sorted by order of least AMSE (Robust-Iteration, Iteration PLSR, Iteration-Robust, and Robust PLSR).

8. Conclusions

1. The three proposed methods address the problem of outliers and noise in PLSR model data.
2. The proposed methods gave better results than the robust PLSR method.
3. The proposed methods provide highly efficient estimators sorted by order of least AMSE (Robust-Iteration, Iteration PLSR, and Iteration-Robust).
4. Increasing the number of observations and independent variables resulted in greater values of the AMSE average and decreased R^2X and R^2Y of all methods used and for all simulation cases.
5. Increasing the number of principal components resulted in lower values of the AMSE average and increases for R^2X and R^2Y of all methods used and for all simulation cases.
6. There is a significant improvement in PLSR models in analyzing the quality of a chemical product using the proposed methods.
7. The proposed methods proved to be more efficient than the conventional method even in the absence of outliers in the analysis of the chemical product quality.

Authors Declaration: Conflicts of Interest: None

References

1. Aggarwal, C.C. and Aggarwal, C.C., 2017. An introduction to outlier analysis (pp. 1-34). Springer International Publishing.
2. Ali, T., Hussein, T. H., Hayawi, H. A. A., & Botani, D. S. I. (2023). Estimation of the bandwidth parameter in Nadaraya-Watson kernel non-parametric regression based on universal threshold level. Communications in Statistics-Simulation and Computation, 52(4), 1476-1489. https://doi.org/10.1080/03610918.2023.2179 912

3. Ali, T. H., & Saleh, D. M. (2022). Proposed hybrid method for wavelet shrinkage with robust multiple linear regression model: With simulation study. Qalaaai Zanist Journal, 7(1), 920-937.
4. Ali, T. H. (2018). Solving multi-collinearity problem by ridge and eigenvalue regression with simulation. Journal of Humanity Sciences, 22(5), 262-276.
5. Ali, T. H., & M., A. S. (2017). Uses of Waveshrink in detection and treatment of outlier values in linear regression analysis and comparison with some robust methods. Journal of Humanity Sciences, 21(5), 38-61.
6. Ali, T. H., Sedeeq, B. S., Saleh, D. M., & Rahim, A. G. (2024). Robust multivariate quality control charts for enhanced variability monitoring. Quality and Reliability Engineering International, 40(3), 1369-1381. https://doi.org/10.1002/qre.3472.
7. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553. https://doi.org/10.1016/j.dss.2009.05.016.
8. Beyaztas, U., & Shang, H. L. (2022). A robust partial least squares approach for function-on-function regression. Brazilian Journal of Probability and Statistics, 36, 199-219. https://doi.org/10.1214/22-BJPS544.
9. Cummins, D. J., & Andrews, C. W. (1995). Iteratively reweighted partial least squares: A performance analysis by Monte Carlo simulation. Journal of Chemometrics, 9, 489-507. https://doi.org/10.1002/cem.1049

10. Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. Analytica Chimica Acta, 185, 1-17. https://doi.org/10.1016/S0003-2670(00)80560-6.

11. Gil, J. A., & Romera, R. (1998). On robust partial least squares (PLS) methods. Journal of Chemometrics: A Journal of the Chemometrics Society, 12, 365-378. https://doi.org/10.1002/(SICI)1099-128X(199805)12:5<365::AID-CEM532>3.0.CO;2-9

12. González, J., Peña, D., & Romera, R. (2009). A robust partial least squares regression method with applications. Journal of Chemometrics: A Journal of the Chemometrics Society, 23(1), 78-90. https://doi.org/10.1002/cem.1150.

13. Hubert, M., & Van Branden, K. V. (2003). Robust methods for partial least squares regression. Journal of Chemometrics: A Journal of the Chemometrics Society, 17, 537-549. https://doi.org/10.1002/cem.783.

14. Hubert, M., Rousseeuw, P. J., & Van Aelst, S. (2008). High-breakdown robust multivariate methods. Statistical Science, 23(1), 92-119. https://doi.org/10.1214/07-STS227.

15. Pell, R. J. (2000). Multiple outlier detection for multivariate calibration using robust statistical techniques. Chemometrics and Intelligent Laboratory Systems, 52, 87-104. https://doi.org/10.1016/S0169-7439(99)00089-6.

16. Pensia, A., Jog, V., & Loh, P.-L. (2024). Robust regression with covariate filtering: Heavy tails and adversarial contamination. Journal of the American Statistical Association, 1-12. https://doi.org/10.1080/01621459.2024.1911182.

مجلة الغري للعلوم الاقتصادية والادارية
مجلد (21) عدد (2) 2025

17. Phillips, G. R., & Eyring, E. M. (1983). Comparison of conventional and robust regression in analysis of chemical data. Analytical Chemistry, 55, 1134-1138. https://doi.org/10.1021/ac00264a001

18. Pirouz, D. M. (2006). An overview of partial least squares. SSRN. https://doi.org/10.2139/ssrn.1631359.

19. Rosipal, R., & Krämer, N. (2005). Overview and recent advances in partial least squares. In International Statistical and Optimization Perspectives Workshop: Subspace, Latent Structure, and Feature Selection (pp. 34-51). Springer.

20. Serneels, S., Croux, C., Filzmoser, P., & Van Espen, P. J. (2005). Partial robust M-regression. Chemometrics and Intelligent Laboratory Systems, 79, 55-64. https://doi.org/10.1016/j.chemolab.2005.03.001.

21. Wold, H. (1975). Soft modeling: The basic design and some extensions. In Perspectives in Probability and Statistics: Papers in Honor of M. S. Bartlett (pp. 1-17). Academic Press.