Indian Journal of Science and Technology, Vol 12(8), DOI: 10.17485/ijst/2019/v12i8/141809, February 2019
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Comparative Analysis of Accuracy on Partial Least Squares and Principal Component Regression methods

M. Sujaritha, M. Kavitha and J. Janet

Department of Centre for Science and Environment, Sri Krishna College of Engineering and Technology, Kuniamuthur, Coimbatore - 641008, Tamil Nadu India; sujaritham@skcet.ac.in, kavitham@skcet.ac.in, jjanet@skcet.ac.in

Abstract

Objective: To associate or compare the goodness of fit of Principal Component Regression (PCR) and Partial Least Squares (PLS) models using metrics such as RMSEP, MSEP and R2. **Methods and Statistical Analysis:** Regression analysis is used in the study that involves investigation of correlation among an independent and dependent variables. Analysis is made simple when researchers understand and use the preeminent suitable method based on type of dependent variables, independent variables and dimensionality of data. Cross-validation method is used in both predictive models (PCR and PLS). **Dataset and Findings**: This study presents the comparative analysis on PCR and PLS by applying these methods on a public dataset named octane dataset, where the spectral data of gasolines with 401 attributes are provided. This study concludes that partial least squares regression model yields better prediction results than Principal Component Regression model since PLS accurately select the principal component. Also the number of principal components identified by the PLS is comparatively less. An analysis on preprocessing is also performed with same regression methods and dataset in this paper. **Improvements:** In this analysis the importance of removing Region of no interest is focused. If Region of No Interest is removed then number of principal component is also reduced which in turn increase the prediction accuracy. The study reveals the number of principal components is high if Region of No Interest is not used, which decreases the prediction accuracy.

Keywords: Pleast-Squares, Preprocessing, Principal Component Regression

1. Introduction

Research organizations / Government / Corporate Agencies are capturing data at each stage for different applications. These data capturing is increased exponentially in the last 5-6 years. The size of the data generated by these applications is huge ranging from few hundreds to lakhs. Visualizing such a huge data is difficult and further computation also takes time. At this particular situation data cleansing and data preprocessing comes into play.

Data preprocessing is the process of reducing the dimensions of data without losing much of information. But question comes in our mind, "Is this possible to diminish the importance of certain attributes without losing the integrity of data? Understanding data is more important for preparing data for any analysis. The important step involved in preparing data for analysis is preprocessing. The basic preprocessing techniques are data cleansing and normalization. The million dollar question 'Why preprocessing?' is answered in this section. Real world data or raw data are generally unreliable and containing outliers or errors. Raw data has to be converted into meaningful data through data cleansing for flawless analysis and prediction.

Identifying and removing Regions of No Interest (RoNI) is a part of data cleansing, many data have regions that should be removed before analysis. There may be regions of the data that simply don't have much informa-

*Author for correspondence

tion; they contribute a noise or outliers and not much more. After preprocessing, the next step is to discover number of principal components that represents the best original variables (independent) in the least square error sense.

The rest of the study is structured as follows. The applications and usage of PLS and PCR is explained in section 2 and section 3 explains these procedures in detail. The result of applying these methods in octane dataset is discussed in section 4. Conclusions and future work discussions are provided in the last section.

2. Literature Survey

Principal component analysis was first proposed by Hotelling in 1933.Principal component regression[1] is a straightforward method with high computational costs. Distinctive works have been delivered in the ongoing years about the algorithm PCR and PLS[2]. The Partial Least-Squares (PLS) regression method is gaining significance importance in numerous fields of science, systematic, physical, clinical science and mechanical process control. Made observational research in worldwide advertising with PLS[3] and proved the impact of PLS.

Predicted the subjective evaluation of a set of five wines using PLS. The reliant factors that will be anticipated for each wine are its amiability, and how well it runs with meat or pastry. The predictors are the price, sugar, alcohol, and acidity content of each wine[4].

Illustrated PLS correlation with an model in which 36 wines were portrayed by a matrix or table X which comprises 5 independent quantities (price, total acidity, alcohol, sugar, and tannin) and by a matrix or table Y which contains 9 sensory quantities (fruity, floral, vegetal, spicy, woody, sweet, astringent, acidic, hedonic).

With the help of PLS Abdi analyzed information common to both the tables and predicted one table or matrix from other[5]. Used PLS approach as a statistical methodology to fault isolation and detection of robot schemers[6]. Applied PLS for the normally distributed data and he suggested to perform a logarithmic transformation prior to PLS analysis if the data have outliers[7]. No used PLS to estimate the yields for the Stock Exchange of Thailand[8] used PLS on neuroimaging data for three gatherings of members with three members in each gathering, Matrix X stores neuroimaging or brain activity data (i.e. amplitudes across time for the vertex electrode) and matrix Y stores the behavioral data from a memory undertakingtask[9].

Investigated important components of fresh raw milk[2] using VIS-NIR spectrometers and PLS[10] method[11] conducted a study on nonlinear multivariate models to give a good fit[11] on regression analysis and produced better coefficient of determination (R^2 = 0.988 & 0.994) for the Research Octane Number (RON). Applied PLS and PCR on nine gasoline models acquired from the Pacific Coast Exchange Group[12] of the ASTM motor.

Ultimately more than 50 research papers have been published so far using PCR and PLS methodologies in different research sectors, but there is no research suggesting the appropriate number of principal components or range of principal components for predicting dependent variable using PCR and PLS regression models. Since the random selection of K-principal components yield erroneousness prediction and smart selection of K-component from the training set provides better prediction on test data, this paper analyzes the performance of PLS and PCR for various number of principal components. The octane dataset with 401 spectral reflectance values for sixty gasoline samples is used for the proposed analysis. The next section provides the fundamentals of normalization, PCR and PLS.

3. Methods and Techniques

Real world dataset contains features that highly vary in range, units and magnitudes. Normalization should be performed when the scale of a feature is irrelevant or misleading and should not normalize when the scale is meaningful. Normalization protects data integrity. Min-max scaling or min-max normalization[13], is the simplest method of rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Choosing the objective range relies upon the idea of the information.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Where x{\displaystyle x}xxx is an original value, {\displaystyle x›}x' is the normalized value

3.1 Principal Component Regression (PCR)

The fundamental point of PCR is to discover the relapse display with the best expectation execution and not the augmentation of the all-out clarified difference of free variable(X). Considering an information framework with

n x m esteems (X) that contains the indicator factors and furthermore a needy variable y with n perceptions, the univariate relapse show is:

$$y = X * b + e$$

Algorithm: PCR

Input: Data Situation

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}, x = \begin{bmatrix} x_{11} & x_{12} & \cdots & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & \cdots & x_{ip} \end{bmatrix}$$

Output: Predict Y value from X

Sequence of operation:

Step 1: Perform Multiple linear regression with X principal components t1, . . . , tx instead of all of the x's

Step 2: How many components: Determine by Cross-validation.

- Leave-out-one of the interpretations
- Fit a prototypical on the remaining(reduced) data
- Predict the left out perception by the model
- Do this in turn for ALL observations and compute the overall performance of the model by Root Mean Squared Error of Prediction (RMSEP)

Step 3: Validate the model (select the one with less RMSEP value and more R2 value)

Step 4. Interpret, conclude, predict future values Y.

3.2 Partial Least Squares (PLS)

Like PCR, PLS selects components that explain the most variance in the model, but unlike PCR, PLS incorporates the response variable. i.e., it includes the feedback. PLS is an amazing multivariate factual tool that estimates the predictive or connecting relationship between variables.

Algorithm: PLS

Input: Data Situation

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}, x = \begin{bmatrix} x_{11} & x_{12} & \cdots & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & \cdots & x_{ip} \end{bmatrix}$$

Output: Predict Y value from X

Sequence of operation:

Step 1: finds a set of components

Step 2: fitting a set of components to X (as in PCA)

Step 3: likewise fitting a set of components to Y

Step 4: The X and Y scores are chosen so that the association between consecutive pairs of scores is as robust as possible(maximum covariance of X and Y

Step 5 Validate the model (select the one with less RMSEP value)

Step 6. Interpret, conclude, and predict future values Y.

The following section provides the result of analysis on PCR and PLS and the accuracy of the models is measured by minimum Root Mean Squared Error of Prediction (RMSEP), Mean Squared Error of Prediction (MSEP) and maximum R^2. The smaller the estimated error becomes the healthier the prediction.

4. Analysis Results

Sixty data frame with Octane number and 401 NIR spectra of gasoline models has been analyzed in this learning. The NIR spectra were measured using diffuse reflectance as log (1/R) from 900 nm to 1700 nm in 2 nm intervals, giving 401wave lengths[14-16].

The likelihood of predicting the octane number by the near infrared spectra was investigated. Min-Max Normalization is applied on the entire data set. All the values are scaled between 0 -1, to process the data. Analytics

environment used is R i386 3.4.0. R is a primary tool for machine learning, statistics, and data analysis.

4.1 Analysis of PCR and PLS on Raw Data

The data frame with 60 observations is separated into training information and test information. Observation 1 to 50 is selected for Training our model using PCR and PLS, remaining 51 to 60 is kept for testing. Figure 1 a shows the picture generated by R i386 3.4.0for the spectral signature of actual or raw data. A PCR-model and PLS-model is generated by applying PCR and PLS on training dataset. With the help of PCR-model and PLS-model, the prediction of octane number in test dataset is carried out and RMSEP, MSEP and R^2values of training record set and test record set are recorded or tabulated for further analysis. Figure 2 shows the RMSEP, MSEP and R^2 value generated by R i386 3.4.0for components ranging from 1 to 10, which helps in selecting the optimum number of components for prediction for both the model PCR and PLS. Table 1 list the RMSEP, MSEP and R^2 value for components ranging from 1 to 10 for PCR and PLS for Raw data.

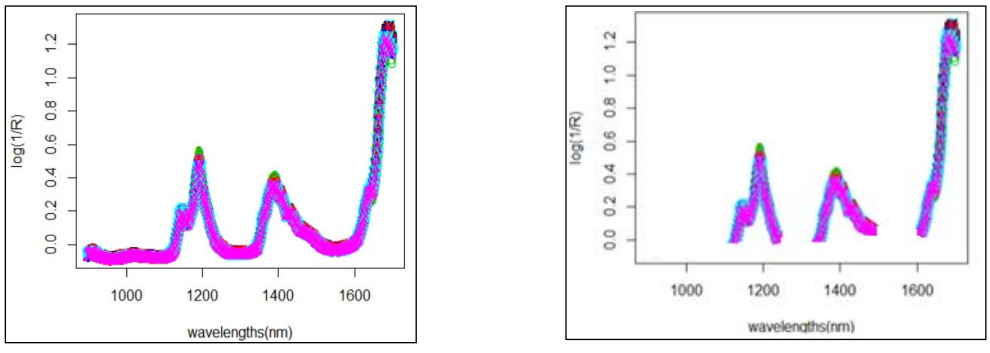
From the plot and table the conclusion made is that for PCR the number of component for accurate predict can be 9, because at component 9 RMSEP (0.23) is minimum and maximum R^2 (0.97). But in case of PLS model at component 6 itself RMSEP (0.23) is minimum and maximum R^2(97). So for further predict or iterative process the best suited number of component for PCR is K=9 and PLS is K=6. In PCR the Principal component can be selected from 7 component to 9(i.e 7>=K<=9). In PLS the Principal component can be selected from 4 component to 9 (i.e. 4>=K<=9).

Figure 3 shows the prediction *vs* measured value of octane number using PCR and PLS generated using R i386 3.4.0. Accuracy of prediction is perfect in PCR when 9 components are taken into account, where as in PLS 6 components are taken into account for prediction.

4.2 Analysis of PCR and PLS on Preprocessed Data

The reflectance variation is found between 1150 to 1250, 1350 to 1450 and 1600 to 1700. The wavelength from 900 to 1149, 1252 to 1348, and 1452 to 1598 are considered as Region of no interest because they contribute redundant information and removed from spectral data. 243 spectral wave lengths is considered as Region of No Interest and are not considered for training or testing purpose. The result reveals that 401 columns are reduced to 158columns. A PCR-model and PLS-model is generated by applying PCR and PLS on training dataset which is preprocess. With the help of PCR-model and PLS-model, the prediction of octane number in preprocessed test dataset is carried out and RMSEP, MSEP and R^2values of training record set and test record set are recorded or tabulated for further analysis.

Figure 1 (b) shows the spectral region after removing region of no interestgenerated using R i386 3.4.0. Figure 4 shows the RMSE,MSEP and R^2value generated using R i386 3.4.0 for components ranging from 1 to 10, which helps in selecting the optimum number of components for prediction for both the model PCR and PLS with Removing Region of No Interest. Table 2 list the RMSEP, MSEP and R^2 value for components ranging from 1 to 10 for PCR and PLS for normalized data and with region of Interest. From the plot and table conclusion made is that for PCR the number of component for accurate predict



(a) (b)

Figure 1. a) Spectral signature of gasoline data, b) Spectral region after removing region of No Interest.

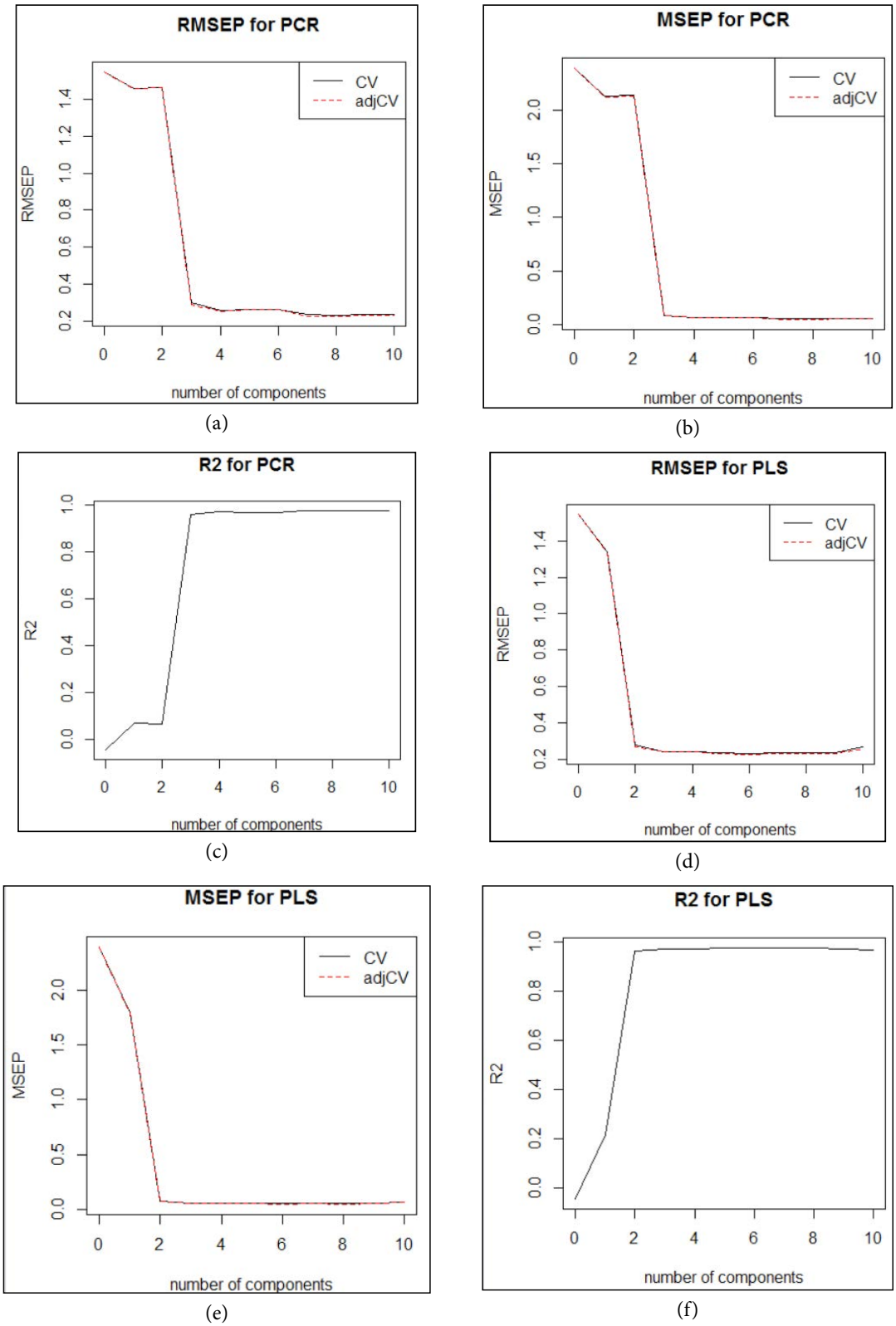
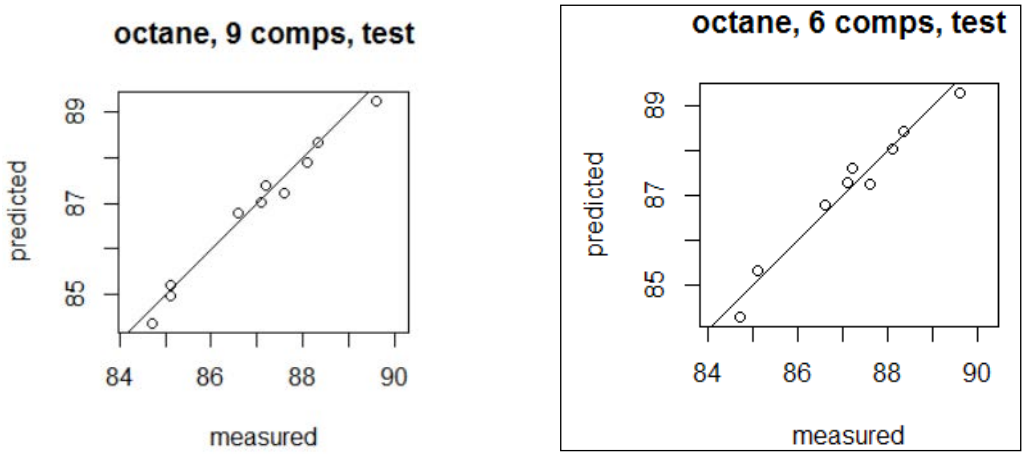


Figure 2. RMSEP, MSEP and R^2 plot for PCR and PLS.

Table 1. RMSEP, MSEP and R^2 values for PCR and PLS for Raw data

Algo-rithm	Number of Compo-nents	Training Data			Test Data		
		RMSEP	MSEP	R^2	RMSEP	MSEP	R^2
PCR	1	1.519	2.306	0.0059	1.3226	1.749	0.234
	2	1.508	2.274	0.0079	1.256	1.579	0.308
	3	0.3131	0.098	0.9572	0.4634	0.214	0.905
	4	0.260	0.067	0.9705	0.224	0.050	0.978
	5	0.270	0.0729	0.968	0.2283	0.021	0.977
	6	0.2755	0.0758	0.966	0.2600	0.067	0.9704
	7	0.2595	0.0673	0.9706	0.2795	0.078	0.965
	8	0.2408	0.0579	0.974	0.243	0.059	0.974
	9	0.2366	0.0559	0.975	0.2290	0.052	0.977
	10	0.2380	0.0566	0.975	0.2881	0.082	0.963
PLS	1	1.335	1.781	0.2231	1.1696	1.3679	0.4011
	2	0.3266	0.1067	0.9534	0.2445	0.0597	0.9738
	3	0.2707	0.0733	0.9680	0.2341	0.0548	0.976
	4	0.2591	0.0671	0.9707	0.3287	0.1080	0.952
	5	0.2342	0.0548	0.9760	0.2780	0.0773	0.9661
	6	0.2327	0.0541	0.9763	0.2703	0.0730	0.968
	7	0.2395	0.0573	0.9749	0.3301	0.1089	0.952
	8	0.2436	0.0593	0.9741	0.3571	0.1275	0.944
	9	0.2509	0.0629	0.9725	0.4090	0.1672	0.926
	10	0.2698	0.0727	0.9682	0.6116	0.3741	0.836



(a) (b)

Figure 3. a) Prediction and measured plot for PCR, b) Prediction and measured plot for PLS.

can be 8, because at component 8 RMSEP (0.03) is minimum and maximum R^2 (0.97). In case of PLS model at component 4 RMSEP (0.03) is minimum and maximum R^2 (97). So for further predict or iterative process the best suited number of component for PCR is 8 and PLS is 4.

From the analysis the conclusion made is that when a preprocessed data improve the prediction with less number of components. 1. Component is reduced in PCR, and 2. Component is reduced in PLS. In PCR the Principal component can be selected from 7 component

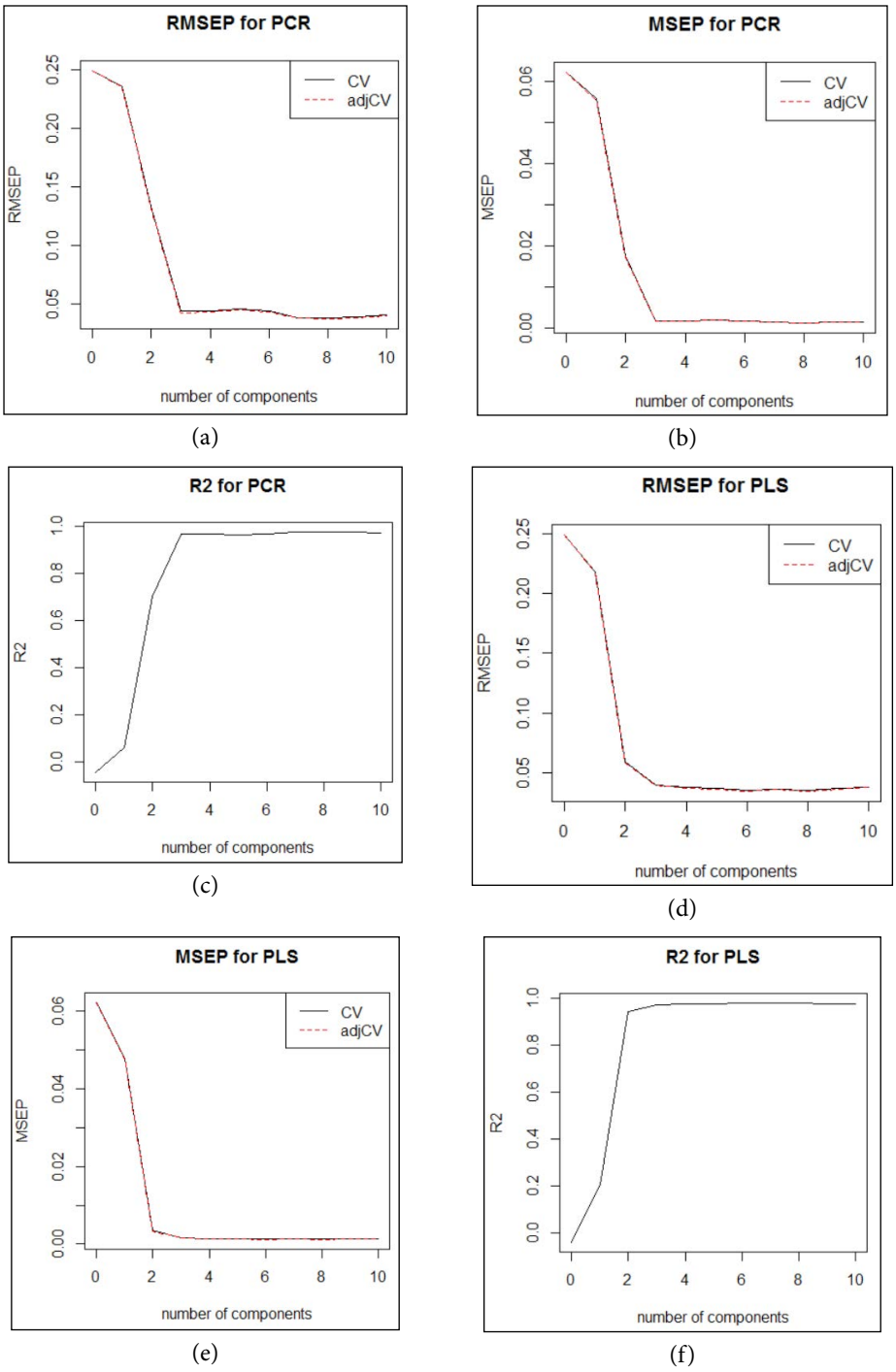


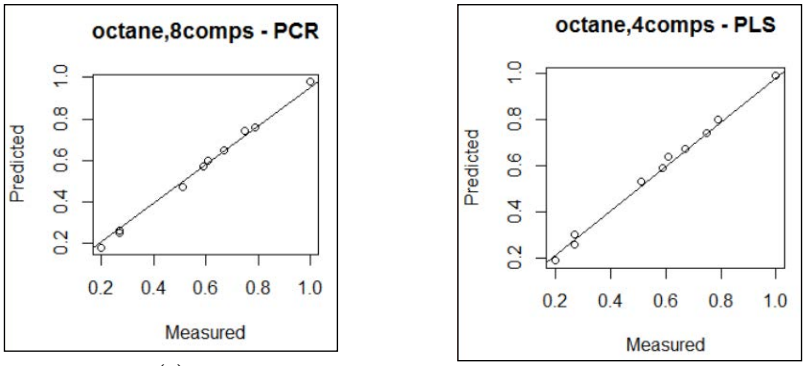
Figure 4. RMSEP, MSEP and R^2 plot for PCR and PLS with preprocessing.

to 8 (i.e 7>=K<=8). In PLS the Principal component can be selected from 4 component to 7 (i.e 4>=K<=7).

Figure 5 shows the prediction *vs* measured value of octane number using PCR and PLS generated using R i386 3.4.0. Accuracy of prediction in PCR is when 8 components are taken into account, where as in PLS 4 components are taken into account for prediction. Over fitting and under fitting is not an issue in our gasoline dataset, but the size of the data is high over fitting the model may happen.

Table 2. RMSEP, MSEP and R^2 values for PCR and PLS for preprocessed data

Algo-rithm	Number of Compo-nents	Training Data			Test Data		
		RMSEP	MSEP	R^2	RMSEP	MSEP	R^2
Min-Max + RONI+ PCR	1	0.238	0.056	0.047	0.222	0.049	0.171
	2	0.136	0.018	0.689	0.123	0.015	0.744
	3	0.044	0.001	0.966	0.038	0.0014	0.975
	4	0.043	0.001	0.968	0.038	0.0014	0.975
	5	0.044	0.001	0.967	0.037	0.0014	0.976
	6	0.043	0.001	0.968	0.035	0.0012	0.978
	7	0.039	0.001	0.974	0.030	0.0009	0.984
	8	0.038	0.001	0.974	0.029	0.0008	0.985
	9	0.039	0.001	0.973	0.0275	0.0007	0.987
	10	0.040	0.001	0.972	0.0270	0.0007	0.987
Min-Max + RONI +PLS	1	0.221	0.049	0.177	0.205	0.042	0.293
	2	0.060	0.003	0.938	0.054	0.002	0.950
	3	0.0412	0.001	0.9715	0.035	0.001	0.979
	4	0.0376	0.0014	0.976	0.028	0.0007	0.98
	5	0.038	0.0014	0.975	0.026	0.0006	0.988
	6	0.039	0.0015	0.973	0.025	0.0006	0.989
	7	0.040	0.0016	0.973	0.0244	0.0005	0.990
	8	0.041	0.0017	0.971	0.023	0.0005	0.991
	9	0.042	0.0018	0.969	0.021	0.0004	0.992
	10	0.042	0.0018	0.969	0.020	0.0004	0.993



(a) (b)

Figure 5. a) Prediction and measured plot for PCR, b) Prediction and measured plot for PLS.

5. Conclusion

Proper application of data cleansing and data pre-processing techniques can reduce analysis time and increase the prediction accuracy. The primary focus of this study was to investigate the feasibility of predicting the K-principal component or suggesting the range of K in PCR and PLS algorithm on raw data and preprocessed data. PLS have: 1. improved predictive accuracy, and 2. a lower risk of correlation. Our study proves PLS is a good alternative to the more classical multiple linear regression and principal component regression methods because it is more strong and healthy. Sound implies that the model parameters don't change especially when new adjustment

tests are taken from the complete populace. Finally comparison of PCR and PLS with raw data and preprocessed data is recorded or tabulated. K-principal component and range of K values for PCR and PLS regression model is suggested. In future the same sequence of analysis can be done for big data. Larger data set show a large variation in selecting number of components for PCR and PLS with raw data and preprocessed data. Preparing the model to generalize from the training record to any record or data from the problem domain, which allows us to make predictions in the future on data the model has never seen.

6. References

1. Chang CW, Laird DA, Mausbach MJ, Hurburgh CR. Near-infrared reflectance spectroscopy - principal components regression analyses of soil properties, Agricultural and Biosystems Engineering. 2001; 480–90.
2. Comparison of principal component regression and partial least square methods in prediction of raw milk composition by VIS-NIR Spectrometry' Application to development of online sensors for Fat, protein and lactose contents. Date accessed: 11.09.2009. http://www.imeko2009.it.pt/Papers/FP_229.pdf.
3. Henseler J, Ringle CM, Sinkovics RR. The use of partial least squares path modeling in international marketing, Advances in International Marketing, 2009; 277–319.
4. Abdi H. Partial least squares regression and projection on latent structure regression (PLS Regression), Wires Computational Statistics. 2010; 2(1):97–106. https://doi.org/10.1002/wics.51.
5. Abdi H, Williams LJ. Partial least squares methods: Partial least squares correlation and partial least square regression, Methods in Molecular Biology. 2013; 930:549–79. https://doi.org/10.1007/978-1-62703-059-5_23. PMid: 23086857.
6. Muradore R, FioriniP. A PLS-based statistical approach for fault detection and isolation of robotic manipulators, IEEE Transactions on Industrial Electronics. 2012; 59(8):3167–75. https://doi.org/10.1109/TIE.2011.2167110.
7. Sawatsky ML, Clyde M, Meek F. Partial least squares regression in the Social Sciences, The Quantitative Methods for Psychology. 2015; 11(2):52–62. https://doi.org/10.20982/tqmp.11.2.p052.
8. Sopipan N. Forecasting the financial returns for using multiple regression based on principal component analysis, Journal of Mathematics and Statistics. 2013; 9(1):65–71. https://doi.org/10.3844/jmssp.2013.65.71, https://doi.org/10.3844/jmssp.2013.29.37.
9. Roon PV, Zakizadeh J, Chartier S. Partial least squares tutorial for analyzing neuroimaging data, The Quantitative Methods for Psychology. 2014; 10(2):200–15. https://doi.org/10.20982/tqmp.10.2.p200.
10. Mevik BH, Wehrens R. The PLS package: Principal component and partial least squares regression in R, Journal of Statistical Software. 2007; 8(2):1–23.
11. Oduola MK, Iyaomolere AI. Development of model equations for predicting gasoline blending properties, American Journal of Chemical Engineering. 2015; 3(2-1):9–17.
12. Kelly JJ, Barlow CH, Jinguji TM, Callis JB. Prediction of gasoline octane numbers from near-infrared spectral features in the range 660-1215 nm, Analytical Chemistry. 1989; 61(4):313–20. https://doi.org/10.1021/ac00179a007.
13. Bing S, Ding X, Liu C, Wu Y. Heteroscedastic max–min distance analysis for dimensionality reduction, IEEE Transactions on Image Processing. 2018; 27(8):4052–65. https://doi.org/10.1109/TIP.2018.2836312. PMid: 29994529.
14. Kalivas JH. Two data sets of near infrared spectra, Chemometrics and Intelligent Laboratory Systems. 1997; 37(2):255–59. https://doi.org/10.1016/S0169-7439(97)00038-5.
15. Robust Principal Component Analysis with Complex Noise. Date accessed: 2014. http://proceedings.mlr.press/v32/zhao14.pdf.
16. Discovering Partial Least Squares with JMP. Date accessed: 2013. https://support.sas.com/content/dam/SAS/support/en/books/discovering-partial-least-squares-with-jmp/65346_excerpt.pdf.