

The Trade-off Between the PLSR and PCR Methods for Modeling Data with Collinear Structure

W.B. Yahya¹, K.O. Olorede¹, M.K. Garba¹, A.W. Banjoko¹ and K.A. Dauda²

¹Department of Statistics, University of Ilorin, P.M.B. 1515, Ilorin, Nigeria.

²Department of Statistics and Mathematical Sciences, Kwara State University, P.M.B. 1530, Malete, Nigeria.

Abstract

This paper investigates the partial least squares regression (PLSR) and principal component regression (PCR) methods as versatile alternative regression techniques when the use of the ordinary least squares method breaks down. Emphasis is more on the situation where the predictor variables are evidently correlated. Data sets with Gaussian non-orthogonal predictor variables were simulated at different sample sizes ranging from 20 to 1000 to examine the performance of the two regression types under varying situations. The data were randomly partitioned into training and test sets with both PLSR and PCR models constructed on the training sets while their performances were evaluated on the test sets using the mean square error of predictions and other indices. At each fit of the models, the leave-one-out cross-validation technique was employed to enhance the efficiency and stability of the fitted models. Results from the simulation studies revealed the goodness of the two regression methods but at varying degrees of accuracy. More importantly, it is evident from the results that though, both the PLSR and PCR techniques yielded good regression models, the PLSR technique is consistently more efficient on the test data in terms of good predictions than the PCR method irrespective of sample sizes. Also in terms of model parsimony, the PLSR technique yielded efficient regression models with relatively fewer latent components than the PCR method. Data sets on the performance of M.Sc. graduates from the Department of Statistics, University of Ilorin, Nigeria during the 2012 academic session were used to validate the results from the Monte Carlo studies.

Keywords: PLS Regression, Principal Component Regression, Root Mean Square Error of Prediction, Leave-One-Out Cross-Validation (LOOCV).

2010 MSC: 62J05, 62J07, 65C20, 68U20.

1.0 Introduction

Within the last two decades, a number of works have evolved on the development and applications of various methods for modeling data with high correlation structure [1-3]. Some of these methods include the ridge regression [4,5], slice inverse regression [6], principal component regression [2,7-9] and the partial least squares regression [3,10-12] methods among others. Of all these methods, the principal component regression (PCR) and partial least squares regression (PLSR) methods, apart from being capable at modeling data with correlated predictors, are suitable tools used to break the 'curse of dimensionality' [13-15], the case of larger p predictors and fewer sample specimens n , with $n \ll p$, in multiple linear regression and classification problems [16].

Multicollinearity is a situation where two or more predictor variables in multiple (linear) regression models are correlated [5,17,18]. If this situation arises, the estimated regression models through the classical ordinary least squares (OLS) techniques becomes very unstable and less efficient [19]. In classical real life situations however, interdependency among the set of predictor variables in regression models is inevitable [5,7,19].

When there is evidence of strong relationship among the predictor variables in multiple linear regression models, it is desirable to adopt any of the appropriate alternative estimation techniques to efficiently model such collinear data. This is to ensure the development of an efficient regression model that would be useful for meaningful inferences.

Corresponding author: W.B. Yahya, E-mail: wbyahya@unilorin.edu.ng, Tel.: +234703-057-8626

The PCR was first suggested by Kendal [20] in his book followed by Hotelling[21] in a different work in the same year as a method for fitting multiple linear regression models by replacing the original predictor variables in the models, most of which are collinear, by their orthogonal latent components. This technique provides regression models that are more efficient and stable. Several modifications and applications of the PCR techniques have evolved [22,23] after its original proposition by Kendal [20] and Hotelling[21].

The PLSR is a method developed by Herman Wold in the 1960s to model data with large numbers of predictor variables most of which were highly correlated [1,10,7,24]. This regression method can be employed on data with many numbers of explanatory variables, even when such is more than the number of observations. The underlying assumption of PLSR method is that the observed data is generated by a system or process that is driven by a smaller number of variables [14, 25,26].

Without loss of generality, many of the real life problems nowadays that are being modeled are fraught with the problem of multicollinearity. Therefore, if the collinear structure in many real life data problems is inevitable, any research effort that would enhance the efficiency of techniques to model such data shall always be a welcome development.

In this paper therefore, the procedures of two of the alternative regression methods, the PLSR and PCR models, primarily developed to model data with inherent collinear structure in the predictor variables were investigated. These two regression methods have been reported to be biased but are efficient regression techniques to model data with evidence of multicollinearity [7,27]. It is then necessary to investigate their relative strengths for better performance. As a result, this work investigates the efficiency of the PLSR and PCR models using both the simulated and real life data sets. Findings from this work would serve as useful guide to researchers and students alike in their future applications of both the PCR and PLSR models to compelling life situations.

2.0 Methodologies

In this section, we provide a brief overview of the PCR and PLSR methodologies as they are employed to model graduate students' scores.

Consider a multiple linear regression model of the form

$$Y = X'\beta + \varepsilon \quad (1)$$

where y is the $n \times 1$ vector of responses (e.g. final weighted scores of graduate students), X $p \times 1$ vector of (mostly) correlated predictor variables (e.g. graduate courses offered in the University by the students) each of length n , β is the $p \times 1$ vector of the regression coefficients while ε is the random noise of the model that is assumed to have independent Gaussian density with zero mean and a constant variance σ^2 .

Since some of the predictors in equation (1) are correlated, the ordinary least squares method would not yield efficient and stable estimates of the regression parameters β [5,25,]. With multicollinearity in the model, the design matrix $X'X$ would be ill-conditioned [28] while the confidence intervals of the regression parameters may be unreasonably wide [15].

2.1 Brief Overview of the PCR and PLSR methods

The PCR technique is one of the multivariate regression methods often employed to model data with correlated predictors. The PCR neutralizes the ill-condition problem in equation (1) by constructing a number of new orthogonal latent variables, often called the Principal Components (PCs) from the original highly correlated predictor variables in matrix X . The newly constructed PCs capture the maximum variations in X , and as a result, optimize the predictive ability of the resulting PCR model from original regression model (1). The PCR is basically a linear multiple regression type in which the response variable Y is regressed on the principal (latent) components [29].

A common practice in fitting the PCR model is to center and scale the column vectors of predictors in matrix X to unit lengths or to have zero means and unit standard deviations [15]. Thus, each predictor variable X_j in matrix X is replaced by a new centered and scaled variable $\tilde{X}_j$ obtained either through the transformation

$$\tilde{X}_j = \frac{x_{ij} - \bar{x}_{ij}}{\sqrt{\sum (x_{ij} - \bar{x}_{ij})^2}} \quad (2)$$

or

$$\tilde{X}_j = \frac{x_{ij} - \bar{x}_{ij}}{\sqrt{\frac{\sum (x_{ij} - \bar{x}_{ij})^2}{n-1}}} \quad (3)$$

The eigenvalues λ_j , $j = 1, 2, \dots, p$ of the $p \times p$ correlation matrix $\tilde{X}'\tilde{X}$ are computed from the centered and scaled original predictor matrix X by solving the determinantal equation [30]

$$|\tilde{X}'\tilde{X} - \lambda_j I| = 0 \quad (4)$$

while the eigenvectors T_j of $\tilde{X}'\tilde{X}$ that correspond to the eigenvalues λ_j are determined through the equation

$$(\tilde{X}'\tilde{X} - \lambda_j I)T_j = 0 \quad (5)$$

Thus, vectors $T = (T_1, T_2, \dots, T_p)$, called the eigenvectors, are the normalized solutions of the set of equations in (5) that satisfy the conditions that $T_j'T_j = 1$ (i.e. T_j has a unit length) and that $T_j'T_{j'} = 0$ for $j \neq j' = 1, 2, \dots, p$ (i.e. T_j and $T_{j'}$ are purely orthogonal).

The $n \times p$ matrix of principal components C is therefore determined by the matrix product

$$C = \bar{X}T \quad (6)$$

Hence, each principal component (PC) C_j in the column of matrix C in (6) is expressed in terms of the entire p predictor variables in $\bar{X}$ and the eigenvector T_j in the form

$$C_j = T_{j1}\bar{X}_{i1} + T_{j2}\bar{X}_{i2} + \dots + T_{jp}\bar{X}_{ip} \quad (7)$$

Usually, the first k principal components $C_k = (C_1, \dots, C_k)$, $k < p$, in (6) might be sufficient to capture almost all the information in the original predictor variables in $\bar{X}$. Thus, the OLS regression of the response variable Y on the set of k PCs results into the model

$$Y = \alpha_0 1 + C_k \alpha_k + \epsilon \quad (8)$$

where the OLS estimator of regression coefficients α_k in (8) is $\alpha_k = (C_k' C_k)^{-1} C_k' Y$. The principal components C_j 's are orthogonal to each other (i.e. $C_j' C_{j'} = 0$, for $j \neq j'$, $j, j' \in \{1, \dots, k\}$) and they are often referred to as artificial (latent) variables [20] in the regression model (8).

The development of the PLSR model is very similar to that of the PCR model. The difference lies in the way the latent components (variables) are constructed in the regression models.

To regress the response variable Y on the set of explanatory variables $X_1, \dots, X_p$, the partial least squares (PLS) seeks to find new variables that will play the same roles as the original X variables in the regression model. Each of these variables is a linear combination of the original predictors $X_1, \dots, X_p$ as it was the case in PCR modeling. However, while the PCR uses only the variations in vector of predictor variables X to construct the PCR latent components (C_k) in model (8), the PLSR uses both the variations in X and Y to construct its latent (PLSR) components that are finally used as predictor variables in the PLSR model [14, 26].

The objective of the PLSR technique is to maximize the information in its constructed components to be used as the new predictor variables in the model by building them using the information in both the response variables Y and that of the original predictor variables $X_1, \dots, X_p$. This technique is expected to improve the quality of the PLSR components in the model and enhance their prediction strength in the model. The intention of PLS is to form components that capture most of the information in both Y and X variables for better performance [31-33].

3.0 Simulation Studies

This section presents the scheme employed for the Monte-Carlo experiments performed in this work as well as the results of the applications of the PCR and PLSR models on the simulated data sets.

3.1 Simulation Scheme

Data sets were simulated for the multiple linear regression model $Y = X'\beta + \epsilon$ in (1) using the following schemes.

Ten non-orthogonal predictor variables in vector $X' = [X_1, X_2, \dots, X_{10}]$ were simulated from multivariate Gaussian distribution with specified mean vector $\mu' = [0, 5, 30, 25, 0, 2, 15, 23, 11, 1]$ and variance-covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{110} \\ \sigma_{12} & & & \sigma_{210} \\ \vdots & & \ddots & \vdots \\ \vdots & & & \vdots \\ \sigma_{110} & \sigma_{210} & \dots & \sigma_{10}^2 \end{bmatrix} \quad (9)$$

at varying sample sizes $n = 20, 40, 60, 80, 100, 150, 200, 300, 500$, and 1000 .

In order to inject collinearity among the set of variables in the predictor vector X in regression model (1), the covariance between predictor variables X_j and $X_{j'}$, $cov(X_j, X_{j'}) = \sigma_{jj'}$, (the off-diagonal element of Σ) was determined by the relation

$$cov(X_j, X_{j'}) = \rho_{jj'} \sigma_j \sigma_{j'} \quad (10)$$

where σ_j is the standard deviation of predictor variable X_j and $\rho_{jj'}$ is the correlation coefficient between X_j and $X_{j'}$, $j, j' = 1, \dots, 10$. The values of $\rho_{jj'}$ were set such that some variable pairs have correlation values within the interval $0 < \rho_{jj'} < 0.5$ while others have their correlation values set within the interval $0.5 \leq \rho_{jj'} < 1$ to ensure that the variance-covariance matrix Σ is positive-definite. However for symmetry, it is expected that similar results would be obtained with the presence of negative multi-collinearity structure in the data for which the values of $\rho_{jj'}$ fall in the interval $-1 < \rho_{jj'} < 0$.

The elements of the vector of regression parameter β were set at $\beta' = [2, 4, 1.5, 2.5, 1, 3, 2, 5, 3, 4.2]$ while the error term ϵ was simulated from a unit Gaussian density. Finally, the response variable Y was simulated directly using the relationship in (1).

For the development of both the PCR and PLSR models, the internal cross-validation approach of the *Leave-One-Out Cross-Validation* (LOOCV) technique was adopted to ensure results' stability. All the simulations and analysis were performed in the environment of R statistical package (<http://www.R-project.org/>) [34].

3.2 Models' Assessment

The relative efficiency of the PCR and PLSR models were accessed by the Root-Mean Square Error (RMSE) and the Mean Square Error of Cross-Validation (MSECV) or the Root-Mean Square Error of Cross-Validation (RMSECV). The RMSE defined by

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (11)$$

measures the deviation of the fitted model from the data where y_i and $\hat{y}_i$ are the observed and predicted responses respectively. The RMSE is a measure of how well the model fits the data. This is in contrast to RMSECV which is a measure of a model's ability to predict new samples. The RMSECV is defined as in equation (11), except that the y_i and $\hat{y}_i$ are the observed and predicted responses of the test samples respectively. Thus, the RMSECV is computed by

$$RMSECV = \sqrt{\frac{PRESS}{n}} \quad (12)$$

where PRESS is the prediction error sum of squares which is calculated via a LOOCV technique where each sample is left out of the model formation and predicted once by the fitted model.

Another metric for model assessment adopted is the Mean-Square Error of Prediction (MSEP) which is only computed when the fitted model is applied to predict new (external) data with known response values. The MSEP is also calculated as in equation (11) over the test data by taken the squares of the estimated values.

Finally, the efficiency of the PLSR and PCR were equally assessed by their coefficient of multiple determinations (R^2), percentage of variance in the response variable Y explained by their latent components and the optimal number of latent components employed to attain better predictions.

4.0 Simulation Results

The regression types were implemented on simulated data for various samples sizes as stated in sub-section 3.1. The maximum number of latent components that can be employed by both the PCR and PLSR techniques is limited to $\max(n, p)$ where n and p are the sample size and number of predictor variables in the data respectively. However, since only ten predictors were simulated at all the sample sizes with the minimum being 20, it follows that the maximum number of components each model can employ cannot exceed 10.

The summary of results of the PCR and PLSR models on the simulated data at each latent component employed up to ten are reported in Tables 1 and 2. Regression results for data sets with small sample sizes $n = 20, 40, 60, 80$ and 100 are presented in Table 1 while results for data sets with large sample sizes $n = 150, 200, 300, 500$ and 1000 are presented in Table 2.

Results presented in Tables 1 and 2 include the amount of variations in the response variable Y captured by the number of components employed (in %) by the two regression models, the cumulative percent of such variation explained at each added latent component in the model (i.e. the R^2), the RMSECV and the RMSE for each fitted model.

Since ten different sample sizes were considered with ten latent components for each regression type in our Monte-Carlo experiment, it therefore shows that 100 data sets were simulated for each of the PCR and PLR models. By implication, a total of 200 PCR and PLRS models were constructed for the Monte-Carlo study.

For small sample size situation, it can be observed from the results in Table 1 that the PLSR model is consistently more efficient than the PCR model as it provides the least RMSECV at all the ten latent components across all the five sample sizes considered. This same good feat of performance was repeated by the PLSR model as compared to the PCR under the large sample situation as evident in Table 2.

Table 1: Results of PCR and PLSR models for simulated data for small sample sizes ($n = 20, \dots, 100$).

Sample Size (n)	No. of Comp. (C)	PCR				PLSR			
		% of Variance in Y	Cum % Var. (R^2)	RMSECV	RMSE	% of Variance in Y	Cum % Var. (R^2)	RMSECV	RMSE
20	1	97.74	97.74	9.186	2.9836	98.49	98.49	7.723	0.5717
	2	0.90	98.64	7.365	0.0036	1.21	99.70	4.356	10.8312
	3	0.87	99.51	6.875	1.2676	0.41	99.84	3.296	14.4322
	4	0.17	99.68	4.825	17.7040	0.08	99.92	3.308	6.6521
	5	0.10	99.78	3.937	31.4076	0.04	99.96	3.037	0.0008
	6	0.02	99.80	4.686	12.0595	0.01	99.97	2.248	0.0506
	7	0.17	99.97	3.204	6.4592	0.01	*99.98	1.773	0.0006
	8	0.00	*99.97	2.233	0.3350	0.00	99.98	2.154	1.6911
	9	0.01	99.98	2.521	1.1028	0.00	99.98	2.565	1.1014
	10	0.00	99.98	3.205	1.4740	0.00	99.98	3.205	1.4741

Cont. of Table 1

40	1	94.83	94.83	10.420	136.7082	97.98	97.98	6.964	21.5897
	2	4.14	98.97	4.905	16.2319	1.56	99.54	3.355	18.2756
	3	0.22	99.19	4.519	5.6045	0.27	99.81	2.320	7.1317
	4	0.05	99.24	4.325	19.9046	0.07	99.88	1.920	5.9618
	5	0.03	99.27	4.528	8.5672	0.07	99.95	1.360	0.3419
	6	0.47	99.74	2.605	14.3541	0.01	99.96	1.105	0.4012
	7	0.00	99.74	2.667	13.5863	0.01	99.97	1.080	0.0052
	8	0.01	99.75	2.774	14.5764	0.00	99.97	1.009	0.0487
	9	0.22	99.97	1.009	0.0427	0.00	99.97	1.007	0.9005
	10	0.00	99.97	1.046	0.0676	0.00	99.97	1.046	0.0676
60	1	97.70	97.70	6.977	31.7501	98.74	98.74	5.197	20.7380
	2	0.08	97.78	6.977	30.9317	0.85	99.59	3.123	6.5228
	3	1.24	99.02	4.772	17.9220	0.23	99.82	2.134	0.9970
	4	0.08	99.10	4.538	9.3448	0.10	99.92	1.509	0.0000
	5	0.26	99.36	3.861	15.3995	0.01	99.93	1.306	0.1405
	6	0.01	99.37	3.910	11.0425	0.01	99.94	1.263	0.3384
	7	0.05	99.42	3.839	16.6983	0.00	99.94	1.262	0.3056
	8	0.49	99.91	1.597	0.5085	0.00	99.94	1.260	0.3610
	9	0.03	99.94	1.263	0.2839	0.00	99.94	1.265	0.2325
	10	0.00	99.94	1.246	0.5988	0.00	99.94	1.246	0.5988
80	1	97.45	97.45	7.116	57.0685	98.52	98.52	5.510	31.6770
	2	0.91	98.36	5.835	31.5636	1.01	99.53	3.263	1.4991
	3	0.10	98.46	5.810	25.4211	0.26	99.79	2.293	0.0024
	4	0.36	98.82	5.562	10.6156	0.14	99.93	1.334	0.0658
	5	0.53	99.35	3.867	0.4689	0.02	99.95	1.108	0.7591
	6	0.04	99.39	3.829	1.9808	0.01	99.96	0.990	2.6163
	7	0.23	99.62	3.067	7.1648	0.00	99.96	0.972	2.7122
	8	0.26	99.88	1.741	0.0385	0.00	99.96	0.966	2.5949
	9	0.08	99.96	0.961	2.6440	0.00	99.96	0.960	2.6689
	10	0.00	99.96	0.963	2.9164	0.00	99.96	0.963	2.9164
100	1	97.95	97.95	6.723	0.0031	98.60	98.60	5.626	0.0331
	2	0.15	98.10	6.574	0.4658	1.11	99.71	2.686	1.3836
	3	0.00	98.10	6.903	0.4475	0.11	99.82	2.136	2.2299
	4	1.41	99.51	3.508	6.6215	0.10	99.92	1.414	0.1368
	5	0.05	99.56	3.328	3.8605	0.03	99.95	1.114	1.0731
	6	0.09	99.65	2.966	0.7900	0.01	99.96	1.072	1.0308
	7	0.01	99.66	2.924	0.8058	0.00	99.96	1.050	1.0851
	8	0.12	99.78	2.416	0.4639	0.00	99.96	1.056	1.0480
	9	0.18	99.96	1.072	1.1090	0.00	99.96	1.040	0.8511
	10	0.00	99.96	1.006	0.4144	0.00	99.96	1.006	0.4144

Table 2: Results of PCR and PLSR models for simulated data for large sample sizes ($n = 150, \dots, 1000$).

Sample Size (n)	No. of Comp. (C)	PCR				PLSR			
		% of Variance in Y	Cum % Var. (R^2)	RMSECV	RMSE	% of Variance in Y	Cum % Var. (R^2)	RMSECV	RMSE
150	1	98.80	94.08	10.78	122.5806	99.18	96.73	8.103	63.2367
	2	0.05	95.66	9.623	28.0013	0.57	99.70	2.553	0.6639
	3	0.17	97.22	7.893	0.6076	0.14	99.83	1.899	0.2034
	4	0.45	99.49	3.352	0.4651	0.05	99.91	1.390	0.3405
	5	0.09	99.67	2.720	0.0389	0.01	99.94	1.171	0.0417
	6	0.00	99.67	2.633	0.0210	0.00	99.95	1.083	0.0063
	7	0.01	99.67	2.654	0.0054	0.00	99.95	1.041	0.1103
	8	0.16	99.78	2.213	1.9018	0.00	99.95	1.040	0.1285
	9	0.22	99.95	1.043	0.1162	0.00	99.95	1.028	0.1164
	10	0.00	99.95	1.026	0.1110	0.00	99.95	1.026	0.1110
200	1	97.75	97.75	6.653	18.8843	98.64	98.64	5.216	12.7675
	2	0.41	98.16	6.080	21.9458	1.05	99.69	2.576	0.9138
	3	0.06	98.22	6.136	28.3692	0.13	99.82	1.967	0.0124
	4	1.08	99.30	4.041	0.0000	0.12	99.94	1.125	0.8485
	5	0.37	99.67	2.627	0.0007	0.01	99.95	1.015	1.3711
	6	0.00	99.67	2.641	0.0023	0.01	99.96	0.969	1.2117
	7	0.00	99.67	2.657	0.0053	0.00	99.96	0.958	1.4623
	8	0.08	99.75	2.325	0.0000	0.00	99.96	0.958	1.4748
	9	0.21	99.96	0.958	1.5244	0.00	99.96	0.956	1.1654
	10	0.00	99.96	0.959	0.8716	0.00	99.96	0.959	0.0011
300	1	97.47	97.47	6.546	165.5247	98.52	98.52	5.045	97.7561
	2	0.52	97.99	5.863	60.7870	1.08	99.60	2.678	13.4256
	3	0.00	97.99	5.948	60.7928	0.19	99.79	1.951	17.6239
	4	0.67	98.66	4.875	141.1364	0.10	99.89	1.392	0.7131
	5	0.83	99.49	2.987	30.3948	0.04	99.93	1.165	1.8413
	6	0.03	99.52	2.910	26.2920	0.00	99.93	1.085	0.6469
	7	0.00	99.52	2.964	26.8651	0.01	99.94	1.083	0.5410
	8	0.18	99.70	2.335	21.9822	0.00	99.94	1.08	0.5302
	9	0.23	99.93	1.083	0.5325	0.00	99.94	1.067	0.7159
	10	0.01	99.94	1.059	0.6828	0.00	99.94	1.059	0.6828
500	1	97.07	97.07	7.454	4.0964	98.37	98.37	5.576	6.6965
	2	0.01	97.08	7.501	3.9374	1.36	99.73	2.271	23.9988
	3	2.34	99.42	3.362	26.6909	0.15	99.88	1.564	15.5900
	4	0.08	99.50	3.094	26.7947	0.05	99.93	1.186	10.9132
	5	0.14	99.64	2.657	24.8838	0.01	99.94	1.081	10.0365
	6	0.00	99.64	2.662	25.6810	0.00	99.94	1.060	9.3783
	7	0.01	99.65	2.635	34.0552	0.00	99.94	1.048	9.3793
	8	0.08	99.73	2.388	15.9034	0.00	99.94	1.047	10.0724
	9	0.21	99.94	1.048	10.0386	0.00	99.94	1.047	9.7564
	10	0.01	99.95	1.036	9.0040	0.01	99.95	1.036	9.0040
1000	1	97.64	97.64	6.402	9.5727	98.60	98.60	4.934	6.2822
	2	0.00	97.64	6.421	12.8692	1.06	99.66	2.434	0.1913
	3	1.24	98.88	4.422	4.5090	0.18	99.84	1.679	0.0778
	4	0.29	99.17	3.807	0.0444	0.09	99.93	1.109	0.2096
	5	0.33	99.50	2.958	0.4851	0.01	99.94	1.029	2.5553
	6	0.02	99.52	2.888	2.9986	0.00	99.94	1.012	2.7005
	7	0.01	99.53	2.887	2.9141	0.00	99.94	1.002	2.7884
	8	0.08	99.61	2.632	1.4897	0.00	99.94	1.001	2.6075
	9	0.33	99.94	1.003	2.6515	0.00	99.94	0.999	2.6774
	10	0.01	99.95	0.985	2.3992	0.01	99.95	0.985	2.3992

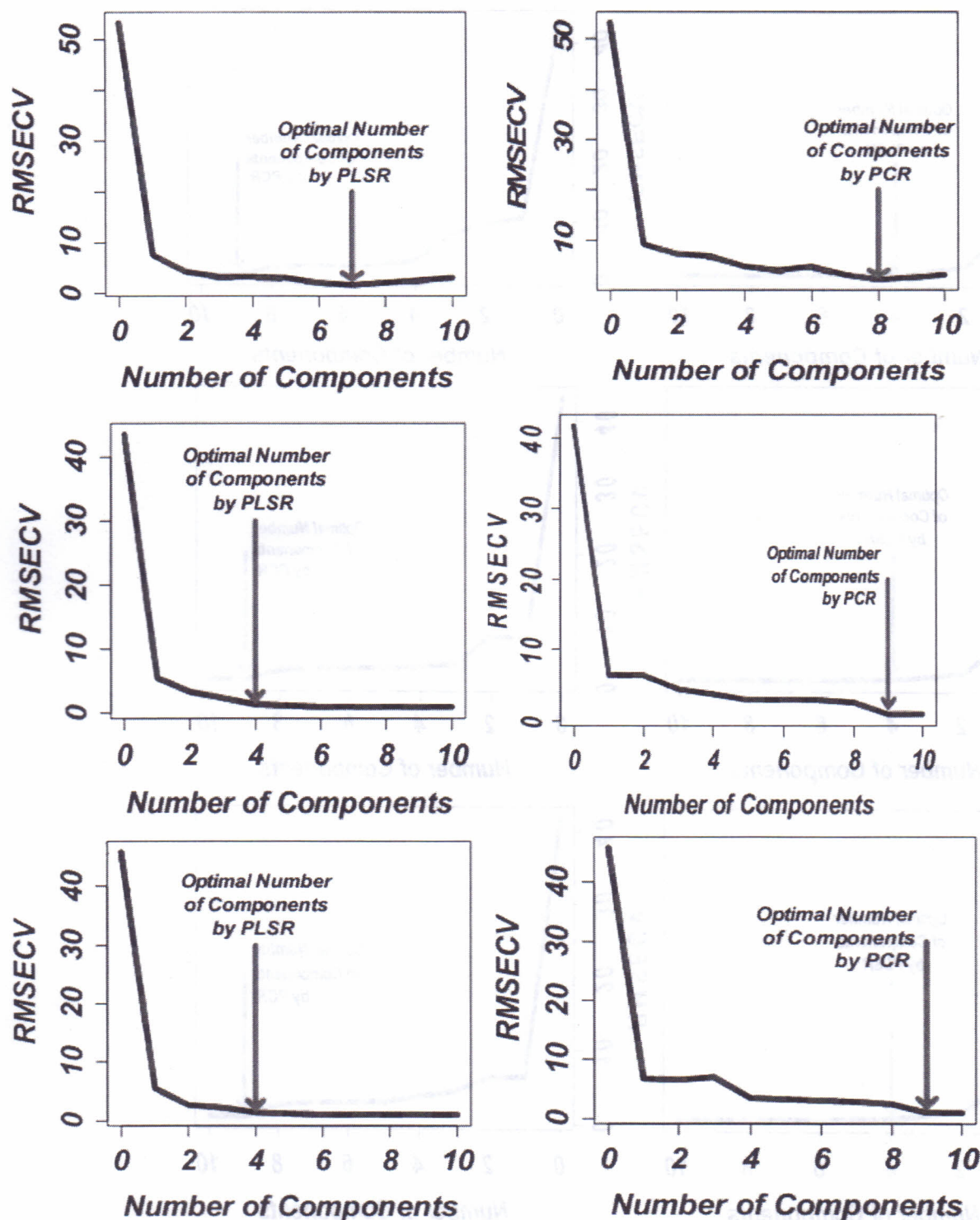


Fig.1: The Scree plots of RMSECV against the number of components used by both the PLSR (left graphs) and PCR (right graphs) models at small sample sizes $n = 20$ (top two graphs), $n = 80$ (middle two graphs) and $n = 100$ (bottom two graphs). The optimal number of components for each model at different sample sizes are indicated by vertical arrows on the graphs.

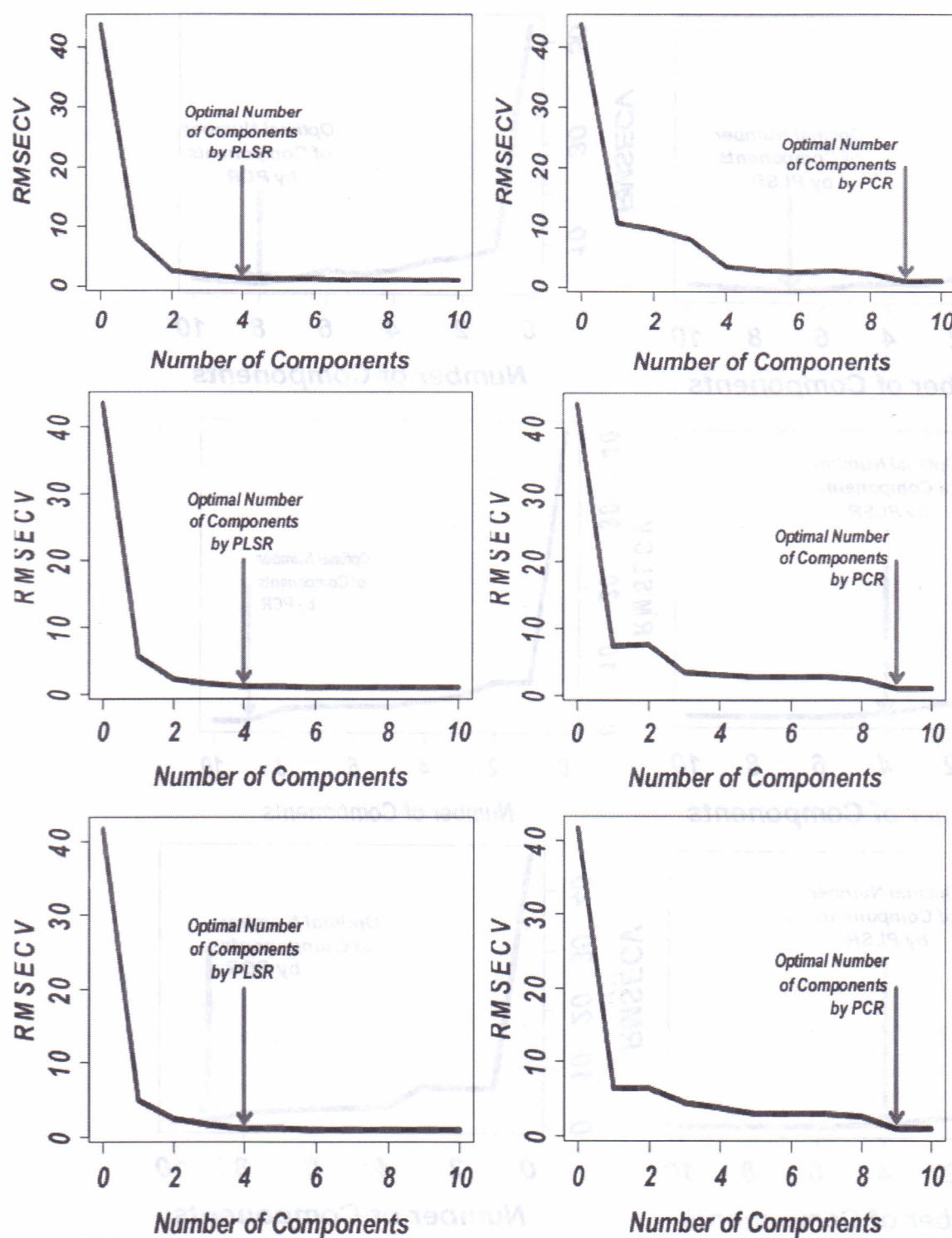


Fig.2: The Scree plots of RMSECV against the number of components used by both the PLSR (left graphs) and PCR (right graphs) models at large sample sizes $n=150$ (top two graphs), $n=500$ (middle two graphs) and $n=1000$ (bottom two graphs). The optimal number of components for each model at different sample sizes are indicated by vertical arrows on the graphs.

In terms of the optimal number of latent components required by the two models to attain the best prediction results, the scree plots of the estimated RMSECV of the models against the number of components used in Figs 1 and 2 show that the PLSR model generally requires fewer numbers of latent components to attain better performance than the PCR model. For instance,

the PLSR model attained its best performance at the seventh component and at sample size $n = 20$ as shown by the graph in Fig 1 (top-left graph). This is the number of components at which the model produces the least RMSECV of 1.773 (see Table 1 at $n = 20$, $C = 7$) where the elbow of the graph almost touches the x-axis before showing evidence of upward movement again as is evident in Fig 1. Thus, the optimal number of components desirable by PLSR model for the data is 7. At this optimal level of performance, the amount of variance in the response variable Y explained by the seven components in the model (R^2) is 99.98% as reported in Table 1. Generally, the choice of optimal component of a model is based on the local minimum rather than the absolute minimum in the curve to avoid *over-fitting*.

However, for the same data set with sample size $n = 20$ as used in PLSR model, the PCR model attained its best performance at the eighth component having the least RMSECV of 2.233 and R^2 of 99.97%. Thus, the optimal number of components of the PCR model for the data is 8 as is evident in Fig 1 (top-right graph).

Evidently, the same good feat of performance above was repeated by PLSR model at sample size $n = 40$ with the optimal number of component $C = 7$ which yielded R^2 of 99.97% as shown in Table 1. Whereas, the optimal number of components of the PCR model increases by 1, from 8 to 9 as the sample size increases from 20 to 40. However, at subsequent sample sizes from $n = 60$, 80 and 100 the optimal number of PLSR components were 6, 4 and 4 respectively, while the optimal number of PCR components at these sample sizes remained 9.

Without loss of generality, it is obvious that both the PLSR and PCR model stabilize at larger sample sizes as evident from the results in Table 2 and the scree plots in Fig 2. While the PLSR model achieves its optimal performance using 4 latent components (with R^2 of about 99.9%) at large sample sizes $n > 100$ as shown by the graphs in Fig 2, the PCR attained the same feat at a relatively higher number of 9 components.

The expected orthogonality condition [35] among the constructed PLSR and PCR components was examined. This is simply achieved by the pair-wise plots of the constructed linear components to ascertain whether any systematic pattern could be noticed on the resulting graphs to suggest possible relationships among the components. The pair-wise plots of the scores of the first four PLSR and PCR components are presented in Fig 3 for clarity. As can be observed in Fig 3, each of the pair-wise plots of latent component scores from the two regression models do not show any evidence of systematic pattern that could suggest that they are correlated in any way.

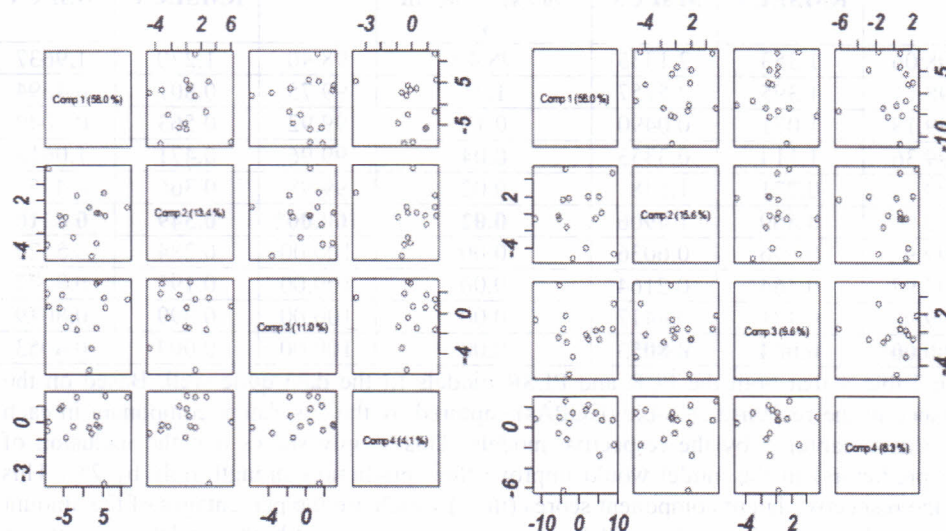


Fig 3: The pair-wise plots of component scores of the first four PLSR (left graph) and PCR (right graph) latent components. All the plots show that the the latent components are purely ethodogonal (uncorrelated).

5.0 Applications to Real Life Data

Applications of both the PLSR and PCR methods on real life data sets are presented in this section.

The datasets considered were secondary data collected from the Department of Statistics, University of Ilorin, Ilorin, Nigeria on results of 19 graduate students of Statistics who enrolled and completed their Master of Science (M.Sc.) in Statistics during the 2011/2012 academic session. The data consisted of final Cumulative Grade Point Average (CGPA) and raw scores of these 19 students in all the 10 courses offered during their M.Sc. programme in the department. The PCR and PLSR methods were employed on the data by regressing the final CGPAs of the students on their respective scores in all the 10 courses offered (predictors). The correlation matrix of the raw scores in all the ten courses (predictor variables) revealed evidence of multicollinearity in the data with pair-wise correlation coefficients ranging between 0.42 and 0.83 as presented in Table 3 (i.e. $0.42 \leq \hat{\rho}_{\square\square'} \leq 0.83, \square \neq \square', \square, \square' = 1, 2, \dots, 10$).

Table 3: The correlation matrix of scores obtained by the students in all the ten courses offered by 19 M.Sc. Statistics graduates of the University of Ilorin. The values reported in the table are the computed Pearson correlation coefficients among the variable pairs. Significant correlation between a pair of courses is taken at 5% level ($\alpha < 0.05$). The asterisk (*) indicate correlation coefficients that are not statistically significant ($\alpha \geq 0.05$).

Courses	STA651	STA652	STA653	STA656	STA657	STA658	STA659	STA661	STA672	STA699
STA651	1.000									
STA652	0.424*	1.000								
STA653	0.660	0.578	1.000							
STA656	0.786	0.446*	0.548	1.000						
STA657	0.515	0.658	0.478	0.528	1.000					
STA658	0.534	0.353*	0.724	0.403*	0.288*	1.000				
STA659	0.663	0.474	0.465	0.598	0.459	0.515	1.000			
STA661	0.614	0.645	0.833	0.591	0.481	0.669	0.511	1.000		
STA672	0.782	0.375*	0.817	0.691	0.282*	0.648	0.539	0.729	1.000	
STA699	0.569	0.525	0.467	0.558	0.209*	0.377*	0.551	0.401*	0.579	1.000

In order to build efficient PCR and PLSR models for these data, all the 19 samples were randomly partitioned into two sets comprising a test sample of size 5 and a training sample of size 14 as adopted elsewhere [16,36-40]. The models were built on the training data using internal cross-validation technique of the LOOCV to ensure models' stability [14]. The efficiency of the two models were examined on the test data using the MSE, R^2 , MSEC and the number of latent components utilized by each model to achieve the same R^2 . Details of these results are presented in Tables 4 to 6.

Table 4: Results of PCR and PLSR models on M.Sc. students Results' data. The performance level at which the two models attained 100% R^2 value are asterisked.

No. of Comp.	PCR				PLSR			
	% exp. Var. in Y	R^2	RMSEC	MSEC	% exp. Var. in Y	R^2	RMSEC	MSEC
1	98.06	98.06	1.353	2.1335	98.40	98.40	1.270	1.9637
2	0.05	98.11	1.395	2.5757	1.35	99.75	0.801	0.4694
3	1.07	99.18	1.071	0.0490	0.17	99.92	0.563	0.7049
4	0.18	99.36	1.114	0.3353	0.04	99.96	0.471	1.0618
5	0.38	99.74	0.771	1.8481	0.02	99.98	0.366	0.4955
6	0.16	99.90	0.569	1.4906	0.02	100.00*	0.349	0.4310
7	0.02	99.92	0.596	0.6036	0.00	100.00	0.284	0.5525
8	0.05	99.97	0.464	0.2164	0.00	100.00	0.195	0.8277
9	0.01	99.98	0.491	0.4417	0.00	100.00	0.099	0.8019
10	0.02	100.00*	0.004	0.8053	0.00	100.00	0.004	0.8053

It can be observed from the results in Table 4 that both the PCR and PLSR models fit the data quite well. Based on the training data used, the amount of variance in the response variable (CGPAs) captured by the first latent component in each model was about 98% of the total variance captured by the respective models. This simply shows that the inclusion of additional nine components (as latent predictors) in the model would improve their prediction strength only by 2%. This result is justified by the scree plots of the respective latent component scores (in %) which are the percentages of the amount of variance in the students' CGPAs captured against the number of latent components employed by the models as shown in Fig 4 for both the PCR (*red solid graph*) and PLSR (*black dotted graph*) models.

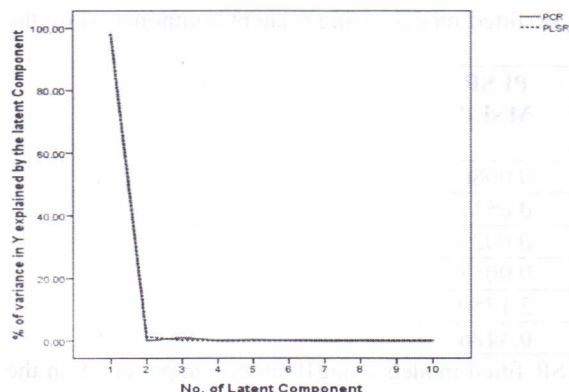


Fig. 4: The scree plot of the PLSR and PCR latent component scores (in %) versus the number of components employed by the models. The first component in the two models (component 1) provided about 98% of the total variation explained by the models while all the other nine components (components 2 to 10) accounted for the remaining 2%.

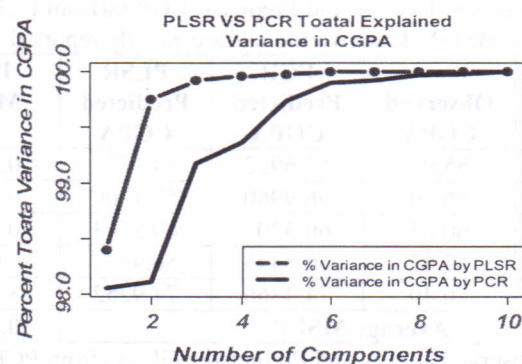


Fig. 5: Plot of the percentage of total variance in CGPA explained by PLSR (red dotted line) and PCR (blue solid line) models at different number of components employed by the models.

However, further assessment of the results in Table 4 revealed that the PLSR model seems relatively more efficient than the PCR model as it yields 100% R^2 using just six latent components as against the PCR model that attained the same feat using all the constructed ten components. It can be observed from the results in the table that the values of R^2 yielded by the PLSR at all the latent components employed are all higher than that of the PCR model. This performance behavior of the two models is made clearer by plotting the R^2 (in %) of each of the models against the number of components employed as shown in Fig 5 where it is obvious that the R^2 of the PLSR model are higher than that of the PCR at all the numbers of components employed.

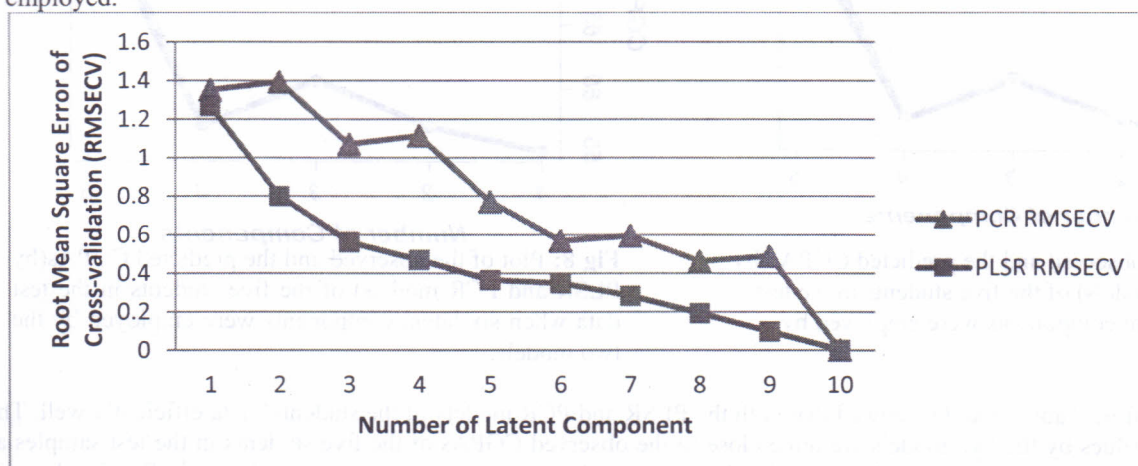


Fig. 6: The plots of the estimated RMSECV against the number of latent components employed by PCR and PLSR models for students' results data.

Additional results on the performances of the PLSR and PCR models on the training data using the RMSECV further showed the relative efficiency of the PLSR over the PCR models. The graphs of the estimated RMSECV by the two models in Fig 6 showed that the RMSECV values of the PLSR model were relatively smaller than that of the PCR model at all the numbers of components employed (except at ten components). Generally, apparent decrease in the estimated RMSECV values as more components were employed by the two models is evident as shown by the two graphs in Fig 6.

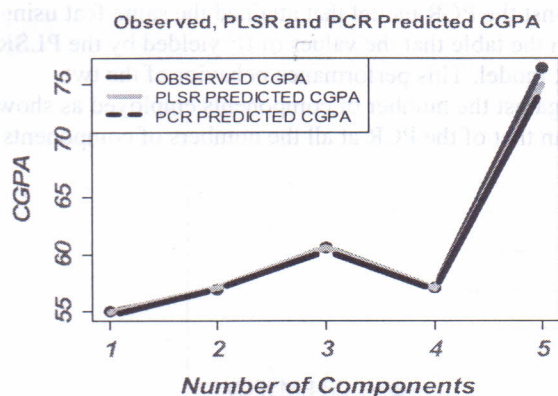
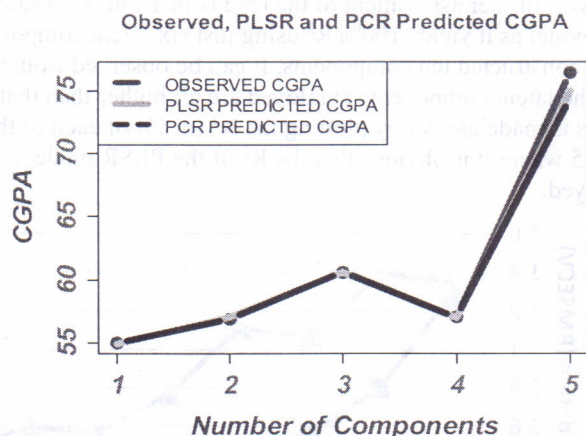
To evaluate the relative performance of the two models on the test (unused) data, each model, as fitted using the 14 training samples, was used to predict the CGPA scores of the five students in the test samples. The observed CGPAs in the test data and their predicted CGPA values by both the PLSR and PCR models are presented in Table 5 (when the two models were fitted using six optimal latent components of the PLSR model) and in Table 6 (when the two models were fitted using all the ten latent components). The mean square errors of predictions (MSEP) yielded by the two models from these predictions are equally reported in Tables 5 and 6.

Table 5: Observed CGPAs and Predicted CGPAs (from PCR and PLSR fitted models using 6 latent components) on the test samples. The MSE of the two models are equally reported.

Test Sample Number	Observed CGPA	PCR Predicted CGPA	PLSR Predicted CGPA	PCR MSE	PLSR MSE
1	55.07	54.6922	54.9771	0.1427	0.0086
2	56.90	56.9960	57.1260	0.0093	0.0511
3	60.63	60.3207	60.5763	0.0957	0.0029
4	57.07	56.8445	56.9974	0.0509	0.0053
5	76.40	74.4866	74.9252	3.6611	2.1750
Average MSE				0.7919	0.4486

Table 6: Observed CGPA and Predicted CGPAs (from PCR and PLSR fitted models using 10 latent components) on the test samples. The MSE of the two models are equally reported.

Test Sample Number	Observed CGPA	PCR Predicted CGPA	PLSR Predicted CGPA	PCR & PLSR MSE
1	55.07	55.06828	55.06825	0.000003
2	56.90	56.89756	56.89756	0.000006
3	60.63	60.63403	60.63403	0.000016
4	57.07	57.06250	57.06250	0.000056
5	76.40	74.40095	74.40095	3.996200
Average MSE				0.799256

**Fig 7:** Plot of the observed and the predicted CGPAs (by PLSR and PCR models) of the five students in the test data when six latent components were employed by the two models.**Fig 8:** Plot of the observed and the predicted CGPAs (by PLSR and PCR models) of the five students in the test data when six latent components were employed by the two models.

Prediction results from Tables 5 and 6 showed that both the PLSR and PCR models fit the students' data efficiently well. The predicted CGPA values by the two models are quite close to the observed CGPAs of the five students in the test samples as clearly shown by the plot of the predicted CGPAs by the two models on all five test samples as shown in Fig 7 (when six latent components are employed by the two models) and Fig 8 (when ten latent components are employed by the two models). This good out-of-sample prediction performance of the two models resulted into low estimates of their MSE which is far below one unit in four of the five test samples.

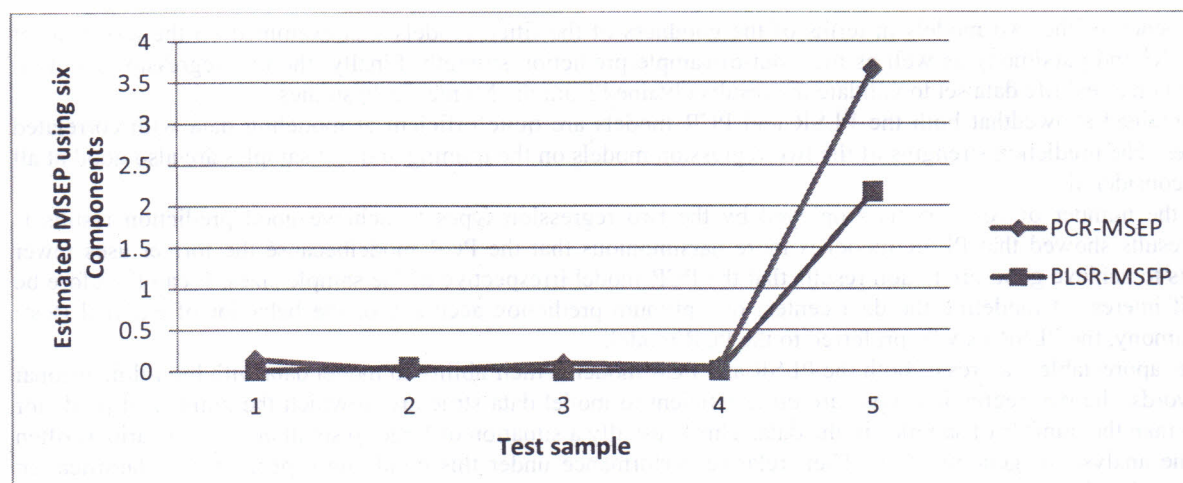


Fig 9: Line graphs of the estimated MSE of PLSR and PCR models using six latent components against the test sample data points.

This excellent performance of the two models notwithstanding, it can still be observed that the PLSR model yielded better prediction results on the test sample than the PCR model as evident in Table 5 when six latent components were employed by the two models. Not only this, the overall average MSE of 0.4486 yielded by the PLSR model is relatively smaller than 0.7919 which was provided by the PCR model for the same data. The superiority of the PLSR over the PCR in terms of out-of-sample prediction performance is made clear by the graphs in Fig 9 which are the plots of the estimated MSEs by both the PLSR and PCR models at each of the test sample data points. From the graphs, it can be observed that the MSE of the PLSR model is smaller than that of the PCR model in all the five test sample cases.

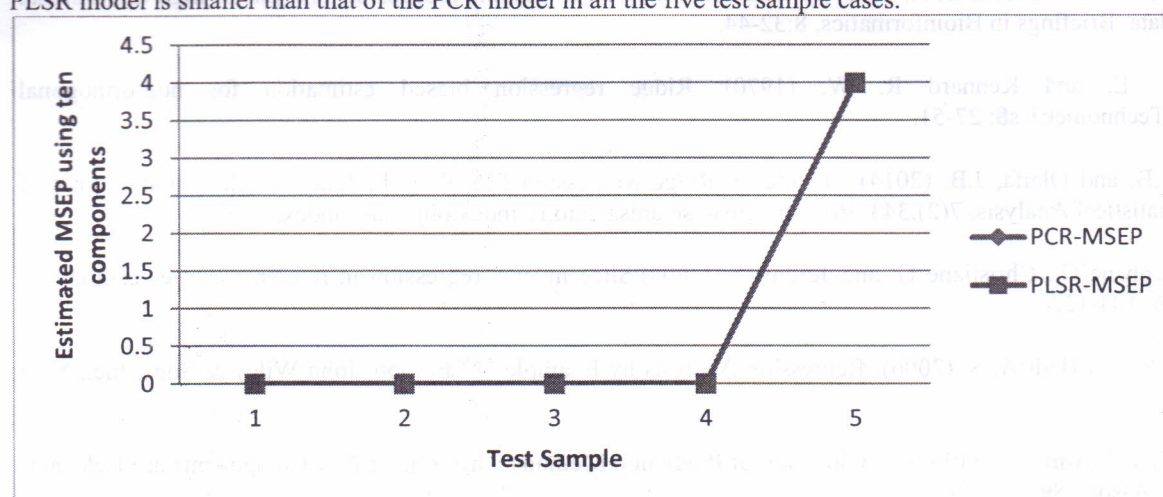


Fig 10: Line graphs of the estimated MSE of PLSR and PCR models using ten latent components against the test sample data points.

Expectedly, when the two models were fitted using all the ten latent components (the optimal number of components desired by PCR model here), the out-of-sample predictions of the two models improved on four of the five test samples as evident from the MSE of PLSR and PCR in Table 6 as compared to that of Table 5 in which only six components were employed by the two models. In fact, both the PLSR and PCR models yielded exactly the same MSE on all the five test samples when ten components were employed by the two models as presented in Table 6 and shown by the line graph in Fig 10.

However, further examination of the results in Tables 5 and 6 revealed that the average out-of-sample prediction performance of PLSR model using six latent components (average MSE of 0.4486) in Table 5 is still better (smaller) than the average MSE of the PCR /PLSR model (average MSE of 0.799256) in Table 6 using ten components. It can therefore be concluded that both regression methods perform better in terms of modeling the response CGPA of students on their respective scores in various courses, but the PLSR model achieves better prediction accuracy more parsimoniously than the PCR model.

6.0 Conclusion

The theoretical formulations and applications of both the PLSR and PCR models are presented in this work. Detailed comparative study of the performances of the two models was undertaken via Monte Carlo experiments.

The relative efficiency of the two models in terms of the goodness of the fitted models was examined on the test data set using RMSECV, R^2 and parsimony as well as their out-of-sample prediction strength. Finally, the two regression methods were implemented on a real life data set to validate the results obtained from the Monte Carlo studies.

All the results obtained showed that both the PLSR and PCR models are quite efficient at modeling data with correlated predictor variables. The prediction strengths of the two regression models on the training and test samples are also good at all the sample sizes considered.

However, when the number of components employed by the two regression types to achieve good prediction results is considered, the results showed that PLSR model is more parsimonious than the PCR model because the former uses fewer latent components to achieve good prediction results than the PCR model irrespective of the sample sizes. It can therefore be concluded that, if interest of modeling the data centers on optimum prediction accuracy of the behavior of external (test) samples and parsimony, the PLSR may be preferred to the PCR model.

Finally, one of the appreciable features of both the PLSR and PCR model is their ability to model data with high-dimensional space. In other words, the two regression types are quite efficient to model data structure in which the number of predictor variables is more than the number of samples in the data. This is usually a situation of larger p small n ($n < p$) scenario as often encountered in the analysis of genomic data. Their relative performance under this condition especially for classification problems shall be undertaken in future works.

7.0 References

- [1] Martens H., T. Næs (1989): *Multivariate Calibration*: John Wiley & Sons, New York.
- [2] Helland I.S. (2000): Model reduction for prediction in regression model. *Scand. J. Stat.*, 27: 1-20.
- [3] Boulesteix, A-L and Strimmer, K. (2007): Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8:32-44.
- [4] Hoerl, A. E. and Kennard R. W. (1970): Ridge regression biased estimation for non-orthogonal problems. *Technometrics* 8: 27-51.
- [5] Yahya, W.B. and Olaifa, J.B. (2014): A note on Ridge Regression Modeling Techniques. *Electronic Journal of Applied Statistical Analysis*, 7(2):343-361 <http://siba-es.unisalento.it/index.php/ejasa/index>.
- [6] Ali G., Stephane G., Christiane G. and Jerome S. (2004): Slice inverse regression in reference curves estimation: *Elsevier* 46: 103-122.
- [7] Chatterjee S. and Hadi A. S. (2006): *Regression Analysis by Example*. 4th Edition, John Wiley & Sons, Inc., New York.
- [8] Helland I.S. and Almoy T. (1994): Comparison of Prediction Methods when only a Few Components are Relevant: *J. Am. Stat. Assoc.*, 89: 583-591.
- [9] Phatak, A. and Jong, S. D. (1997): The geometry of Partial Least Squares: *Journal of Chemometrics* 11: 311-338.
- [10] Wold H. (1985): Partial Least Squares, in: S. Kotz, N.L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, Wiley, New York, 6: 581-591.
- [11] de Jong S. (1993): SIMPLS: An Alternative Approach to Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, 18: 251-263.
- [12] Wold S. (2001): Personal Memories of the Early PLS Development: *Chemometrics and Intelligent Laboratory Systems*, 58(2): 83-84.
- [13] Mevik B-H, Cederkvist H. R. (2004): Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR), *Journal of Chemometrics*, 18(9):422-429.

- [14] Mevil, B-H and Wehrens, R. (2007): The pls package: Principal component and partial least squares regression in R: Journal of Statistical Software, 18(2):1-18.
- [15] Yeniyay O. and Gokta A. (2002): A comparison of partial least squares regression with other prediction methods. Journal of Mathematics and Statistics, 31: 99-111.
- [16] Yahya WB, Oladiipo MO and Jolayemi ET (2012): A fast algorithm to construct neural networks classification models with high-dimensional genomic data. Annals. Computer Science Series, 10 (1): 39-58.<http://anale-informatica.tibiscus.ro/>
- [17] Myers, R. H.(1986): Classical and Modern Regression with Applications. PWS-KENT Publishing Company, Massachusetts.
- [18] Karakach, T.K. (2002): Comparison of Linear and Nonlinear Multivariate Calibration Methods. M.Sc. thesis, Dalhousie University, Halifax, Canada.
- [19] Yahya, W.B., Adebayo, S.B., Jolayemi, E.T., Oyejola, B.A. and Sanni, O.O.M. (2008): Effects of non-orthogonality on the efficiency of seemingly unrelated regression (SUR) models. InterStat Journals, 1-29.
- [20] Kendall, M. G. (1957): A Course in Multivariate Analysis. London: Griffin.
- [21] Hotelling, H. (1957) The relations of the newer multivariate statistical methods to factor analysis. Brit. J. Stat. Psychol., 10, 69-79.
- [22] Massy, W.F. (1965): Principal Component Regression in Exploratory Statistical Research: Journal of the American Statistical Association 60: 234-246.
- [23] Wentzell, P. D. and Montono L.V. (2003): Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures: Chemometrics and Intelligent Laboratory Systems 65: 257-279.
- [24] Venkatesan, P., Dharuman, C. and Gunasekaran, S. (2011): A comparative study of principal component regression and partial least squares regression with application to FTIR diabetes data. Indian Journal of Science and Technology, 4(7): 740-746.
- [25] Gunst, R. F. and Mason, R. L. (1979); Some considerations in the evaluation of alternate prediction equations, Technometrics, 21: 55-63.
- [26] Helland I.S. (2011): Some theoretical aspects of partial least squares regression. Chemometrics and intelligent laboratory systems, 58: 97-107.
- [27] Ozgur Y. and Atilla G. (2002): A Comparison of Partial Least Squares Regression with other Prediction Methods: Hacettepe Journal of Mathematics and Statistics 31: 99-111.
- [28] Vajargah, B. F. and Moradi, M. (20012): Diagonal scaling of ill-conditioned matrixes by genetic algorithm. Journal of Applied Mathematics, Statistics and Informatics, 8 (1):49-53. DOI: 10.2478/v10294-012-0005-3
- [29] Jolliffe, I. T. (1982): A Note on the Use of Principal Components in Regression. Appl. Statist.,31(3): 300-303
- [30] Draper, N.R. and Smith, H. (1981): Applied regression analysis. 2nd edition. John Wiley and Sons, New York.
- [31] Garthwaite, P.H. (1994): An interpretation of Partial Least Squares, Journal of the American Statistical Association, 89: 122-127
- [32] Nguyen, D.M. and Rocke, D. M. (2004): On partial least squares dimension reduction for microarray-based classification: a simulation study. Computation Statistics & Data Analysis,46: 407-425.

- [33] Ergon R. (2006): Reduced PCR/PLSR models by subspace projections. *Chemometrics and Intelligent Laboratory Systems* 81: 68-73.
- [34] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.
- [35] Martens H. (1985): Multivariate Calibration: Dr. technical thesis: Technical University of Norway, Trondheim.
- [36] Yahya, W.B. (2009): Sequential dimension reduction and prediction methods with high-dimensional microarray data: Universitätsbibliothek, Ludwig-Maximilians-Universität, München, Germany. Ph.D. Thesis. URL: <http://edoc.ub.uni-muenchen.de/10254/>
- [37] Yahya W.B. Ulm K. Fahrmeir L. and Hapfelmeier A. (2011): k-SS: A sequential feature selection and prediction method in microarray study: *Int'l. Journal of Artificial Intelligence*, 6 (S11): 19-47. <http://www.ceser.in/ceserp/index.php/ijai/article/view/912>.
- [38] Hapfelmeier A, Yahya WB, Rosenberg R and Ulm K. (2012): Predictive modelling of gene Expression data. In: *Handbook of Statistics in Clinical Oncology*, 3rd ed., Edited by Crowley, J. and A. Hoering. Chapman and Hall/CRC, New York: 463-475.
- [39] Yahya WB (2012): Genes selection and Tumour Classification in Cancer Research: A new approach. Lambert Academic Publishing, Säbruck, Germany.
- [40] Yahya WB, Rosenberg R and Ulm K (2014): Microarray-based Classification of Histopathologic Responses of Locally Advanced Rectal Carcinomas To Neoadjuvant Radiochemotherapy Treatment. *Turkiye Klinikleri Journal of Biostatistics*, 6(1):8-23