doi:10.20944/preprints202111.0549.v1

PCR, PLS, or OPLS

Evaluation of different regression techniques for hypothesis generation

Avani Ahuja

Affiliation: Georgetown Day School, 4200 Davenport St NW, Washington, DC 20016

Abstract

In the current era of ‘big data’, scientists are able to quickly amass enormous amount of data in a limited number of experiments. The investigators then try to hypothesize about the root cause based on the observed trends for the predictors and the response variable. This involves identifying the discriminatory predictors that are most responsible for explaining variation in the response variable. In the current work, we investigated three related multivariate techniques: Principal Component Regression (PCR), Partial Least Squares or Projections to Latent Structures (PLS), and Orthogonal Partial Least Squares (OPLS). To perform a comparative analysis, we used a publicly available dataset for Parkinson’s disease patients. We first performed the analysis using a cross-validated number of principal components for the aforementioned techniques. Our results demonstrated that PLS and OPLS were better suited than PCR for identifying the discriminatory predictors. Since the X data did not exhibit a strong correlation, we also performed Multiple Linear Regression (MLR) on the dataset. A comparison of the top five discriminatory predictors identified by the four techniques showed a substantial overlap between the results obtained by PLS, OPLS, and MLR, and the three techniques exhibited a significant divergence from the variables identified by PCR. A further investigation of the data revealed that PCR could be used to identify the discriminatory predictors successfully if the number of principal components in the regression model were increased. In summary, we recommend using PLS or OPLS for hypothesis generation and systemizing the selection process for principal components when using PCR.

Keywords: Principal Component Regression, Partial Least Squares, Orthogonal Partial Least Squares, multivariate regression, hypothesis generation, Parkinson’s disease

Introduction

Many scientific investigations involve generating experimental data for multiple input or dependent variables (termed as predictors) with the hope that one would be able to find the predictors responsible for the observed trends in the response variable. Multiple linear regression (MLR) is a common technique that is used widely to establish a relationship between the response variable and predictors. However, a prerequisite to performing MLR is that the predictors should be orthogonal to each other. Also, to perform MLR, the number of observations must exceed the number of predictors. The MLR technique is primarily concerned with explaining the variation in Y as a function of X, it however does not address collinearity that may exist amongst predictors. The condition of orthogonality can be achieved in systematically designed experiments but is not possible in cases where multidimensional data is generated with limited number of experiments or when predictors are correlated to each other. In such situations, to enable a rigorous analysis, the correlated variables can be transformed to a new multivariate space containing uncorrelated variables known as principal components. This transformation helps reduce the number of necessary dimensions (or the new primary axes) for describing the data and facilitates an enhanced understanding of correlations within predictors and relationships between the predictors and the response variable. In this article, we explored three relevant techniques, PCR, PLS, and OPLS. We chose these techniques as they are all based on notion of latent variables, the variables that are linear combinations of the original variables. The brief descriptions of these techniques are described below:

PCR: In this methodology, Principal Component Analysis (PCA) (Jackson, 1991) is first applied to the X data to derive the latent variables or principal components. The first principal component is chosen so that the R^2X value is maximized. The successive principal components must be orthogonal to the preceding ones and try to maximize R^2X with every iteration. Note that the principal components are linear combinations of the predictors and are mathematically related to them by the corresponding loading values. The loading values can be interpreted as the coefficients in the said linear combination and vary between -1 and 1. Also, the observations in the new multivariate space are characterized by the scores associated with principal components where the scores can be interpreted as the coordinates of observations in the new multivariate space containing principal components as the primary axes.

After principal components are selected, they are subjected to MLR to generate a PCR model. Note that the derivation of orthogonal components allows the application of MLR since the principal components are orthogonal to each other. Also, in situations where the number of predictors far exceed the number of observations, reducing the number of dimensions allows generating a regression model that would not have possible with the original data. For a detailed review of the technique, the reader is directed to some of the earlier publications on the subject (Jeffers, 1967; Hawkins, 1973; Mansfield *et al.*, 1977).

PLS: The PLS technique attempts to maximize the X-Y covariance, while assuming the existence of a small number of latent variables in the X-data that predict the response variable. In other words, this technique involves maximizing the relationship between X and Y data while maintaining the correlation amongst predictors at the same time. Mathematically, the principal components in PLS are chosen so that at each step, the algorithm finds a new principal component that maximizes the product of the variance of the predictors multiplied by the square of their correlation to the response variable (Hastie *et al.*, 2017a). Please refer to relevant publications (Wold *et al.*, 1993; Wold, 2001) for statistical details about the technique.

OPLS: Consistent with the PLS methodology, OPLS technique aims to maximize the relationship between X and Y while retaining latent X variables; additionally, it also attempts to filter out the X information that is unrelated to Y. Mathematically, the OPLS method uses a technique named orthogonal signal correction (Wold *et al.*, 1998), where the first principal component, termed as the predictive component, maximizes available X-Y covariance. The succeeding principal components capture variance in the orthogonal predictors, i.e. the ones that are not statistically correlated to the response variable. While PLS does a reasonable job in reducing random noise in the data, the OPLS technique allows removing the structured noise present in X data that is uncorrelated to Y. This helps simplify the model by reducing the number of principal components and allows the analysis of the main sources of orthogonal variation (Goueguel, 2019). For details about the technique, please refer to (Eriksson *et al.*, 2006a).

Although PLS is widely used for data mining purposes and is offered in a number of data mining platforms, the equivalent cannot be said about PCR and OPLS. The objective of this work was to compare these techniques and provide recommendations regarding which technique should be used for hypothesis generation.

Materials and Methods

The data used in this article was borrowed from a Kaggle data base (Kaggle Inc., 2021) and originally belonged to the work performed by Hlavnička *et al*. (2017), where the researchers showed that latent parkinsonian speech aberrations can be captured even in patients with Rapid Eye Movement (REM) behavior disorder. The dataset included 30 patients with early untreated Parkinson's disease (designated as PD), 50 patients with REM sleep behavior disorder (designated as RBD) that were at high risk for developing Parkinson's disease or other synucleinopathies, and 50 healthy controls (designated as HC). A professional neurologist with experience in movement disorders examined the patients and provided them with a clinical score, Unified Parkinson’s Disease Rating Scale (UPDRS) score (designated as UPDRS III total (-)). The patients were also evaluated by a speech specialist where they read standardized, phonetically balanced text of 80 words and monologized about their current activities, interests, family, or job for approximately 90 seconds. Table 1 lists all the predictors in the data set that were used for this work. Note that the data for three of the predictors available in the dataset - Antiparkinsonian medication, Antipsychotic medication, and Levodopa equivalent (mg/day) - were not used as that did not contain significant variability between patients. The current work also ignored the 50 HC observations since they did not contain the corresponding Y (UPDRS III total (-)) data. Also, Hoehn & Yahr scale (-) response was not included as a Y parameter since the related scores were available for the PD patients only. Overall, the modeled data included 33 X and 1 Y variable.

Table 1: List of variables and corresponding numerical designation

Parameter	Variable # (X or Y)	Description	Designation	Variable type	Numerical transformation
Disease category	1 (X)	Parkinson’s disease (PD) or	DIS	Categorical	PD: 1, RBD: 0

Parameter	Variable # (X or Y)	Description	Designation	Variable type	Numerical transformation
		REM sleep behavior disorder (RBD)			
Demographic Information	2 (X)	Age	AGE	Continuous	
	3 (X)	Gender	GND	Categorical	Female: 0, Male: 1
Clinical information	4 (X)	Positive history of Parkinson's disease in family	HPF	Categorical	No: 0, Yes: 1
	5 (X)	Age of disease onset (years)	ADO	Continuous	
	6 (X)	Duration of disease from first symptoms (years)	DDF	Continuous	
Medication	7 (X)	Antidepressant therapy	ANTD		No: 0, Yes: 1
	8 (X)	Benzodiazepine medication	BENZ		No: 0, Yes: 1
	9 (X)	Clonazepam (mg/day)	CLON	Continuous	
Speech examination: speaking task of reading passage	10 (X)	Entropy of speech timing (-)	EST	Continuous	
	11 (X)	Rate of speech timing (-/min)	RST	Continuous	

Parameter	Variable # (X or Y)	Description	Designation	Variable type	Numerical transformation
	12 (X)	Acceleration of speech timing (-/min^2)	AST	Continuous	
	13 (X)	Duration of pause intervals (ms)	DPI	Continuous	
	14 (X)	Duration of voiced intervals (ms)	DVI	Continuous	
	15 (X)	Gaping in-between voiced intervals (-/min)	GVI	Continuous	
	16 (X)	Duration of unvoiced stops (ms)	DUS	Continuous	
	17 (X)	Decay of unvoiced fricatives (‰/min)	DUF	Continuous	
	18 (X)	Relative loudness of respiration (dB)	RLR	Continuous	
	19 (X)	Pause intervals per respiration (-)	PIR	Continuous	

Parameter	Variable # (X or Y)	Description	Designation	Variable type	Numerical transformation
	20 (X)	Rate of speech respiration (-/min)	RSR	Continuous	
	21 (X)	Latency of respiratory exchange (ms)	LRE	Continuous	
Speech examination: speaking task of monologue	22 (X)	Entropy of speech timing (-) (monologue)	EST-M	Continuous	
	23 (X)	Rate of speech timing (-/min) (monologue)	RST-M	Continuous	
	24 (X)	Acceleration of speech timing (-/min^2) (monologue)	AST-M	Continuous	
	25 (X)	Duration of pause intervals (ms) (monologue)	DPI-M	Continuous	
	26 (X)	Duration of voiced intervals (ms) (monologue)	DVI-M	Continuous	
	27 (X)	Gaping in-between voiced intervals (-/min) (monologue)	GVI-M	Continuous	

Parameter	Variable # (X or Y)	Description	Designation	Variable type	Numerical transformation
	28 (X)	Duration of unvoiced stops (ms) (monologue)	DUS-M	Continuous	
	29 (X)	Decay of unvoiced fricatives (‰/min) (monologue)	DUF-M	Continuous	
	30 (X)	Relative loudness of respiration (dB) (monologue)	RLR-M	Continuous	
	31 (X)	Pause intervals per respiration (-) (monologue)	PIR-M	Continuous	
	32 (X)	Rate of speech respiration (-/min) (monologue)	RSR-M	Continuous	
	33 (X)	Latency of respiratory exchange (ms) (monologue)	LRE-M	Continuous	
Clinical score (Overview of motor examination)	1 (Y)	UPDRS III total (-)	UPDRS	Continuous	

We used MATLAB (Version R2019a) to run the PCR model, SIMCA-P (Version 16) to run the PLS and OPLS models, and JMP (Version 15) to run the MLR models. After compiling the predictors in a suitable format, mean-centering and univariate scaling were used for centering and scaling the data ((Eriksson *et al.*, 2006b). To determine the principal components that should be retained in PCA, PLS, and OPLS algorithms, a cross validation approach similar to a methodology described elsewhere (Eastment and Krzanowski, 1982) was employed using predefined rules in SIMCA. The technique involved determination of R^2 and Q^2 values (R^2X and Q^2X for PCA and R^2Y and Q^2Y for PLS and OPLS) for each of the components. The R^2 values reflected the model fit and were estimated by coefficient of determination (R^2) based on the predicted and actual values for given number of principal components. For calculating the Q^2 values, the data was divided into 7 parts and PCA, PLS, or OPLS models were run iteratively seven times using six parts of the data, and the resulting models were used to predict the X values for PCA and Y values for PLS and OPLS for the remainder of the data. Using the original and predicted values for seven subsets, Q^2X and Q^2Y values (1 – PRESS statistic) were calculated; the PRESS statistic here stands for the Predicted Residual Error Sum of Squares and it was calculated based on the original and predicted values for the observations in seven subsets. While the R^2(cum) values continued to increase as the number of principal components increased, Q^2(cum) values first increased and then decreased as the number of principal components increased. This trend became the basis of how the principal components were selected. The R^2(cum) and Q^2(cum) values reflect the cumulative R^2 and Q^2 values using the given number of principal components.

The parameters were ranked in their influence on the models according to the regression coefficients in the MLR models (considering their magnitude as well as their variability) and Variable Influence on Projection (VIP) values in the PLS/OPLS models. The VIP value for a given predictor in a PLS/OPLS model summarizes its contribution to the model and predictors having large VIP values are considered most relevant. The calculation of a VIP value considers the relationship the predictor has with the selected principal components in combination with Y variation that is explained by each of the selected components. For the PCR model, after determining the coefficients in the regression model of Y as a function of different principal components (designated as t1, t2, etc.), referred to as Y-T coefficients herein, where T is the matrix containing the score data of 80 observations on the selected principal components, the X-

Y coefficients were determined by multiplying the Y-T coefficients by the loading matrix (with k x q dimension; k refers to the number of predictors and q refers to the number of selected principal components) that defined the relationship between the X variables and selected principal components.

Results

PCR

Prior to running the PCR regression model, PCA was applied to the X data, resulting in 33 principal components. However, based on the predefined rules in SIMCA, only the first 2 components were retained. Figure 1(a) shows the R^2 and Q^2 values associated with the selected components and Figure 1(b) shows the loadings of different predictors for the first principal component. The X data was clearly correlated considering many predictors had similar absolute loading values. The 5 predictors that influenced the direction of the first principal component were GVI-M, RST-M, DVI-M, GVI, and RST. After selecting the principal components, MLR was applied to the score data of 80 observations on the two components. The resulting regression model yielded an R^2=0.012 and a p-value (Prob > F) of 0.62, indicating a poor fit and that none of the principal components was significant in explaining the variation in Y data (Figure 2). Note that the p-value for the regression model is based on an overall F-test which tests whether any of the predictors in the regression model improves the fit. By using the procedure described in ‘Material and Methods’ section, the regression coefficients from the above model were multiplied with the appropriate loadings to transform them into their counterparts for the X-Y relationship. The prediction of the Y values based on the X-Y coefficients and comparting them to the original Y values resulted in the same R^2 value (R^2=0.012) as obtained earlier. Despite the low R^2 and high p-values, we investigated the contribution of top predictors to the X-Y relationship and found that the top contributors were RST-M, PIR, RSR-M, DDF, and RLR-M.

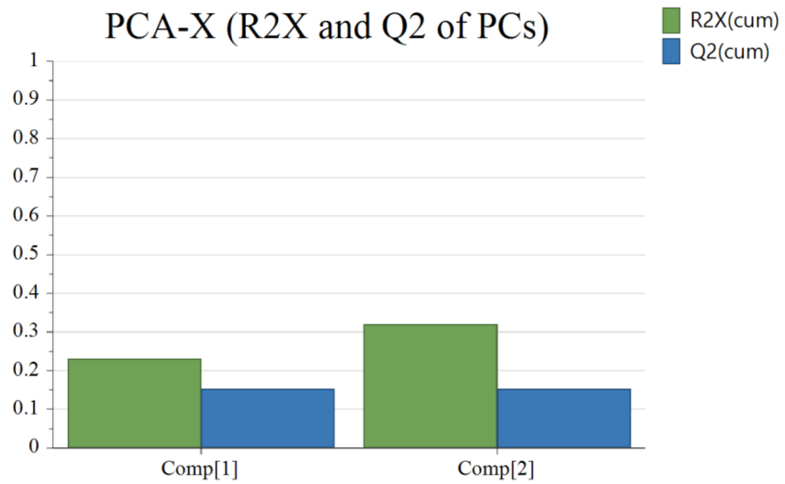


Figure 1(a): R^2X and Q^2 of principal components selected based on cross-validation

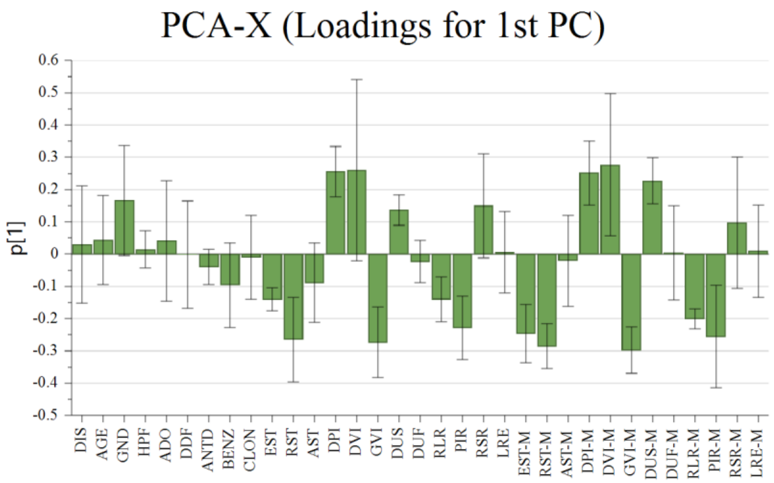
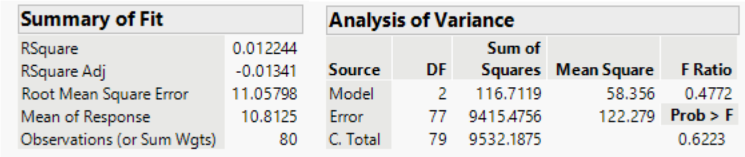


Figure 1(b): Loadings of different predictors for the first principal component

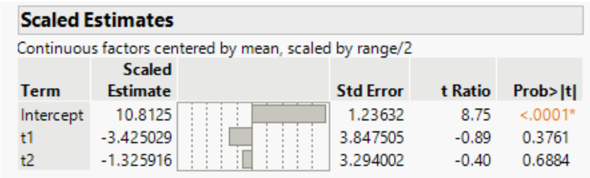


Summary of Fit

RSquare	0.012244
RSquare Adj	-0.01341
Root Mean Square Error	11.05798
Mean of Response	10.8125
Observations (or Sum Wgts)	80

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	116.7119	58.356	0.4772
Error	77	9415.4756	122.279	Prob > F
C. Total	79	9532.1875		0.6223



Scaled Estimates

Continuous factors centered by mean, scaled by range/2

Term	Scaled Estimate		Std Error	t Ratio	Prob>|t|
Intercept	10.8125		1.23632	8.75	<.0001*
t1	-3.425029		3.847505	-0.89	0.3761
t2	-1.325916		3.294002	-0.40	0.6884

Figure 2: Summary of fit, Analysis of variance, and Scaled estimates for PCR model using 2 principal components

PLS

The PLS algorithm yielded 2 principal components based on the cross-validation rules described earlier. The corresponding R^2Y(cum) and Q^2Y(cum) values were 0.61 and 0.29, respectively (Figure 3(a)), with R^2X and R^2Y values of 0.10 and 0.52 for the first principal component, and R^2X and R^2Y values of 0.19 and 0.09 for the second principal component, respectively. A review of VIP plot indicated DIS, HPF, DDF, RLR-M, and AST-M as the top 5 predictors in influencing the model (Figure 3(b)).

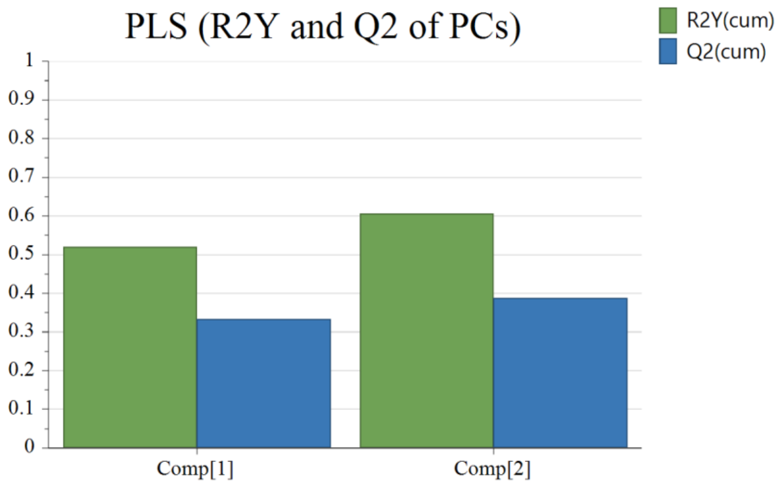


Figure 3(a): R^2Y and Q^2Y values of principal components for PLS model based on cross-validation

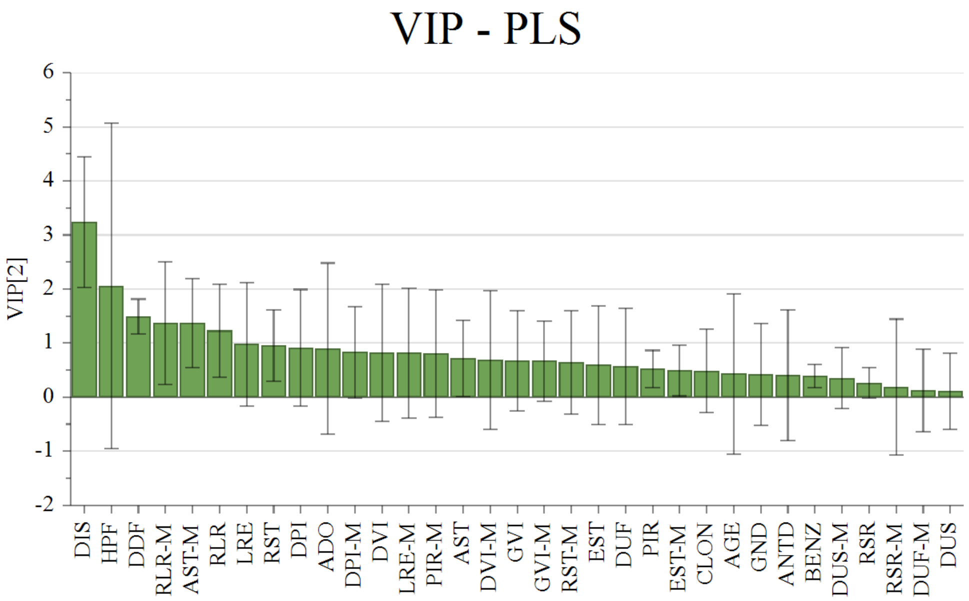


Figure 3(b): VIP plot for PLS model. The VIP values shown here are sorted according to the predictors' importance in influencing the PLS model.

OPLS

The OPLS algorithm resulted in 1 predictive component and 1 orthogonal component based on the cross-validation rules and resulted in R^2Y(cum) and Q^2Y(cum) values of 0.61 and 0.39, with R^2X and R^2Y values of 0.07 and 0.61 for the predictive component, and R^2X and R^2Y values of 0.23 and 0.00 for the orthogonal component, respectively. According to the VIP(total) plot, the top 5 predictors that influenced the model were DIS, DDF, HPF, RLR-M, and RLR (Fig 4(a)). The VIP(total) values shown therein are for the complete OPLS model containing both the predictive and orthogonal components and are sorted according to the predictors' importance in influencing summarization of X data as well as correlation to Y. Figure 4(b) and 4(c) also show the VIP plots for predictive and orthogonal components of the OPLS model, respectively. The top 5 predictors influencing the predictive component were DIS, DDF, HPF, ADO, and AST-M and those influencing the orthogonal component were GVI-M, RST-M, GVI, DVI-M, and RST.

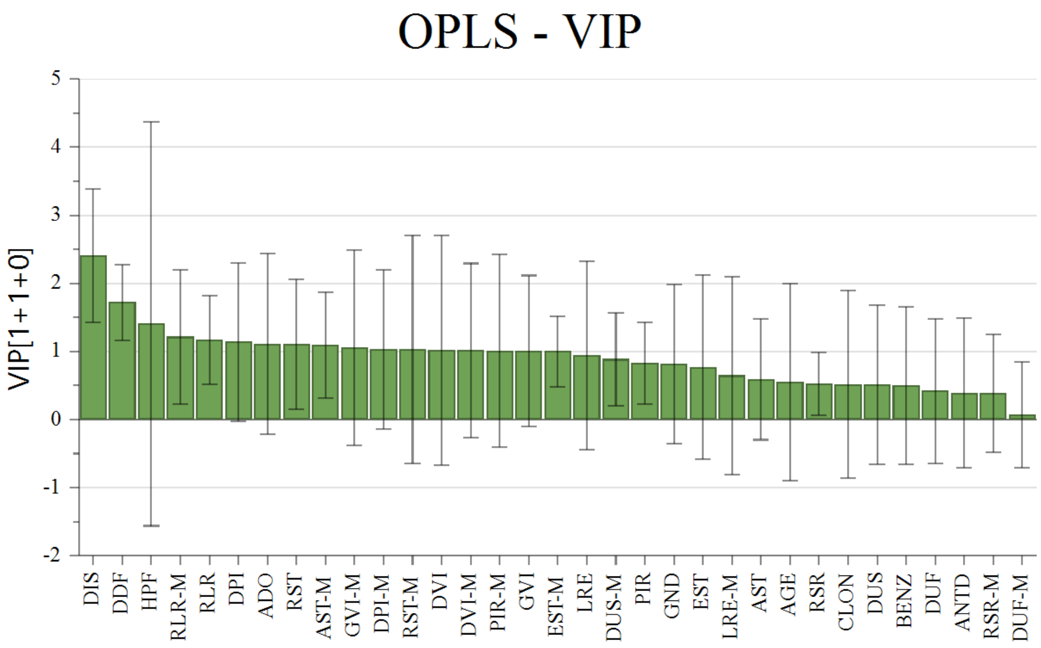


Figure 4(a): VIP(total) model for OPLS model. The VIP values shown here are for the complete OPLS model and are sorted according to the predictors' importance in influencing summarization of X data as well as correlation to Y.

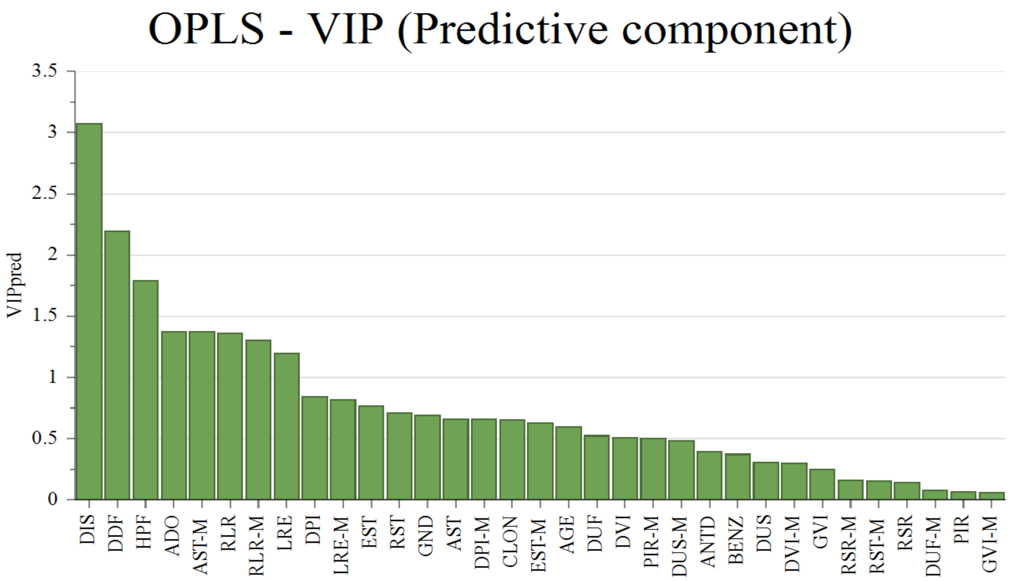


Figure 4(b): VIP (pred) plot for predictive component of OPLS model. The VIP values shown here are for the predictive component of the OPLS model and are sorted according to the predictors' importance in their correlation to Y.

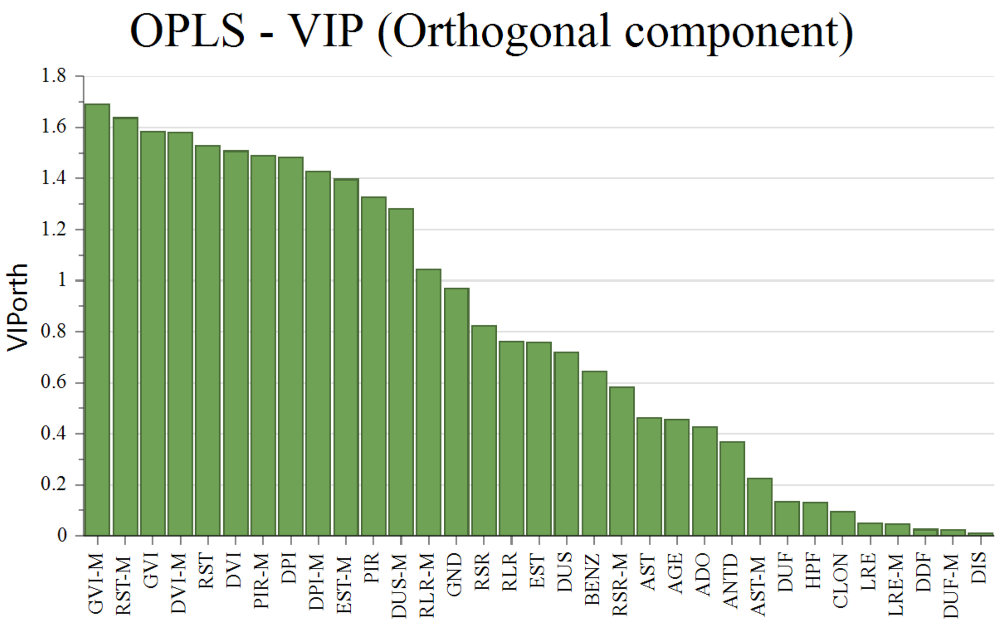
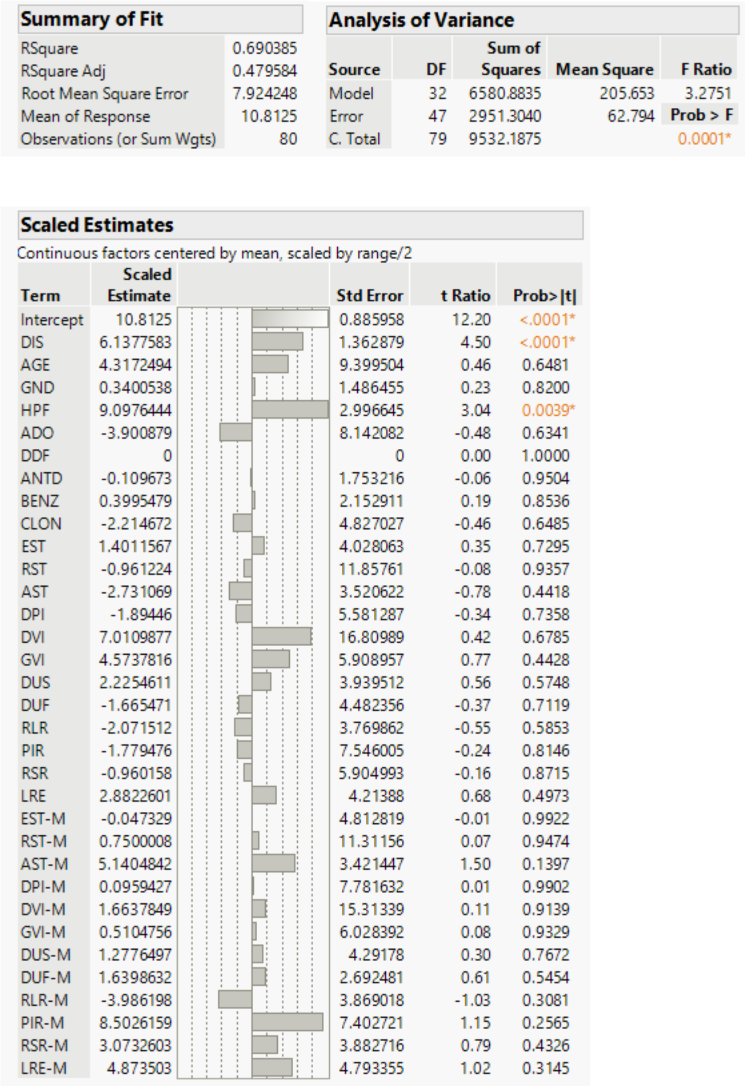


Figure 4(c): VIP (orth) plot for orthogonal component of OPLS model. The VIP values shown here are for the orthogonal component of the OPLS model and are sorted according to the predictors' importance in explaining variation orthogonal to Y.

MLR

Considering that the R^2X(cum) value was only 0.32, which does not reflect a strong correlation between predictors, we also explored MLR to evaluate the predictors related to the response variable. The regression model yielded a moderate fit (R^2=0.69) and a p-value (Prob > F) of 0.0001, indicating that at least one predictor had a significant effect on Y. The scaled estimates of the regression model are shown in Figure 5. DIS, HPF, AST-M, RLR, and PIR-M were identified as the most important predictors in influencing the model (Figure 5). If one selects the predictors strictly on the basis of p-value using a strict criterion (say significance level of 0.05), the analysis shown in Figure 5 would conclude only 2 significant predictors. This points to the limitation of MLR when dealing with correlated data. Since MLR analysis only focuses on establishing a relationship between X and Y without any summarization of the X data, the analysis can result in weighing correlated variables equally and in turn excluding discriminatory predictors that otherwise might be important.



Summary of Fit

RSquare	0.690385
RSquare Adj	0.479584
Root Mean Square Error	7.924248
Mean of Response	10.8125
Observations (or Sum Wgts)	80

Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	32	6580.8835	205.653	3.2751
Error	47	2951.3040	62.794	Prob > F
C. Total	79	9532.1875		0.0001*

Scaled Estimates

Continuous factors centered by mean, scaled by range/2

	Scaled			
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	10.8125	0.885958	12.20	<.0001*
DIS	6.1377583	1.362879	4.50	<.0001*
AGE	4.3172494	9.399504	0.46	0.6481
GND	0.3400538	1.486455	0.23	0.8200
HPF	9.0976444	2.996645	3.04	0.0039*
ADO	-3.900879	8.142082	-0.48	0.6341
DDF	0	0	0.00	1.0000
ANTD	-0.109673	1.753216	-0.06	0.9504
BENZ	0.3995479	2.152911	0.19	0.8536
CLON	-2.214672	4.827027	-0.46	0.6485
EST	1.4011567	4.028063	0.35	0.7295
RST	-0.961224	11.85761	-0.08	0.9357
AST	-2.731069	3.520622	-0.78	0.4418
DPI	-1.89446	5.581287	-0.34	0.7358
DVI	7.0109877	16.80989	0.42	0.6785
GVI	4.5737816	5.908957	0.77	0.4428
DUS	2.2254611	3.939512	0.56	0.5748
DUF	-1.665471	4.482356	-0.37	0.7119
RLR	-2.071512	3.769862	-0.55	0.5853
PIR	-1.779476	7.546005	-0.24	0.8146
RSR	-0.960158	5.904993	-0.16	0.8715
LRE	2.8822601	4.21388	0.68	0.4973
EST-M	-0.047329	4.812819	-0.01	0.9922
RST-M	0.7500008	11.31156	0.07	0.9474
AST-M	5.1404842	3.421447	1.50	0.1397
DPI-M	0.0959427	7.781632	0.01	0.9902
DVI-M	1.6637849	15.31339	0.11	0.9139
GVI-M	0.5104756	6.028392	0.08	0.9329
DUS-M	1.2776497	4.29178	0.30	0.7672
DUF-M	1.6398632	2.692481	0.61	0.5454
RLR-M	-3.986198	3.869018	-1.03	0.3081
PIR-M	8.5026159	7.402721	1.15	0.2565
RSR-M	3.0732603	3.882716	0.79	0.4326
LRE-M	4.873503	4.793355	1.02	0.3145

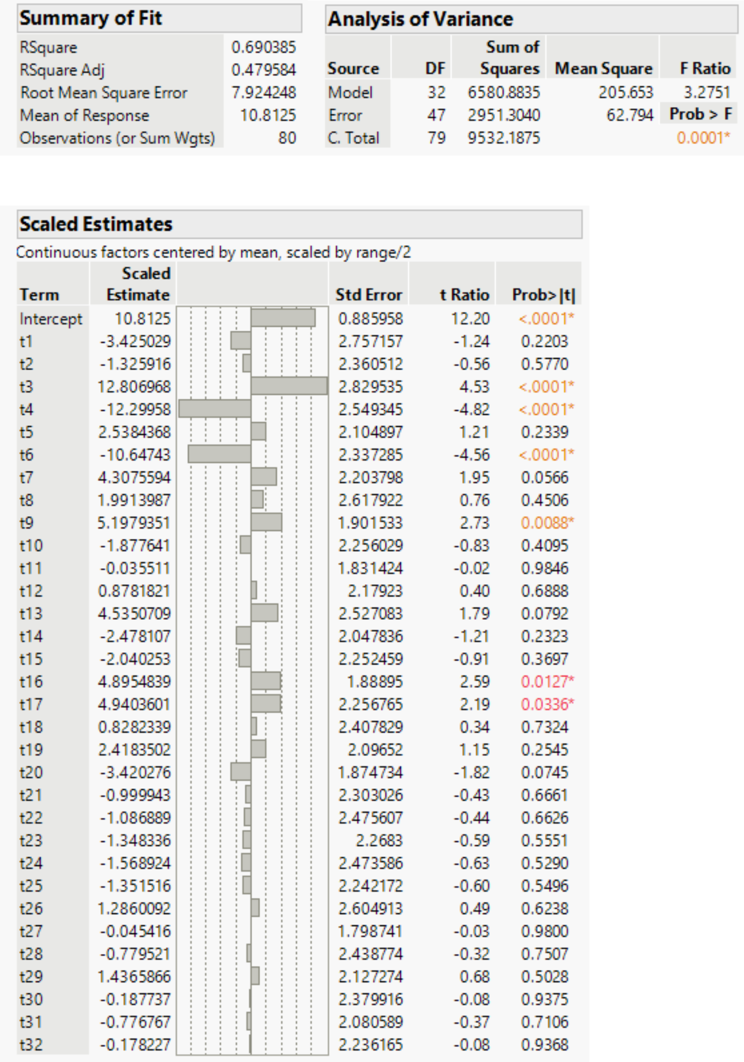
Figure 5: Summary of fit, Analysis of variance, and Scaled estimates for MLR model

Comparative analysis of different approaches using cross validated number of principal components

Comparison between different algorithms showed that the PLS and OPLS algorithms yielded similar results, since four out of the top five predictors between the two models were identified to be the same. MLR shared three of the common top predictors for PLS and OPLS. Two of the three common predictors between the three techniques were also shared by PCR.

Evaluation of increased number of principal components in PCR model

As described earlier, the PCR model with two components did not result in a significant fit of the data, and none of the two principal components demonstrated a significant effect on Y (Figure 2). We checked if the data fit and model significance can be improved by increasing the number of principal components. We attempted to generate an MLR model with all 33 principal components, but this algorithm was not successful, since the scores corresponding to the last (33rd) component were extremely low for all the observations. Note that the first component explained the maximum variation in X and the 33rd component explained the least X variation. Removing the last component resulted in a model that could successfully fit the remaining 32 principal components. The model resulted in a moderate R^2 value of 0.69 and yielded a p-value (Prob > F) of 0.0001, indicating that at least one principal component had a significant effect on Y. Six principal components, namely t3, t4, t9, t16, and t17 were found to have statistically significant effects (Figure 6). Similar to the procedure we adopted earlier, the regression coefficients from this model were multiplied with appropriate loadings to determine the coefficients for the X-Y relationship. Subsequently, the Y values were predicted based on the X-Y coefficients and compared with the original Y values. An R^2 value of 0.69 was obtained based on the predicted and original Y values (same as the R^2 value determined by the regression model with principal scores). Based on the X-Y coefficient values, the top contributors to variation in Y were DIS, HPF, PIR-M, AST-M, and DVI. This new list of discriminatory variables shared 3 common entries with PLS and OPLS results and 4 common entries with MLR results, demonstrating that the addition of principal components to the regression model based on cross-validated number of principal components improved the PCR model’s ability to identify the discriminatory predictors.



Summary of Fit

RSquare	0.690385
RSquare Adj	0.479584
Root Mean Square Error	7.924248
Mean of Response	10.8125
Observations (or Sum Wgts)	80

Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	32	6580.8835	205.653	3.2751
Error	47	2951.3040	62.794	Prob > F
C. Total	79	9532.1875		0.0001*

Scaled Estimates

Continuous factors centered by mean, scaled by range/2

	Scaled			
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	10.8125	0.885958	12.20	<.0001*
t1	-3.425029	2.757157	-1.24	0.2203
t2	-1.325916	2.360512	-0.56	0.5770
t3	12.806968	2.829535	4.53	<.0001*
t4	-12.29958	2.549345	-4.82	<.0001*
t5	2.5384368	2.104897	1.21	0.2339
t6	-10.64743	2.337285	-4.56	<.0001*
t7	4.3075594	2.203798	1.95	0.0566
t8	1.9913987	2.617922	0.76	0.4506
t9	5.1979351	1.901533	2.73	0.0088*
t10	-1.877641	2.256029	-0.83	0.4095
t11	-0.035511	1.831424	-0.02	0.9846
t12	0.8781821	2.17923	0.40	0.6888
t13	4.5350709	2.527083	1.79	0.0792
t14	-2.478107	2.047836	-1.21	0.2323
t15	-2.040253	2.252459	-0.91	0.3697
t16	4.8954839	1.88895	2.59	0.0127*
t17	4.9403601	2.256765	2.19	0.0336*
t18	0.8282339	2.407829	0.34	0.7324
t19	2.4183502	2.09652	1.15	0.2545
t20	-3.420276	1.874734	-1.82	0.0745
t21	-0.999943	2.303026	-0.43	0.6661
t22	-1.086889	2.475607	-0.44	0.6626
t23	-1.348336	2.2683	-0.59	0.5551
t24	-1.568924	2.473586	-0.63	0.5290
t25	-1.351516	2.242172	-0.60	0.5496
t26	1.2860092	2.604913	0.49	0.6238
t27	-0.045416	1.798741	-0.03	0.9800
t28	-0.779521	2.438774	-0.32	0.7507
t29	1.4365866	2.127274	0.68	0.5028
t30	-0.187737	2.379916	-0.08	0.9375
t31	-0.776767	2.080589	-0.37	0.7106
t32	-0.178227	2.236165	-0.08	0.9368

Figure 6: Summary of fit, Analysis of variance, and Scaled estimates for PCR model using 32 principal components

Conclusion

Using a cross-validated number of principal components for PCR, PLS, and OPLS, a comparison of the top 5 discriminatory variables identified by the abovementioned techniques and MLR showed a significant overlap between the results from PLS, OPLS, and MLR with three techniques demonstrating a minimum overlap with the results obtained by PCR. Further exploration of the data revealed that PCR could be used to identify the discriminatory variables too if the number of principal components used in the regression model were increased. In summary, it is recommended to use PLS or OPLS for hypothesis generation and perform a thorough analysis of principal components prior to running the PCR.

Discussion

Using a cross-validated number of principal components, PCR resulted in poor fit of the data and did not yield the same results as PLS and OPLS. This may be explained since the principal components extracted by PCA are aligned with the predictors with the maximum variation and the extracted principal components may not directionally align with the response variable at all. In case of PLS and OPLS, the principal components are extracted with a predefined objective of extracting principal components that would optimize variation in Y and summarize X at the same time. So, in case of PCR, if the directions of the extracted components do not align with the response variable, the resulting model will not be able to establish a relationship between X and Y variables. This phenomenon clearly applies to the current scenario. The top 5 X predictors influencing the direction of the first principal component were GVI-M, RST-M, DVI-M, GVI, and RST (according to the loading plot shown in Figure 1(b)), while those most aligned with the response variables were DIS, HPF, ASTM, RLR, and PIR-M (according to the MLR results in Figure 5). Considering that there was no overlap between the X predictors with the maximum variation and those whose variation most aligned with the response variable, it is not surprising that PCR resulted in poor fit of the data.

Curiously, there was a significant overlap between the results of PLS and OPLS and both models resulted in the same R^2X(cum), R^2Y(cum), and Q^2(Y(cum) values. At the same time, there was a clear difference in how the first principal component was selected for the two models. While the first principal component for PLS accounted for 10% of X variation and 52% of Y variation, the corresponding values in the OPLS model were 7% and 61%, respectively. Also, the second principal component for the PLS model captured 19% of X variation and 9% of Y variation and the orthogonal component for OPLS accounted for 22% of X variation and 0% of Y variation. These numbers are consistent with the mechanics of PLS and OPLS algorithms; while the PLS model tries to summarize X variation and capture X-Y covariance simultaneously in the first principal component, the OPLS model attempts to maximize the X-Y covariance in the predictive component. Curiously, the review of the literature shows that the PLS and OPLS

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0549.v1

models result in the same predictive power as long as the number of components used in the analysis remains the same (Eriksson *et al.*, 2006a). However, the key advantage of OPLS over PLS lies in the interpretability of the results as one can use the predictive component to look into parameters that are correlated to Y and the orthogonal component gives insight into the predictors that have little influence on Y.

Finally, we demonstrated that the number of selected principal components is critical to the accuracy of the results obtained with the PCR. There can be many ways by which the principal components are selected prior to running the PCR regression model. Clearly, as shown in the current work, selecting the first few principal components based on cross validation is not the optimal approach. The fact that the six significant principal components from the 32-components model were not the components that explained the maximum variation in X shows that selecting the principal components in the order of their ability to summarize X is not the right approach to efficiently capture X-Y covariance. One can choose a higher number of principal components as has been done in this work, using an excessive number of principal components can result in overfitting of the data. Alternate approaches for component selection could be selecting the principal components based on subset selection methods such as stepwise regression or shrinkage based methodologies such as Ridge Regression or Lasso (Hastie *et al.*, 2017b). Appropriate selection of principal components to enable PCR has been the subject of prior work (Næs and Martens, 1998; Sutter *et al.*, 1992,) and this should be revisited to develop an efficient algorithm that can successfully identify discriminatory predictors.

Notation

ANOVA	Analysis of variance
MLR	Multiple linear regression
OPLS	Orthogonal Partial Least Squares
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares or Projections to Latent Structures

PRESS Predicted Residual Error Sum of Squares

Q^2 fraction of the total X or Y variation (estimated by cross validation) that can be predicted by the selected principal component

Q^2 fraction of the total X or Y variation (estimated by cross validation) that can be predicted by all the selected principal components

R^2X fractional sum of squares X data explainable by the selected principal component

R^2Y fractional sum of squares Y data explainable by the selected principal component

R^2Y(cum) fractional sum of squares Y data explainable by all selected principal components

VIP Variable influence on projection

doi:10.20944/preprints202111.0549.v1

References

Eastment, H., & Krzanowski, W. (1982). Cross validatory choice of the number of components from a principal component analysis. *Technometrics*, *24*, 73-77.

Eriksson, L., Johansson, E., Kettanah-Wold, N., Wikstrom, C., & Wold, S. (2006a). Orthogonal PLS (OPLS). In: Multi- and Megavariate Data Analysis – Part II. Umeå, Sweden. *Umetrics*, *AB*, 113-136.

Eriksson, L., Johansson, E., Kettanah-Wold, N., Trygg, J., Wikstrom, C., & Wold, S. (2006b). Centering and Scaling. In: Multi- and Megavariate Data Analysis – Part I. Umeå, Sweden. *Umetrics*, *AB*, 207-219.

Goueguel, C. L., PhD. (2019, December 20). *An Overview of Orthogonal Partial Least Squares*. Retrieved October 31, 2021, from https://towardsdatascience.com/an-overview-of-orthogonal-partial-least-squares-dc35da55bd94

Hastie, T., Tibshirani, R., & Friedman, J. (2017a). Subset Selection and Shrinkage Methods. In*:* Elements of Statistical Learning*:* Data Mining, Inference, and Prediction (12th ed.). In *Elem Stat Learn/printings/ESLII_print12* (pp. 57-79). https://web.stanford.edu. https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12_toc.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (2017b). Partial Least Squares. In: Elements of Statistical Learning: Data Mining, Inference, and Prediction. In *Elem Stat Learn/printings/ESLII_print12* (pp. 80-82). https://web.stanford.edu. https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12_toc.pdf.

Hawkins, D. M. (1973). On the investigations of alternative regressions by principal component analysis. *Appl. Statist*, *22*, 275-286.

Hlavnička, J., Čmejla, R., Tykalová, T., Sonka, K., Růžička, E., & Rusz, J. (2017). Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. *Scientific reports*, *7*(1), 12.

Jackson, J. E. (1991). *A User's Guide to Principal Components*. New York: John Wiley & Sons.

Jeffers, J. N. (1967). - Two case studies in the application of principal component analysis. *Appl. Statist*, *16*, 225-236. https://doi.org/10.1038/s41598-017-00047-5

Mansfield, E. R., Webster, J. T., & Gunst, R. F. (1977). An analytic variable selection technique for principal component regression. *Appl. Statist*, *36*, 34-40.

Naes, T., & Martens, H. (1998). Principal Component Regression in NIR Analysis: Viewpoints, Background Details and Selection of Components. *Journal of Chemometrics*, *2*, 155-167. https://doi.org/10.1002/cem.1180020207

Kaggle Inc. (n.d.). *Early Biomarkers of Parkinson's Disease*. Retrieved October 31, 2021, from https://www.kaggle.com/ruslankl/early-biomarkers-of-parkinsons-disease. 2021. Kaggle Inc.

Sutter, J. M., Kalivas, J. H., & Lang, P. M. (1992). Which Principal Components to utilize for Principal Component Regression. *JOURNAL OF CHEMOMETRICS*, *6*, 217-225.

Wold, S., Antti, H., Lindgren, F., & Ohman, J. (1998). Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, *44*(1-2), 175-185. https://doi.org/10.1016/S0169-7439(98)00109-9

Wold, S., Johansson, E., & Cocchi, M. (1993). PLS - Partial Least-Squares Projections to Latent Structures, In: Kubinyi, H. 3D-QSAR in Drug Design, Theory, Methods, and Applications. Ledien. *ESCOM Science*, 523-550.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-Regression: A Basic Tool of Chemometrics. *J Chemometr.*, *58*, 109-130.