REVSTAT – Statistical Journal
Volume 0, Number 0, Month 0000, 000-000
https://doi.org/00.00000/revstat.v00i0.000

The Utilization of Partial Least Squares for Simultaneous Feature Selection and Extraction

Authors: Elisavet Beki
– Laboratory of Statistics and Data Analysis,
University of the Aegean, Greece
elsampeki@hotmail.com

Kimon Ntotsis ✉
– NIHR Leicester Biomedical Research Centre – Respiratory, United Kingdom
Department of Respiratory Sciences, University of Leicester, United Kingdom
kn170@leicester.ac.uk

Received: Month 0000 Revised: Month 0000 Accepted: Month 0000

Abstract:

- Several works concerning the utilization of Partial Least Squares as a supervised dimension reduction technique have been developed over the years in the field of chemometrics, among others, for regression purposes. However, Partial Least Squares can be a challenging procedure especially in the case of multivariate multiple regression due to data characteristics and complexity. Thus, in this work we propose the use of Partial Least Squares method as a variable selection technique in linear regression tasks that involve high dimensional spectral data sets. More precisely, we suggest the exploitation of the regression coefficients that Partial Least Squares estimates in order to identify and eject the insignificant predictor variables from the analysis. In such manner we are able to remove the uninformative variables and obtain in most cases better results than the classical Partial Least Squares regression but with simpler structure. We compare our proposed technique with the classical Partial Least Squares and Principal Component Analysis in both univariate and multivariate regression.

Keywords:

- *partial least squares; beta constrain; dimension reduction; variable selection.*

AMS Subject Classification:

- 62H25, 62J05, 62H12, 62H20.

✉ Corresponding author

1. INTRODUCTION

Recent advances in computer science and engineering enable scientists to collect and store datasets, typically characterized by high dimensionality. Such data sets, also referred to as Big Data, consist of numerous measured variables, the successful process of which aim to provide production of accurate predictions, informative visualizations and foremost, the in-depth understanding of the underlying patterns and relations between the variables of a dataset. However, the complexity of Big Data raises new challenges in their effective handling. The mitigation of noise accumulation and multicollinearity are two of the most common obstacles that the researcher is called upon to address and resolve. By noise, it is meant the part of data that brings insignificant information, while the term multicollinearity is used to describe the existence of high intercorrelations between variables. Further, in many cases, scientists have to deal with the existence of a small number of available observations compared with the larger number of measured variables, a situation referred as "small n, large p problem" (Ntotsis et al. [15]).

Therefore, Dimensionality Reduction Techniques, hereafter DRT, have been developed and utilized to overcome these potential situations that can severely effect the modelling process when analyzing Big Data. DRT are defined as techniques that process high-dimensional datasets to produce representations of lower dimension, that retain as much as possible information from the original ones. Dimensionality reduction is used in statistical tasks that serve various purposes, such as classification, linear regression, etc, and it can be achieved either by feature selection or by feature extraction.

Any technique that selects and subsets a number of the initial variables -that are considered statistically significant, and formulates a new dataset containing only these ones can be characterized as feature selection. By this process, variables' interpretation is preserved. Nevertheless, despite the advantage of interpretability, information captured in interactions and correlations between retained and removed variables is lost (Li and Zeng [13]). Some of the most frequently used feature selection techniques, based either on wrapper methods (e.g. Forward Feature Selection, Backward Feature Elimination), or either on regularization methods (e.g. Least Absolute Shrinkage and Selection Operator). On the contrary, in feature extraction the produced data set is consisted of transformed variables and each variable is considered as a combination of all the initial ones, that project data points to a low dimensional space. These transformations are called components and they summarize information of initial variables by using them to form linear combinations. Although, in this case the interpretability of the initial variables is lost, feature extraction methods are very popular because they effectively replace initial variables by a few components that compress the relevant information contained in the data and make visualizations feasible. Among the most popular feature extraction techniques are Partial Least Squares Method (PLS), Principal Component Analysis (PCA), and Linear Discriminant

Analysis (LDA) and Canonical Correlation Analysis (CCA).

Despite their different approach, both types of dimensionality reduction techniques can successfully mitigate noise accumulation and multicollinearity. Further, analysis of data of reduced complexity leads to models with improved accuracy/precision generated with less computational power and time, due to the parsimonious descriptions of the available observations. As a result, their implementation in Big Data is substantial and the increased scientific interest about them has resulted in the proposal of various techniques.

In this article we attempt to evaluate the proficiency of a multilevel dimension reduction scheme: In the first place we estimate the regression coefficients of original variables utilizing PLS method. Combining testing thresholds and model assessment criteria, we are able to identify and remove uninformative variables for the modelling of response variable. Further, the dimension reduction is completed with the integration of Partial Least Squares Regression on the modified reduced data set. The efficiency of the proposed algorithm is investigated through the implementation of Principal Component Regression (PCR) on the reduced data set, independently from PLS, to compare the performance of the resulted models, in terms of AIC, Adjusted R^2 value and RMSECV information criteria.

In the past, many researchers have modified PLS and proposed refinements of the classical algorithm in order to achive optimized results. Moving window partial least squares regression (MWPLS; Jiang et al. [9]), group partial least squares (gPLS; Liquet et al. [14]), sparse partial least squares (sPLS; Chung and Keles [3] and Lé Cao et al. [11]) and sparse group partial least squares (sgPLS; Liquet et al. [14]) are among these attempts. In our proposal, the main idea is based on the fact that in linear regression those variables which are strongly related with the outcome and valuable for its prediction are associated with a large PLS regression (PLSR) coefficient (Wold et al. [18]).

In Section 2 the basics of PCA is presented while Section 3 discuss the proposed feature selection PLS (FS-PLS) method. Some model assessment criteria are provided in Section 4. The application of FS-PLS and its comparison with PCA and PLS for the univariate and multivariate case is provided in Section 5. The paper ends with the concluding remarks and possible future expansions of this work.

2. PRINCIPAL COMPONENT ANALYSIS

Principal Components Analysis (PCA) is one of the most widespread dimensionality reduction techniques. It is a multivariate feature extraction method through which it is achieved the transformation of a data matrix X into a low dimension matrix. The newly generated matrix contains the principal components, i.e. the transformed variables of X matrix, that have been generated as linear combinations of the original variables. These components have the following desirable properties: they are uncorrelated and their number is significantly reduced compared to the corresponding number of the initial matrix. Thus, due to the construction algorithm, they summary the majority of the initial information discharged by noise.

Geometrically, through PCA data points are projected onto a low dimension space, the coordinate system of which is oriented in the directions of maximized variance of data points, the directions of Principal Components. The vectors containing the weights of the original variables in the linear combinations that define these directions are called loadings, while the coordinates of the available observations in the new space are called scores.

Making use of these quantities, an X matrix of size $n \times p$ can be written as:

$$X = T_m P_m^{\intercal} + E$$

where T, P, E are correspondingly the matrices of scores, loadings and residual errors. The latter expresses the information that is lost through this analysis and it is the cost of the dimensionality reduction process. The index m indicates the dimension of the new space -it is the number of selected Principal Components and it is defined by the analyst.

Finally, the Principal Component Analysis of the original matrix is:

$$X = T_m P_m^{\intercal}.$$

This representation is an approximation of the initial matrix and can be used for modelling.

2.1. Construction algorithm

The main elements of PCA are scores and loadings matrices and they can be computed based on the Eigen-decomposition of either covariance matrix or correlation matrix of the initial variables (Ntotsis and Karagrigoriou [16]).

Covariance matrix is used when all variables in X express the same measurement unit. In opposite case, correlation matrix is used, since correlation is independent from the scale of the variables. Regardless of the selected matrix, data standardization is highly recommended in the presence of extreme multicollinearity. It is achieved by replacing each x_{ij} element in X matrix by:

$$\frac{x_{ij} - \bar{x}_j}{s_j}$$

where s_j indicates the standard deviation of the j^{th} variable.

The selection and computation of the selected matrix is followed by the computation of its eigenvectors and its eigenvalues (λ_i). The latter are ordered and based on them, the determination of the number of the retained PCs (m) in the model is possible according to:

- Cumulative Percentage of total variation: The inclusion of PCs is interrupted when the first m of them achieve to absorb $80\% - 90\%$ of total variation of X. The cumulative percentage of the PCs variation is computed by:
$$\sum_{j=1}^{m} q_j = 100 \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$
- Kaiser's rule: According to this rule the dimension of the summarized matrix is equal to the number of eigenvalues that are larger than one.
- The scree graph: This graph illustrate the rank of eigenvalues against their value. Usually, the curve that connects the points forms an elbow-like shape, with the rank of the point located on its angle indicate the dimension of the new matrix.

The formation of the $p \times m$ loadings matrix P follows, which consists of the eigenvectors that correspond to the greatest m eigenvalues, set as columns. This matrix is useful for the computation of the matrix T, the matrix of Principal Components through the formula:

$$T = XP$$

where they constitute its columns, and they can be used in linear regression tasks according to the following schema (Wehrens [17]):

$$Y = XB + E = (TP^{\intercal})B + E' = T(P^{\intercal}B) + E' = TA + E'$$

where A matrix indicates the regression coefficients of components and they arise with the use of Ordinary Least Squares Method:

$$A = (T^{\intercal}T)^{-1}T^{\intercal}Y.$$

These coefficients can be transformed to coefficients of the original variables:

$$B = PA = P(T^{\intercal}T)^{-1}T^{\intercal}Y.$$

3. FEATURE SELECTION PARTIAL LEAST SQUARES (FS-PLS)

The FS-PLS procedure can be considered as an expansion of the original Partial Least Squares Method (PLS). Thus, its definition arises through PLS.

The PLS is a dimension reduction technique that achieves feature extraction. It is quite popular in chemistry or chemometrics due to high correlations frequently encountered among variables in those fields. It shares a common approach with PCA: via its algorithm, few new uncorrelated variables that summarize the information in X matrix are produced. They are called latent variables and they arise as linear combinations of the initial ones. The difference in this approach is that these variables aim to summarize information that is directly related to the response variables in a regression problem. To accomplish this, both X and Y matrices are being analyzed through an iterative procedure that generates latent variables (LV) oriented in the direction that maximizes the covariance between the involved matrices. On the contrary, PCA focuses on the explanation of the variability in X matrix independently from Y, an approach that can result in loss of information, that could be proved valuable for prediction purposes. Furthermore, another difference between the two methods is that PLS disposes of a variation, that makes feasible the simultaneous modelling of multiple response variables, where the intercorrelations among them are considered. Instead, PCA runs regression steps multiple times to model multiple responses. PLS is often considered more appropriate than PCA in cases with small sample sizes, multicollinearity and missing data.

From a geometrical point of view, similarly to PCA, X but here also Y data are projected to low dimensional spaces defined by the latent variables. Making use of the scores and loadings of the observations in the new spaces, an X data matrix of size $n \times p$ and a Y data matrix of size $n \times k$ can be written as:

$$X = T_m P_m^\intercal + E$$

$$Y = U_m Q_m^\intercal + F$$

where T and U are score matrices, P and Q are loading matrices and E and F are matrices representing the information loss. The index m expresses the dimension of the new space and it is determined by the analyst. Eventually, the summarized approximation of the initial X matrix through PLS is:

$$X = T_m P_m^\intercal$$

and it can be used in linear regression.

3.1. Construction algorithm

The construction of a PLS model can be completed by a number of relative algorithms. In this article we use the Nonlinear Iterative Partial Least Squares algorithm (NIPALS), which was proposed by Herman Wold [18].

According to the NIPALS algorithm [5], a PLS model can be produced by a process in every iteration of which a latent variable emerges. More specifically, initially an X score vector (t) and a Y score vector (u) with maximized covariance are generated. Their directions are determined by weight vectors, w and c that correspond to X and Y matrices, respectively:

$$t = Xw$$

$$u = Yc/(c^{\intercal}c).$$

Next the computation of a loading vector p follows:

$$p = X^{\intercal}t/(t^{\intercal}t)$$

that is used in a deflation process of the X matrix and through which the information explained by the produced latent variable is subtracted:

$$X_{new} = X_{old} - tp^{\intercal}.$$

This deflated matrix is then used in the next iteration of the algorithm for obtaining a new latent variable.

The procedure of emerging new variables is terminated based on the indications of model selection criteria. The computation of their values follows the implementation of the regression step, demonstrating that in PLS dimension reduction and regression run simultaneously. The scheme in Partial Least Squares Regression (PLSR) is identical to Principal Component Regression (PCR):

$$Y = XB + E = (TP^{\intercal})B + E' = T(P^{\intercal}B) + E' = TA + E',$$

where

$$A = (T^{\intercal}T)^{-1}T^{\intercal}Y.$$

Here, T and P matrices are formed by the apposition of the output scores vectors (t) and loading vectors (p) as columns, respectively. The i^{th} column in A matrix contains the regression coefficients for the i^{th} response variable in Y matrix. These coefficients refer to the extracted latent variables, although the reference to the original variables is attainable by the formula:

$$B = RA = R(T^{\intercal}T)^{-1}T^{\intercal}Y.$$

The vectors in R matrix represent the weights of every original variable of X matrix at the extracted latent variables, unlike the weight vectors w_i that refer to the deflated matrices X_i and their apposition forms W matrix, which is related to R:

$$R = W(P^{\intercal}W)^{-1}.$$

As stated before, PLS is considered to be an effective dimension reduction technique when it comes to obtaining an optimal statistical model. However, like many similar feature extraction techniques, we end up with a model that involves all original variables, significant, or not. What if there was a way to take advantage of PLS algorithm in order to utilize it as a variable selection technique? The FS-PLS is a novel approach that allows the researcher to use the

PLS procedure to remove non-significant variables from the original dataset and obtain a statistically significant model with minimum dimension when PLS is applied. FS-PLS provides a new dataset with simpler structure than the original one and still when its implementation is compared to PLS and PCA, the model arises from FS-PLS is more efficient than the corresponding models of PLS and PCA. This "superiority" is due to the fact that the constructed model of FS-PLS is easier to interpret since all irrelevant variables have been removed.

The beta coefficients (β) that emerge from the PLSR in conjunction with the number of selected latent variables can be seen as a general rule of thumb for disregarding variables from a dataset. We propose the following rule to determine if a variable is significant:

Let us assume a model with X_j, $j = 1, \ldots, m$ independent variables and let v be the number of latent variables that have been selected as optimal from the PLS regression of the aforementioned model. Let us also assume that β_j^v being the corresponding coefficient of X_j variable in the v-latent variable (each latent contains all original variables). Now, let us define Equation 3.1 as follows:

$$|\beta_j^v| \leq c, \tag{3.1}$$

where $c \in [0.05, |max\{\beta_j^v\}/2|)$ is a pre-determined non-negative value close to zero and $|max\{\beta_j^v\}|$ is the maximum (absolute) value that exists in the coefficient matrix of the selected v latent variables. If Equation 3.1 is satisfied for the j-th variable, i.e. $|\beta_j^1| \leq c$, and $|\beta_j^2| \leq c$, $\ldots$ and $|\beta_j^v| \leq c$, then this variable can be labelled as non-significant. By integrating this β-based constraint in the PLS regression, it is feasible to discard the insignificant variables and still maintain a robust model. A fixed value c is expected to complement effectively all other aspects (purpose of the study, researcher's judgement, etc.) of the decision-making process. In that sense, it can be considered as a rule of thumb and is in the judgement of the researcher which value of c is the one that results the optimal PLS model without underfitting or overfitting the model under consideration. We recommend a step procedure of 0.05 units (i.e. 0.05, 0.10, 0.15, etc.) until model underfitting is observed based on the model selection criteria.

The FS-PLS algorithm consists of a two level implementation of the PLSR algorithm. Initially, PLS method is applied on the original dataset and the regression coefficients of the original variables are estimated with the use of models consisted of up to three latent variables. Those variables with absolute values of regression coefficients lower than the testing threshold in all three models are considered insignificant for the prediction of response variable and they are removed from the dataset. This distinction between the variables is followed by the application of Partial Least Squares Regression to generate predictive models. Their competency is evaluated based on information criteria, such as AIC, Adjusted R^2 (R^2_{adj}), RMSECV and Adjusted Wold's R criterion, that lead to the final model selection.

The following algorithm displays the proposed procedure

Algorithm 1 Pseudocode for FS-PLS

```
Input: A data set consisted by a $n \times p$ matrix X and a $n \times 1$ matrix Y, where each $X_j$ and $Y$ column represents a variable, and a constant threshold $c$.
Output: A data set consisted of the minimum variables that can result in the optimal PLS model.

    Step 1: Application of PLSR on original data for the evaluation of regression coefficients.

    Step 2: Usage of model selection criteria for number of optimal latent variables determination

    Step 3: Application of the constrain proposed in Equation 3.1 for the location of the statistically insignificant variables.

    Step 4: Removal from the input dataset the variables that Step 3 indicate as insignificant

    Step 5: Repetition of Step 1 on the minimized original data
```

4. MODEL ASSESSMENT

In this section we briefly discuss classical model assessment criteria for PLS and/or PCA.

The selection of the optimal number of latent variables to retain in a PLS model is determined on the basis of the following criteria:

- Wold's R criterion: It is a criterion specially designed to evaluate PLSR models by comparing the contribution of a new extracted variable with the previous one, to the predictive ability of the model. For this purpose, cross validation technique is involved to compute Predicted Error Sum of Squares (PRESS) statistic and R ratio as follows (Li et al. [12]):

  $$R = \frac{PRESS(m+1)}{PRESS(m)}$$

  where m denotes the number of retained latent variables in the model. The inclusion of the latent variable that makes R greater than one, terminates the construction algorithm and the produce of new latent variables. The first m of them are then included in the model.

- Adjusted Wold's R criterion: In this permutation of Wold's R criterion the ratio R is compared to the values 0.90 ($R_{adj}^{0.90}$) and 0.95 ($R_{adj}^{0.95}$) rather than

1, as in the original version. As it has been proven in Li et al. [12], these variations give better results due to sample variability.

In many cases, when researchers deal with high dimensional datasets, variable selection leads up to the construction of a PLS and/or a PCA model, in order to remove insignificant variables in a preparatory level. As a result, the production of sets of models that differ in the number of predictors they arise from and also differ in terms of complexity, occurs. The selection of the optimal model can emerge from various model selection criteria (Faraway [6]). The most frequently utilized criteria that one can use when in PCA, PLS and similar techniques have been documented below:

- R-squared (R^2) value: It expresses the percentage of the explained variability in the response variable and it is computed by:

  $$R^2 = \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{\sum_i^n (y_i - \bar{y})^2}$$

  It varies from 0 to 1 and higher values indicate more sufficient model performance.

- Adjusted R^2 value: It is a modification of R^2 criterion that penalizes models of higher complexity. It is computed by:

  $$R^2_{adj} = 1 - \frac{(n-1)}{(n-p-1)}(1 - R^2)$$

  where n refers to the number of available observations and p to the number of retained components.

- Akaike's Information Criterion: This measure consider both the predictive accuracy and the parsimony of a model to evaluate it, according to the following formula:

  $$AIC_m = -2\textit{(maximum log-likelihood)} + 2m$$

  where m is the number of the retained predictors in the model. Among the comparable models, the most sufficient performance is presented by the model with the lowest AIC value.

- RMSECV: This measure involves cross validation to give an estimation of the variation/divergence of the predicted values from the true values of unseen observations, in lack of available data that could be used as a test set. The criterion uses the cross-validation approach and its value is computed as:

  $$RMSECV = \sqrt{\frac{\sum_j \frac{\sum_i (y_{ij} - \hat{y}_{ij})^2}{N_j}}{k}}$$

  where $\hat{y}_{ij}$ is the estimation of y_{ij}, N_j is the number of observations in the j^{th} fold and k is the number of folds in cross validation procedure.

Lower RMSECV values indicate better predictive capacity of the compared models.

By utilizing R^2_{adj}, we aimed to assess the models' ability to explain the variability in the dependent variable while considering the complexity of the regression models. This metric allows for a fair comparison between the PLS and PCA regression models, taking into account their respective number of latent variables or principal components and potential differences in the predictive performance. Moreover, adjusted R-squared aligns with the principle of parsimony, which encourages the use of simpler models that explain the data effectively. The inclusion of adjusted R-squared as one of our evaluation criteria supports the selection of models that strike a balance between explanatory power and model complexity.

We acknowledge that R^2_{adj} should not be the sole criterion for model evaluation, and we have also incorporated other well-established metrics such as AIC, Wold's R criterion, and RMSECV. These metrics provide complementary insights into model performance, including goodness of fit, model complexity, and prediction accuracy.

By utilizing a combination of these metrics, we aimed to provide a comprehensive evaluation of the PLS and PCA regression models, taking into consideration various aspects of model performance. This approach ensures a robust and thorough assessment of the models and allows for informed comparisons between them.

To address the issue of information control in our proposed methodology, we evaluate the same chemometrics datasets that have been extensively used in previous studies on PLS regression. These studies, have already addressed the concern of information preservation within PLS models. For instance, in first row of Table 1 (case where c is $-$), we present a typical PLS model used as a baseline in previous works. Our goal is to simplify this model by removing variables while maintaining its information content and robustness. To ensure information control, we employ the Wold's R criterion, specifically developed for evaluating PLS models. Additionally, we utilize other established metrics such as AIC, adjusted R-squared, and RMSECV to comprehensively assess the performance of our methodology. By incorporating these evaluation metrics and techniques, we ensure that the simplification process retains the essential information captured by the original PLS model. This allows us to strike a balance between model complexity and interpretability while preserving the predictive performance of the PLS regression. It is important to note that the information control aspect has already been addressed in the relevant literature on PLS regression, which serves as the foundation for our work.

5. NUMERICAL APPLICATIONS

In this section the application of the FS-PLS on near infrared (NIR) spectroscopy data is presented.

5.1. Univariate FS-PLS regression – FS-PLSR

In the first case, in the *gasoline* dataset, which is found in the `pls` package, X matrix includes 401 diffuse reflectance measurements and Y matrix is consisted of one response variable, that corresponds to the number of octanes of the total 60 observations (Kalivas [10]). Due to the multicollinearity and the rate of available observations to X-variables, dimensionality reduction is demanded in order to generate a linear regression model. Applying the FS-PLS optimization, we first computed the estimators of PLS-regression coefficients of all 401 variables in models built with up to three components. Their absolute values were then compared with predefined constant c of 0.10, 0.20, 0.25 and 0.30. The final X data matrices contextually included only the predictive variables with absolute values of PLS-coefficients higher than the testing threshold in one-, two- and three-component models (1LV, 2LV, and 3LV). In the next step we reapplied the PLSR method to the selected variables and the resulted models were evaluated based on AIC, R^2_{adj}, $R^{0.90}_{adj}$, and $R^{0.95}_{adj}$ and RMSECV. Table 1 and Table 2 summarize the results:

					1 LV		2 LV		3 LV		4 LV		5 LV		6 LV	
c	Attributes	$R^{0.90}_{adj}$	$R^{0.95}_{adj}$	RMSECV	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}
-	401	4	4	7	203	30%	52	94%	3	97%	-2	97%	-25	98%	-36	98%
0.10	374	4	4	7	203	30%	52	94%	4	97%	-3	97%	-25	98%	-36	98%
0.20	307	4	4	7	203	31%	52	94%	6	97%	-3	97%	-24	98%	-35	98%
0.25	245	4	6	6	202	31%	51	94%	9	97%	-4	98%	-23	98%	-35	98%
0.30	217	4	4	6	202	31%	51	94%	12	97%	-5	98%	-22	98%	-35	98%

Table 1: Information criteria values of FS-PLSR models, where Attributes is the number of original variables.

In Table 1, the reduction in AIC values in all two-component models and the simultaneous increase of their R^2_{adj} values is noteworthy. These changes strongly indicate the outstanding enhancement of the corresponding models when the second component is retained in the model. Further, the most sufficient FS-PLSR model is proposed, the four-component model, which is based on the 0.30 testing threshold and it includes only 217 variables in X matrix, which consist of 46% of the initial observations. This choice is established in accordance with the adjusted Wold criterion, which is specialized to evaluate PLS models, complemented by the high R^2_{adj} value and the significant reduction in AIC value. It should be noted that AIC values tend to decrease as more components are added to the model.

However, the rate of decrease is approximately fixed after the addition of the fourth component. Moreover, the criteria values of the models that resulted from the thresholds 0.25 and 0.30 are alike, though the latter constraint conveys to further dimensional reduction. At this point, it should be mentioned that more restrictive thresholds were tested; they were found to lead to over-fitted models and rejected.

The results of the PCA regression (PCR) models, generated with the datasets arising from the aforementioned thresholds, are displayed in Table 2. As the most adequate model is proposed the five-component model of the last threshold, since R^2_{adj} value is close to 1 and AIC value does not change sufficiently with the addition of more components in the model. In contradiction to the FS-PLSR models, the inclusion of the second component does not improve the model performance in any case, while the minimization of RMSECV values proposes much more complicated models than in FS-PLSR cases.

		1 LV		2 LV		3 LV		4 LV		5 LV		6 LV	
c	RMSECV	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}
-	17	213	17%	215	17%	192	43%	6	97%	6	97%	8	97%
0.10	17	213	17%	215	17%	189	46%	6	97%	6	97%	7	97%
0.20	15	213	17%	214	17%	174	58%	6	97%	5	97%	7	97%
0.25	15	213	17%	214	18%	147	73%	7	97%	5	97%	5	97%
0.30	14	213	17%	214	18%	133	79%	7	97%	5	97%	5	97%

Table 2: Information criteria values of PCR models

Artigue and Smith [2] have examined several challenges related to PCR. However, our study specifically focuses on PLSR, which addresses a key limitation of PCR by considering both the independent and dependent variables, thereby offering a solution to the mentioned issue which is one of the main arguments of the article.

Additionally, taking into account the R^2_{adj} criterion and the percentages of explained variability in the models, as displayed in Figure 1, we conclude that in FS-PLSR the two-component and three-component models can lead to reliable results, preserving the advantage of visualization. Note that all c constrains resulted in similar explained variability and thus only one's results are being presented in Figure 1. These FS-PLSR models expose high R^2_{adj} values, while they leave unexplained a negligible percentage of the response variable. In PCR instead, the inclusion of the first four components fails to provide a model with sufficient performance. Finally, the comparison of these methods in terms of AIC values verifies the predominance of FS-PLSR against PCR: all AIC values in PCR models (with up to three components) are significantly smaller than the corresponding FS-PLSR model (Table 2).

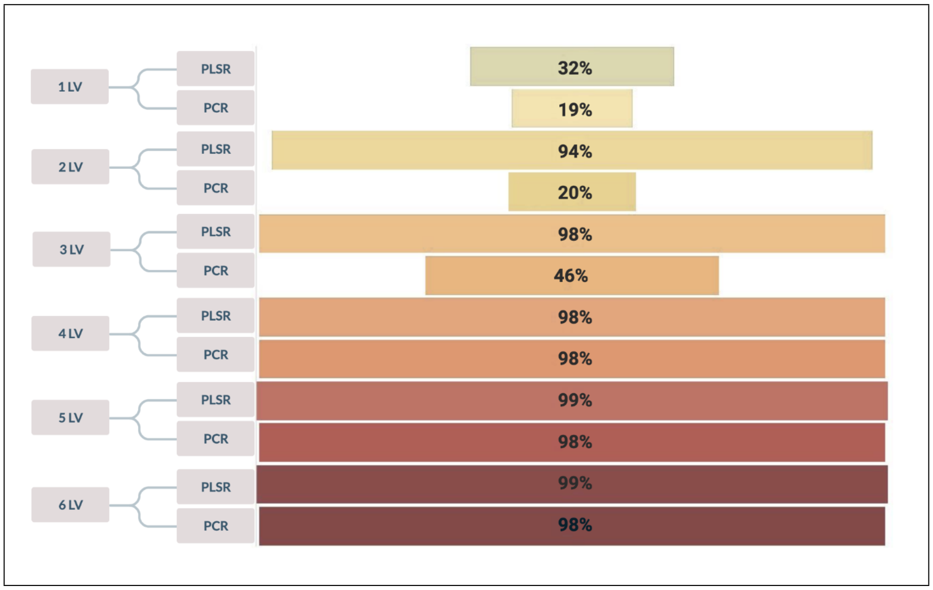

Figure 1: Percentage of explained variability of FS-PLSR and PCR models

5.2. Multivariate FS-PLS regression – FS-MPLSR

In the second case, the performance of the proposed FS-MPLSR optimization over a multivariate response is investigated (the multivariate case of FS-PLSR will be addressed by FS-MPLSR in the remaining article). In the *corn data set* [4], that we processed, the Y matrix consists of four variables, -moisture, oil, protein and starch, and X matrix includes 700 NIR spectroscopic attributes. Note that the FS-MPLSR algorithmic procedure is similar to the FS-PLSR with the only deference to be the number of response variables that form the latent variables. In the multivariate case, the modelling process aims to reveal and enable chemists to predict the moisture, oil, protein and starch content in different samples. In this situation, the implementation of Ordinary Least Squares as a linear regression method would be an inappropriate choice, since the X matrix is characterized by the existence of multicollinearity. Its mitigation is achieved through dimensional reduction, based on the absolute values of the FS-PLS-regression coefficients, in a similar way as in the univariate case. The FS-MPLSR algorithm was applied on the initial dataset to estimate these values. The computation of Adjusted Wold's R criterion $R_{adj}^{0.90}$ led to the conclusion that the sufficient modelling of the four Y-variables requires the inclusion of first 5, 21, 7 and 8 FS-PLS-components respectively. Based on this conclusion and the use of testing thresholds we defined the final reduced set of predictors as the intersection of the following four subsets:

- The first subset included the variables considered as statistically significant for Y1. The absolute values of regression coefficients of these variables are higher than the tested thresholds in one- to five-component models.

- The second subset included the variables considered as statistically significant for Y2. The absolute values of regression coefficients of these variables are higher than the tested thresholds in one- to twenty one-component models.
- The third subset included the variables considered as statistically significant for Y3. The absolute values of regression coefficients of these variables are higher than the tested thresholds in one- to seven-component models.
- The fourth subset included the variables considered as statistically significant for Y4. The absolute values of regression coefficients of these variables are higher than the tested thresholds in one- to eight-component models.

The thresholds that we tested were 2, 2.25, 2.50 and they resulted in the removal of 44, 69, and 99 variables from the original dataset, correspondingly. The new reduced data matrices were then processed via the FS-MPLSR and PCR methods. Based on the values of the aforementioned model selection criteria we inferred that the third threshold examined (2.50) generated the most efficient models. Table 3 and Table 4 summarize the values. The other options led to over-fitted or under-fitted models.

	1 LV		2 LV		5 LV		7 LV		8 LV		omitted LV	21 LV	
	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	$\cdots$	AIC	R^2_{adj}
Y1	1	51%	2	52%	-50	80%	-142	96%	-166	97%	$\cdots$	-366	99%
Y2	-62	27%	-62	27%	-68	37%	-87	55%	-154	86%	$\cdots$	-214	95%
Y3	76	17%	65	32%	-35	88%	-56	92%	-60	92%	$\cdots$	-181	99%
Y4	153	0%	145	13%	63	78%	28	88%	7	92%	$\cdots$	-131	99%

Table 3: Information criteria values of the FS-MPLSR model based on the remaining 601 attributes (700-99).

	1 LV		2 LV		5 LV		7 LV		8 LV		omitted LV	21 LV	
	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	$\cdots$	AIC	R^2_{adj}
Y1	1	52%	1	53%	-115	93%	-153	96%	-151	96%	$\cdots$	-371	99%
Y2	-62	27%	-61	26%	-69	38%	-116	72%	-124	76%	$\cdots$	-158	88%
Y3	76	17%	74	22%	22	68%	-24	86%	-56	92%	$\cdots$	-112	97%
Y4	153	0%	151	0%	120	44%	76	74%	52	83%	$\cdots$	-47	97%

Table 4: Information criteria values of the PCR model based on the remaining 601 attributes.

The optimum FS-MPLSR model retained twenty one components. The model complexity was determined in accordance with the theory (Wold et al. [18], which states that the FS-MPLSR model should include every component that is found to be significant for at least one variable of the set of responses. In this way information in Y matrix is significantly explained, as the percentages of the explained variability are 99.95% for Y1, 97.06% for Y2, 99.43% for Y3 and 99.55% for Y4, while the overall information of the new X matrix is utilized. We can infer that the substantial dimensionality reduction that we achieved through the PLS-optimization, resulted in the generation of a unique model capable to predict the four responses at the same time, with the cost of an insignificant percentage of unexplained information. Nevertheless, a less strict consideration of the theoretical frame would yield an eight-component model with the profit of further dimensionality reduction and with the cost of a less accurate, but yet sufficient, prediction of Y2 response variable.

On the contrary, the PCR method generated four individual models, one for the prediction of each response variable, that needed twenty one components to capture 99.95%, 92.51%, 98.21% and 98.17% of the variability of the responses. These percentages, in combination with the results of the information criteria presented in Table 3, demonstrate that FS-MPLSR model is more adequate in all four responses.

6. CONCLUDING REMARKS

The aim of this study is to introduce PLS as a method for variable selection in a variety of fields, including time series analysis. Although this method is commonly used in a regression analysis, it can also be implemented in various other applications such as discriminant analysis, and hierarchical modelling. It can handle complex data sets and situations that cannot be solved by standard methods.

FS-PLS is considered optimal in evaluating more complex structures with a more realistic and holistic view. It has been proved to be a non-time consuming process and statistically efficient method with high prediction accuracy. As a recently found technique in the field, many aspects of its underlying mechanism have recently been revealed and yet, there is no strictly defined frame for its application. As a result, the method is considered to be very flexible and many modifications and experimentations can be tested. In this work the utilization of PLS approach was used as a variable selection criterion and by expansion as a dimension reduction technique. The FS-PLS procedure was able to remove up to 45% and 14% of the original variables in two frequently used datasets in chemometrics, one univariate set and one more complex multivariate one.

Although PLS is considered to be useful in small datasets, through the FS-PLS methodology it has been found to be useful in high-dimensional and/or big data analysis. Although, the applications is chosen from the field of chemometrics, the applicability was quite wide covering biology, physics, chemistry, business, and social sciences among others.

In the univariate case, the final selected model is based on only 217 predictors out of an initial set of 401. The three-component model, which is suggested as optimum, explains the major part of information captured in the data, while it is parsimonious, with high prediction ability and can easily be used for visualizations. The comparison with the corresponding PCR model, which was based on information criteria AIC, R^2_{adj}, and RMSECV, demonstrates that FS-PLSR model gave more sufficient results.

In the multivariate case, the problem appears to be more complicated. Initially FS-MPLSR was implemented on the data out of necessity, due to the fact that correlations were observed between the response variables. We estimated regression coefficients and we determined the significant components for each response variable. We compared the absolute values of the coefficients in significant components with thresholds and then, we defined four sets of predictors, which contained the important predictors for the individual responses, respectively. Their intersection consisted the final set of predictors for the multivariate regression model. This way, in the final selected model 99 less predictors than in the initial set were included. The simultaneous process of the response variables generated a single regression model with AIC values lower than the individual PCR models in all four response variables. The increased number of constructed models in the PCR method is associated with high complexity and computational cost of the whole analysis. This, in combination with the fact that less variability is explained in the second response variable with the PCR method, leads to the suggestion that a FS-PLSR model is optimum also in the multivariate case.

While our work primarily focuses on the predictive performance and feature selection aspects of the proposed FS-PLS methodology, we acknowledge the importance of addressing the issue of interpretability in regression analysis. It is worth noting that interpretability is a complex aspect in high-dimensional settings, and various techniques have been proposed in the literature to enhance it. In future research, we can explore the application of these techniques, such as sparse PLS (SPLS), non-negative matrix factorization (NMF), and independent component analysis (ICA) have been proposed in the literature to enhance interpretability in high-dimensional settings and improve techniques like PCA and PLS. By incorporating these approaches, we can potentially provide a more comprehensive analysis that combines predictive accuracy, feature selection, and enhanced interpretability. This avenue of research holds promise for advancing the field of PLS regression and its applicability in practical domains.

ACKNOWLEDGMENTS

We express our appreciation to the two referees and the Associate Editor for providing valuable suggestions and recommendations that greatly improved both the quality and the presentation of the manuscript. Furthermore, we gratefully acknowledge the invaluable guidance and supervision of Professor Alex Karagrigoriou from the Laboratory of Statistics and Data Analysis, which greatly contributed to the successful completion of this work.

REFERENCES

[1] Ajana, S.; Acar, N.; Bretillon, L.; Hejblum, B.P.; Jacqmin-Gadda, H. and Delcourt C. (2019). Benefits of dimension reduction in penalized regression methods for high-dimensional grouped data: a case study in low sample size, *Bioinformatics*, **35**, 19, 3628–3634.

[2] Artigue, H. and Smith. G. (2019). The principal problem with principal components regression, *Cogent Mathematics and Statistics*, 6: 1622190.

[3] Chung,D. and Keles, S. (2010). Sparse Partial Least Squares classification for high dimensional data, *Statistical Applications in Genetics and Molecular Biology*, **9**, 1, Article 17.

[4] Corn data available from: `http://software.eigenvector.com/data/index.html`.

[5] De Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression, *Chemometrics and Intelligent Laboratory Systems*, **18**, 3, 251–263.

[6] Faraway, J.J. (2002). *Practical Regression and Anova using R*, Retrieved November 1, 2020, from:
`https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf`.

[7] Höskuldsson, A. (1988). PLS regression methods, *Journal of Chemometrics*, **2**, 3, 211–228.

[8] James, G.; Witten, D.; Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R. Springer (1st ed.)*, Springer.

[9] Jiang, J.-H.; Berry, R. J.; Siesler, H. W. and Ozaki, Y. (2002). Wavelength interval selection in multicomponent spectral analysis by moving window Partial Least-Squares regression with applications to mid-infrared and near-infrared spectroscopic data, *Analytical Chemistry*, **74**, 14, 3555–3565.

[10] Kalivas, J. H. (1997). Two data sets of near infrared spectra, *Chemometrics and Intelligent Laboratory Systems*, **37**, 2, 255–259.

[11] Lê Cao , KA.; Rossouw, D.; Robert-Granié, C. and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data, *Statistical Applications in Genetics and Molecular Biology*, **7**, 1, Article 35.

[12] Li, B., Morris, J. and Martin, E.B. (2002). Model selection for partial least squares regression, *Chemometrics and Intelligent Laboratory Systems*, **64**, 1, 79–89.

[13] Li, G.-Z. and Zeng, X.Q. (2009). Feature selection for Partial Least Square based dimension reduction, *Foundations of Computational Intelligence*, **5**, 3–37.

[14] Liquet, B.; de Micheaux, P. L.; Hejblum, B. P. and Thiébaut, R. (2016) Group and sparse group partial least square approaches applied in genomics context, *Bioinformatics*, **32**, 1, 35–42.

[15] Ntotsis, K.; Artemiou, A. and Karagrigoriou, A. (2021). Interdependency pattern recognition in econometrics: a penalized regularization antidote, *Econometrics*, **9**, 44.

[16] Ntotsis, K. and Karagrigoriou, A. (2021). *The impact of multicollinearity on big data multivariate analysis modeling*. In “Applied Modeling Techniques and Data Analysis 1: Computational Data Analysis Methods and Tools”, iSTE Wiley and Sons, 187–202.

[17] Wehrens, H.R.M.J. (2011). *Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences*, Springer, New York.

[18] Wold, S., Sjöström, M. and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, **58**, 2, 109–130.