UNIVERSITY OF THE AEGEAN

School of Sciences
Department of Statistics and Actuarial - Financial Mathematics

MSc in Statistics and Actuarial - Financial Mathematics
Area of Statistics and Data Analysis

Recent Advances on Dimensionality Reduction based on Partial Least Squares Optimization

BEKI ELISAVET

2022

Abstract

High-dimensional data sets give researchers from different fields, the ability to answer various scientific questions. However, their commonly complicated structure and other features, like multicollinearity and noise accumulation, set their handling with classic methods insufficient. New computational and statistical techniques are required to conduct analysis, which can result in reliable conclusions. Such a method is considered to be Partial Least Squares. It is a method with many applications in a wide spectrum of fields, that serves various statistical purposes (such as regression, classification, etc) and though, till recent years, was unfamiliar to a major part of statisticians. The aim of this master thesis is to investigate the prospects of Partial Least Squares Method in both univariate and multivariate level as a reliable method for the analysis of high-dimensional data. The theoretical basis of this dimensionality reduction method is presented along with Principal Component Analysis, to further compare their performance in linear regression problems.

Περίληψη

Τα σύνολα δεδομένων μεγάλων διαστάσεων δίνουν τη δυνατότητα σε ερευνητές διαφόρων πεδίων να απαντήσουν σε επιστημονικά προβλήματα. Ωστόσο, η συχνά πολύπλοκη δομή τους και άλλα χαρακτηριστικά τους, όπως η πολυσυγγραμμικότητα και ο 'θόρυβος', καθιστούν τον χειρισμό τους με κλασσικές μεθόδους, ανεπαρκή. Νέες υπολογιστικές και στατιστικές τεχνικές απαιτούνται ώστε να διεξάγουμε αναλύσεις, οι οποίες μπορούν να οδηγήσουν σε αξιόπιστα συμπεράσματα. Μια τέτοια μέθοδος θεωρείται η Μέθοδος Μερικών Ελαχίστων Τετραγώνων. Πρόκειται για μια μέθοδο με πολλές εφαρμογές σε ευρύ φάσμα επιστημονικών πεδίων, που εξυπηρετεί ποικίλους σκοπούς στατιστικής ανάλυσης (όπως παλινδρόμηση, κατηγοριοποίηση, κλπ.) και, παρ΄ όλα αυτά, παρέμεινε άγνωστη σε μεγάλη μερίδα στατιστικών, μέχρι πριν από λίγα χρόνια. Ο σκοπός αυτής της διπλωματικής εργασίας είναι να διερευνήσει τις δυνατότητες της Μεθόδου Μερικών Ελαχίστων Τετραγώνων στην μονομεταβλητή και την πολυμεταβλητή διάσταση, ως μια αξιόπιστη μέθοδος ανάλυσης δεδομένων μεγάλων διαστάσεων. Στην παρούσα εργασία παρατίθεται η θεωρητική βάση της συγκεκριμένης μεθόδου μείωσης διάστασης, παράλληλα με την μέθοδο Ανάλυσης Κυρίων Συνιστωσών, προκειμένου να συγκριθεί η απόδοσή τους σε προβλήματα γραμμικής παλινδρόμησης.

Acknowledgements

Before we begin, I would like to attribute my acknowledgments to some people for their support in this project.

Firstly, I would like to thank Professor Alexandros Karagrigoriou, of the University of the Aegean, for being my supervisor and guiding me scientifically. His intelligent way of thinking and his rare kindness make him a prototype of a scientist and a teacher.

I would also like to thank Mr. Kimon Ntotsis. I am grateful for all the time he dedicated to my thesis, his continuous willingness to share his knowledge and his support last months.

I could not omit to give thanks to Evangelos Kipouridis. For all the times, in all these years, he tirelessly explained my questions and turned the incomprehensible into simple.

Finally, I would like to express my gratitude to my family and especially George.

Contents

Chapter 1

Dimensionality Reduction

Over the last three decades advances in technology have allowed the collection and storage of datasets, the size of which exceeds what was earlier considered as large, since both their dimension and their sample size have been significantly grown. By dimension, it is meant the number of variables that are measured and their number in these high-dimensional datasets can be hundreds or more. The need for their management and utilization resulted in the bloom of Big Data Analytics, which is a scientific field that searches for ways to systematically extract information from too large or too complicated data sets [33]. In other words, this area is referred to the search of modeling techniques to successfully reveal underlying relationships between available variables, make predictions on similar data and make sufficient summarizing visualizations. Typical fields that generate Big Data are bioinformatics, neuroscience, medicine, health care, Web search, social media analysis, economics and finance.

One of the challenges of Big Data, which demands special handling, is heterogeneity, which is present when values of investigated variables differ and form subpopulations with special properties. Samples that belong to small subpopulations may have been categorized as outliers in small samples, but large data sets give a more realistic picture of the systems that are being analyzed. The phenomenon is commonly noticed when data are collected from different or multiple sources.

However, it is commonly accepted that in Big Data Analytics the problem of a large number of samples can be solved with more convenience than an extremely large number of measured variables. Despite what someone would expect, data sets with too many variables do not always result in

more effective models than lower dimension data sets.

One of the main problems researchers deal when analyze high-dimensional data and refers to the number of available variables, is noise accumulation. Noisy data are the part of a data set that brings meaningless information in it. "*Such a noise accumulation effect is especially severe in high dimensions and may even dominate the true signals*"[8]. Noisy data can significantly effect the models unless they are identified and removed from the dataset before the analysis. They come from bad performance in measurement tools, as mistaken recordings or just random errors. Furthermore, quite often high-dimensional data sets include irrelevant variables that negatively effect the process of analysis.

Additionally, a very common problem is the handling of data which are defined by collinear variables. Multicollinearity, or just collinearity in a dataset is the situation where two or more variables are linearly associated. Performance of models based on multicollinear variables are severly affected by them; this phenomenon is further discussed in subsection Multicollinearity.

Dimensionality Reduction Techniques (also referred to as Dimension Reduction) are used to successively manage these potential problems. They are defined as techniques that reduce the dimension of a dataset and transform it into a set of lower dimension that retains as much as possible information from the original dataset. They are divided into two main categories: Feature extraction and Feature selection methods.

Feature selection methods include algorithms that aim to find irrelevant and/or redundant variables of a dataset. Then, these variables are removed. A new dataset has a lower dimension, as it consists of a subset of the variables of initial dataset. In most techniques in this category, all variables are matched with a value, arising from a criterion. According to its value, every variable is evaluated and it is decided either on its removal or its selection. The physical meaning of retained variables does not change at all. Despite the advantage of interpretability, information captured in interactions and correlations between selected and removed variables is lost [23]. Some popular techniques of this kind are Information Gain, Relief, Fisher Score, Forward Feature Selection, Chi-square Test, Backward Feature Elimination, Lasso and Elastic Net.

On the contrary, application of feature extraction techniques results in transformation of a dataset -data points are projected to a low dimensional space. This is achieved with the use of original variables as elements of

combinations that summarize information from initial variables [23]. This is true because these newly generated variables, also called Latent Variables in the literature, are correlated with the original variables. Among the most applied techniques are Principal Component Analysis, Partial Least Squares Method, Canonical Correlation Analysis and Linear Discriminant Analysis.

Benefits of datasets which have been processed with these techniques are multiple: data complexity can be reduced significantly. As a result models produced by them demand less computation power and time. Furthermore, overfitting is avoided and they can generalize better. Their performance is also improved in terms of accuracy. Finally, their parsimonious representations make data visualizations feasible.

Although, the two mentioned dimension reduction categories have different approaches, there are techniques that can be used so *"feature extraction (transformation) methods can be converted into feature selection methods"* [1][26]. In any case, they can both effectively serve the aim of dimension reduction.

In this thesis, it is investigated a comparison between two feature extraction methods, Partial Least Squares method and Principal Component Analysis. Their performance refers to regression analysis problems and their ability to achieve dimension reduction when high-dimensional datasets are being analyzed. Hence, it is considered necessary to include in the beginning, the basic elements of Linear Regression Analysis, which are recorded in chapter 2. Next, in chapter 3 the theoretical framework these methods are based on is presented. Finally, in chapter 4 these two methods are going to be applied to datasets from the field of chemometrics. The conclusions of their comparison are stated there along with the future work.

Chapter 2

Linear Regression Analysis

2.1 Simple Linear Regression

Linear regression is a statistical method that allows to build a linear model, so as to relate a dependent variable Y to an independent variable X. Then, we can use the linear model to either describe the kind of linear relation between these two variables or to predict the values of Y, given the values of X. The mathematical expression of a linear regression model is

$$Y = b_0 + b_1 X + e$$

where Y and X are the variables intent to be related, b_0 is the intercept and b_1 is the slope. Intercept represents the value of Y when X variable is zero. When b_1 is positive, it expresses that one unit increase in X variable is estimated to increase $|b_1|$ units the response variable Y, and when b_1 is negative, it expresses that the response variable is estimated to decrease $|b_1|$ units. The random variable e is called residual and it contains the difference between Y value and the term $(b_0 + b_1 X)$.

In order to build this linear model, a training set which is nothing but a set of samples (objects) with known X and Y values, is used. For example, let's say that our training set consists of n samples. Theoretically, applying the previous relation, we would get n equations:

$$Y_i = b_0 + b_1 X_i + e_i$$

where Y_i and X_i are the Y and X values of the i^{th} out of a sample of size n.

The next step is to find estimates of the regression coefficients b_0 and b_1 to be able to use the model for predictions and understanding of this relation. The most popular and widely used criterion to achieve that is the minimization of the sum of squared residuals, an approach called Ordinary Least Squares Method (OLS):

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2.$$

Setting the partial derivatives with respect to b_0 and b_1 equal to zero, we get the coefficients estimates $\hat{b}_0$ and $\hat{b}_1$ that can be used to make the predictions of unknown Y values given X values according to the next relation:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X.$$

The previous relation can always be applied, although the reliability of our predictions can be ensured if only the assumptions of regression are fulfilled (see section 2.5).

However, when an analyst defines a linear regression model makes the assumption that in some way X is related to Y. A hypothesis test, where:

$$H_0 : b_1 = 0$$

vs

$$H_1 : b_1 \neq 0$$

can be used to validate this state. For this purpose, t-statistic formed as:

$$t = \frac{\hat{b}_1 - 0}{SE(\hat{b}_1)}$$

where $SE(\hat{b}_1)$ is the standard error of $\hat{b}_1$, and is used to check if $\hat{b}_1$ differs from zero. If that is true, the assumption of relation between X and Y is also true, and therefore X is a statistically significant predictor of Y.

2.2 Multiple Linear Regression

Despite the simplicity of Simple Linear regression, most real-world problems are described by more than just one predictor X. Today, it is very often for statisticians to handle data sets where every sample is described by a set of measurements, say $x_{i1}, x_{i2}, x_{i3}..., x_{ip}$, which are the values of variables $X_1, X_2, ...X_p$ respectively and the value y_i of the response variable, where i is the index for the i^{th} out of n observations of the training set. In this case, the mathematical expression of the linear model that relates Y variable to p X variables is:

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p + e.$$

It is obvious that each additional independent variable is combined with a new regression coefficient. The interpretation of Y, b_k with *k= 1, ..p*, and e are analogous to the simple regression model.

Again, in order to compute the regression coefficients in Multiple Linear Regression (MLR), as this case is called, we make use of the available data, resulting in the following n equations:

$$\begin{aligned} y_1 &= b_0 + b_1x_{11} + b_2x_{12} + \cdots + b_px_{1p} + e_1 \\ y_2 &= b_0 + b_1x_{21} + b_2x_{22} + \cdots + b_px_{2p} + e_2 \\ &\vdots \\ y_n &= b_0 + b_1x_{n1} + b_2x_{n2} + \cdots + b_px_{np} + e_n \end{aligned}$$

which can be summarized to

$$Y = XB + e$$

where Y is the column vector containing the y_i values, $Y = (y_1, y_2, \dots y_n)^\intercal$, B is a column vector containing the regression coefficients $B = (b_0, b_1, \dots b_p)^\intercal$, e is the column vector containing the residuals $e = (e_1, e_2, \dots e_n)^\intercal$ and X is a matrix of size $n \times (p + 1)$, frequently called design matrix. The first column includes n values of 1 and the remaining p columns include the sample values

as shown below:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

Applying the Least Squares Method, where the sum of squared residuals $e^\intercal e$ is minimized, the estimation for the regression coefficients can be computed and summarized in a column vector by the type:

$$\hat{B} = (X^\intercal X)^{-1} X^\intercal Y$$

where $\hat{B} = (\hat{b}_0, \hat{b}_1, \dots \hat{b}_p)^\intercal$.

Further, fitted values can be predicted for a new unseen sample, as long as the values of independent variables are available:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_p X_p.$$

It should be mentioned that it's common practice to use mean-center design matrix to reduce computational cost. The first column of X matrix, the column of ones, is omitted and the linear model is formed as:

$$Y = b_1 X_1 + b_2 X_2 + \cdots + b_p X_p + e.$$

The intercept can be finally computed by taking the difference between the mean of the y values and the predicted values

$$b_0 = \bar{Y} - \hat{Y}.$$

2.3 Multivariate Linear Regression

Finally, there are problems where it is needed to relate a set of independent variables $X_1, X_2, \dots X_p$ to a set of dependent variables $Y_1, Y_2, \dots Y_m$. The statistical problem in multivariate regression, as this kind of regression

is called, is written as:

$$Y = XB + E$$

where Y is a matrix of size $n \times m$, B is a $(p+1) \times m$ matrix of regression coefficients, X is a data matrix as in multiple regression, and E is the $n \times m$ matrix of the residuals. The previous expression compresses m relations of the type:

$$Y_j = Xb_j + e_j$$

where Y_j, b_j and e_j denote the j^{th} column of Y, B and E matrices respectively. According to Ordinary Least Squares (OLS) Method regression coefficients can be estimated by:

$$\hat{B} = (X^\intercal X)^{-1} X^\intercal Y$$

and fitted values can be predicted by:

$$\hat{Y} = X\hat{B}.$$

2.4 Evaluation of Model Performance

When analyzing high dimensional data, it is common to produce models that include different subsets of the original predictors. Once the regression models are defined and possible collinearity issues are faced, one should evaluate the performance of the resulted models. This is usually done using one of the following measures:

- R-squared (R^2) value: It is called the coefficient of determination and denotes the percentage of the variability in the response variable that can be explained by the model. Its formula is:

$$R^2 = \frac{ESS}{SSTo}$$

where *ESS* is the Explained Sum of Squares and *SSTo* is the Total Sum

of Squares [15], since based on the relevant theory

$$ESS = SSTo - RSS,$$

so that R^2 can be written as:

$$R^2 = 1 - \frac{RSS}{SSTo}$$

where RSS is the Residual Sum of Squares. It can be concluded that R^2 values range from 0 and 1, with values closer to 1 indicating better model performance. In univariate and multiple linear regression with a unique response variable a unique R^2 is generated to evaluate model performance, while in multivariate regression an index R^2 corresponds to each response variable.

- Adjusted R^2 value: It is a modification of R^2 value that penalizes more complicated models, meaning models that include more predictors. Its formula is:

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

where n refers to the number of samples and p to the number of included predictors.

- Akaike's Information Criterion: It is an information criterion utilized to compare models with different complexity scales, meaning models that include different number of predictors for a given set of data. Its computation formula for each model is:

$$AIC_{(p)} = -2(maximum\ log - likelihood) + 2p$$

where p is the number of included predictors. In former equation, the first term represents model accuracy and the second, *2p*, relates to model parsimony. In a set of compared models, the model that performs better is the one with the lowest AIC value. As in case of R^2 and R^2_{adj}, in multivariate response variables models, computed AIC indexes should be as many as the response variables, one index for each

response variable.

2.5 Assumptions of Regression Analysis

The predictions arising from the previous models are reliable under specific conditions. In all regression cases, either simple or multiple regression, the most important assumptions that should always be fulfilled are:

- Normality of residuals

  In regression analysis, it should be confirmed that residuals are normally distributed with mean 0, regardless of the distribution of variables X and Y. Normal distribution of residuals can be detected graphically with a Q-Q plot, where the dots, that represent the samples, are expected to form a straight line. Except from this, there are also hypothesis tests, like Shapiro Wilk and Lilliefors that can be used to check the normal distribution of residuals. In both statistical tests as null hypothesis is set, the normality of residuals. In case of non-normal residuals, the problem can be solved by transformation of the response variable, like logarithm. However, short-tailed distributions of residuals do not have a severe impact on regression reliability and the violation of this assumption could be ignored. Further, as far as large data sets and Big Data, *"with a large dataset, even mild deviations from non-normality may be detected, but there would be little reason to abandon least squares because the effects of non-normality are mitigated by large sample sizes"* [9].

- Independence of residuals

  The assumption that error terms are uncorrelated is very important. It means that the sign of e_i is not associated with the sign of error term e_{i+1}. Graphically, it can be detected by plotting the residuals against time. Durbin-Watson statistical test can be used for this purpose as Run Test, too. In the latter, the null hypothesis being checked is that residuals are randomly ordered. In the first case, the null hypothesis is that residuals are not autocorrelated against a correlated special structure. It is applied after validation of normality of residuals and is based on the argument that normally distributed and uncorrelated residuals are independent [11].

- Homoscedasticity of residuals

  This assumption concerns residual variance. More precisely, in the case of regression analysis, the residuals must have common variance (σ^2), and by extension the same standard deviation (σ). To confirm that this assumption is not violated, Levene's test is used to check the null hypothesis of equality of all population variances. Additionally, the general picture can be captured in a plot of residuals against fitted values ($\hat{Y}$).

- Linearity

  It is the only assumption not referring to residuals. It is a fundamental assumption that linear relationship between the variables should exist in order for regression results to be statistically trustworthy. Linearity can be confirmed by a scatterplot of fitted values against the residuals. It is expected an approximately straight line through the middle of the plot. Note that in the case of Big Data Analytics, where a large amount of variables exists, this assumption is fulfilled by a random sampling of approximately 10%-30% of the possible combinations, and if linearity exists in the random sample, then the generalization of the assumption in the total population can be made.

Additionally, another factor that can negatively affect the regression model is the existence of "anomalous" data in the training set. Such observations lie away from the regression line and consequently have large residuals, but also, they are points with high leverage, meaning they differ significantly from the rest training set. Cook's distance is a measure used to evaluate the influence of an observation on the regression model and to indicate such potential points.

2.5.1 Multicollinearity

Another issue, often met in multivariate data sets is the existence of multicollinearity among predictors. Multicollinearity, or Collinearity as is also called, is present in a data set when two or more predictors are correlated.This can be expressed as follows: let's make the assumption that j out of p variables of X data matrix are correlated, meaning:

$$\sum_{i=1}^{j} a_i X_i = 0$$

where there are a_j different from zero that make the equation true. This case is called perfect or structural multicollinearity and *"it caused by generating predictors with the use of already existing ones"*[29]. It could be said that it arises as a non-appropriate definition of a regression problem set by the analyst, and as such it can be easily handled, by identifying and removing the variables that can be predicted by others as a linear combination of them.

However, in real-world regression problems there is also the case that the relation between the predictors is approximately linear, a case that is also described by the term multicollinearity, and it can be written as:

$$\sum_{i=1}^{j} a_i X_i \approx 0$$

for a_j different from 0. This case is called data-based multicollinearity, or simply high multicollinearity and it is presented when data come from observational experiments. It is attributed to the structure of available data. It is met more often than perfect multicollinearity and needs different dealing.

In the first case, the matrix $X^\intercal X$ becomes singular, since two or more of its columns are dependent. As a result, the OLS method can no longer be used to estimate regression coefficients b_j, since there is no inverse matrix $(X^\intercal X)^{-1}$.

In the second case, when the inverse matrix $(X^\intercal X)^{-1}$ can be computed successfully, problematic situations are created:

It is difficult to distinguish the effects of predictors on the response variable. A regression coefficient expresses the impact of a one-unit change in the correspond predictor, when the rest of the predictors are unchanged. However, when predictors are correlated, they have a common variability pattern, meaning that they simultaneously increase, or oppositely decrease. As a result, it is difficult to estimate how much correlated predictors effect on the response variable. The apparent effects are misleading.

Additionally, when multicollinearity is presented there are many combinations of estimated coefficients that all result in similar predictions and RSS statistic. This means that the standard error of the estimates is high. Consequently, the t-statistic, $t = \hat{b}_j / SE(\hat{b}_j)$ is declined, which reduces the power of hypothesis test $H_0 : b_j = 0$. This implies that zero coefficients may not be

detected as so, but mistakenly appear to be statistically significant, a fact that leads to the existence of redundant variables. In these cases, R^2, the measure to evaluate the goodness of fit (i.e. the performance) of a model, will be large and the produced model will be interpreted as sufficient. However, redundancy leads to overfitting of the regression model, a situation where the regression model is unable to perform sufficiently for samples that do not belong in the training set. Further, since the produced estimates of coefficients are not precise, the predictions of samples lying far from the space covered by the training set will also be imprecise.

Moreover, the regression model becomes unstable. Small changes in training set could lead to large changes in estimates. It is even possible to result in estimates with inverse signs.

Examination of the correlation matrix of the predictors can reveal pairwise collinearities. High values of correlation between two variables may be an index of collinear relation between them. Nonetheless, Variance Inflation Factors (VIFs) are typically used to reveal such problematic cases. This measure can be computed as:

$$VIF_k = \frac{1}{1 - R^2_{(-k)}}$$

where $R^2_{(-k)}$ is the R^2 value of the regression model that uses X_k as the response variable and the rest variables of the dataset as predictors. When X_k is correlated to other variables, it can be easily linearly predicted by them. $R^2_{(-k)}$ will be close to one and VIF index will be large.

This measure *"expresses the rate at which the variance of the estimator increases when collinearity exists"* [29] and it is obvious in the following [9]:

$$var(\hat{b}_k) = \sigma^2 \frac{1}{S_{x_j x_j}} (\frac{1}{1 - R^2_{(-k)}}) = \sigma^2 \frac{1}{S_{x_j x_j}} VIF_k$$

where

$$S_{x_j x_j} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2.$$

It is revealed that increased VIF_j, as is the case in multicollinearity,

means increased variance of the estimator. A widespread threshold that is used to infer if a predictor is collinear, is the value 5 (a more relaxed threshold is 10). When the VIF index is higher than the set threshold, the predictor is considered collinear.

Another way to detect multicollinearity is to make use of the condition number and the condition index. According to this method the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ of $X^\intercal X$ matrix are computed and their maximum λ_{max} is identified. The ratio

$$k_s = \sqrt{\frac{\lambda_{max}}{\lambda_s}}$$

for s in (1,... p), is called condition index. The largest of the s indices is the condition number and is associated with the detection of multicollinearity. When this value is more than 10 and less than 30, multicollinearity exists in the data set. When it exceeds 30, multicollinearity is severe and effects regression results significantly. Additionally, when more than one condition index is high, it means that the phenomenon is caused by more than one linear combination.

Furthermore, as stated in [29] a predictor which appears to be significant in simple linear regression with the response variable and insignificant in multiple linear regression with the same response variable could be an indication of multicollinearity. An other indication is significant changes in the estimated coefficients when a predictor is removed from the design matrix X.

As revealed from above, multicollinearity can negatively affect resulting estimations and inferences when present in a regression problem. In the case of perfect multicollinearity, it is logically concluded that a re-definition of the description of the problem, by removing the predictors, which are linear combinations of other predictors, eliminates the problem. In the case of high collinearity, high correlated predictors should be removed after their identification. In Big Data analysis, where the number of predictors can be hundreds, it is more sufficient to eliminate multicollinearity with Dimensionality Reduction Techniques. More specifically, Partial Least Squares Method and Principal Component Analysis can be used as a pre-processing step before the regression analysis, to eliminate the problem by generating uncorrelated variables produced by combinations of original variables.

Chapter 3

Dimension Reduction Techniques

Scientists in every scientific field often deal with datasets with special properties that make Ordinary Least Squares Method an inappropriate choice for modeling. Such a case is a situation where analysts have access to a small number of observations (let's say n) described by a larger number of variables (p). Another problematic situation is a dataset with highly correlated predictors that contain similar information. As explained in subsection 2.5.1, multicollinearity leads to unstable OLS regression models or, even worse, makes it impossible to calculate the model since the $(X^\intercal X)$ matrix will be singular. Feature selection methods, like Forward Feature Selection, Backward Feature Elimination, Lasso and Elastic Net have been proposed to overcome these problems. However, these techniques are often time-consuming and include high risk of omitting significant variables. An alternative sufficient approach to overcome these situations and manage dimension reduction are Latent Variable methods, like Principal Component Analysis (PCA) and Partial Least Squares (PLS), which are feature extraction methods. As stated in [34], *"latent variables are variables that are not directly observed but are rather inferred from other variables that are observed (directly measured)"*. They arise as linear combinations of the original variables so as to compress information from given data. This property, in combination with the fact that in PCA and PLS methods uncorrelated Latent Variables are produced, makes it feasible to extract information from data and further relate it to response variables, when the OLS method fails.

PCA is a method first described by Pearson in 1901 [30] and Hotelling in

1933 [14] independently. Pearson approached the main idea from a geometric point of view, while Hotelling described it more algebraically. Development of computers was the determining factor in widespread application of the method in other sciences, since processing of large datasets with this method was finally feasible. In PCA, latent variables are called Principal Components or simply components. These three terms are equivalent and in the text may alternate.

PLS or Projection to Latent Variables, as sometimes is called, is also a method that uses latent variables to model complicated data, in terms of dimension and collinearity. It was first developed by Herman Wold around 1975 and applied in econometrics. Later, Svante Wold and Harald Martens applied it in chemometrics [40]. As stated in [31], *"PLS regression used to be overlooked by statisticians and is still considered rather an algorithm than a rigorous statistical model"*. However, nowadays it is a method with many variations used in numerous fields. Here, Latent variables are also called (Principal) Components, a strong evidence of the common philosophy that PLS shares with PCA.

3.1 Principal Component Analysis

PCA is a commonly applied dimension reduction method. It is a multivariate technique in which a data matrix X, which includes correlated variables, is transformed into a new one. Variables in the new matrix, also called Principal Components, are uncorrelated and ordered so as to contain variance of the original X matrix on a declining scale, starting from the first one. The beneficial property of the new matrix is that most variation of X matrix is compressed in the first few new variables. These variables, the number of which is selected by the user, form a low-dimension matrix, an approximation of X that can be used for modelling purposes.

As stated in [38] *"PCA can be seen as a method to compute a new coordinate system formed by the latent variables, which is orthogonal, and where only the most informative dimensions are used."* From a geometrical point of view, an X matrix is projected/mapped to a new space (hyperplane, plane or line), the coordinate system of which is formed by the Principal Components. Additionally, Principal Components are oriented in the direction of maximized variance of data points. *"The coordinates of the samples in the new space are called scores, often indicated with the symbol T. The new dimensions are linear combinations of the original variables and are called loadings (symbol P)."*[39].

The Principal Components Analysis of a X matrix of size $n \times p$ is:

$$X = T_p P_p^\intercal$$

where T is the matrix containing the scores of the samples, P is the matrix containing the loadings, and superscript $^\intercal$ indicates the transpose of a matrix. Subscript p indicates the number of latent variables that can be computed. However, as only a few Principal Components are almost always used for modeling, since they suffice to explain most of the variance included in X, the original matrix can be written as:

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

where m, ranging from 1 to p, indicates the number of selected latent variables and E is the matrix containing the residual error. Geometrically, that is, the perpendicular distance of each point onto the hyperplane formed by loading vectors [6]. These quantities represent the loss of information because of the projection of X data points into a low-dimension space. Finally, the new,

low-dimension matrix can be written as:

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

where $\tilde{X}$ indicates the approximation of X, that can be used for modeling purposes discharged of noise.

Steps to build a PCA model

The first step in PCA is centering data matrix X in order to remove arbitrary bias from measurements [6]. This is achieved by replacing each x_{ij} element by:

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

where $(\bar{x}_j)$ indicates the mean value of column j. After this process, in the mean-centered matrix X, each column has a mean of zero. This technically means that data points have been moved to the center of the coordinate system while the distances between them do not change at all.

In some cases, datasets include variables of different magnitudes, because they are measured in different units. As a result, some variables have different statistical weights in the analysis. This problem can be solved by replacing each x_{ij} element by:

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

where s_j indicates the standard deviation of the j_{th} variable, a process called scaling. In this way, the final columns in X have a mean of zero and a unit variance, and it should be noted that in this case the relative distance between data points is changed. However, if predictors are measured in the same units, scaling could cause the inflation of noise in uninformative variables [39].

Next step is to compute matrices of Principal Components.

Singular Value Decomposition

A common applied technique to compute scores and loadings is the Singular Value Decomposition (SVD), according to which the mean-centered X matrix is decomposed as:

$$X = UDV^{\intercal}$$

where U is a matrix of size $n \times n$ and its columns are the left singular vectors of X, while V is a matrix of size $p \times p$ and its columns are the right singular vectors of X. Matrices U and V are orthogonal, meaning each column is orthogonal to the others. Matrix D is a diagonal $n \times p$ matrix, where diagonal elements d_i are related to variances of corresponding principal components. These quantities can be computed by:

$$\lambda_i = \frac{d_i^2}{n-1}.$$

Finally, matrices U, D, V, T, and P are related to each other as follows:

$$X = (UD)V^{\intercal} = TP^{\intercal}$$

meaning that the matrix of loadings P is set equal to matrix V, while the matrix of scores T is set equal to matrix (UD).

Eigen Decomposition

In the case of data sets with many original variables, the SVD process is considered computationally demanding and it is avoided. Instead, an other method, called Eigen decomposition is applied to either covariance matrix Σ or correlation matrix $\rho(X)$.

Each element in a covariance matrix represents the covariance between two variables, a quantity that measures the joint variability of them [35] and it is computed by:

$$cov(X_i, X_j) = E[(X_i - E(X_i))(X_j - E(X_j))]. \tag{3.1}$$

Given a data set that includes $X_1, X_2, \ldots, X_p$ variables, covariance matrix is symmetric as shown below:

$$\Sigma = \begin{bmatrix} cov(X_1, X_1) & cov(X_1, X_2) & ... & cov(X_1, X_p) \\ cov(X_2, X_1) & cov(X_2, X_2) & ... & cov(X_2, X_p) \\ \vdots & \vdots & & \vdots \\ cov(X_p, X_1) & cov(X_p, X_2) & ... & cov(X_p, X_p) \end{bmatrix}$$

Diagonal elements of the Σ matrix represent variances of variables, since by definition

$$Var[X_i] = E[(X_i - E(X_i))^2] = E[(X_i - E(X_i))(X_i - E(X_i))] = cov(X_i, X_i).$$

As a result, this matrix is also called the variance-covariance matrix.

However, in PCA, the original X variables are mean-centered and equation (3.1) becomes:

$$cov(X_i, X_j) = E[X_i X_j].$$

In terms of matrix, that includes all p variables, this can be written as $E[X^\intercal X]$ and when variance-covariance matrix refers to sample data set it is equal to $X^\intercal X$ matrix.

Correspondingly, each element in the correlation matrix represents a correlation between two variables, a measure of the linear relationship between them, and it is computed by:

$$cor(X_i, X_j) = \frac{cov(X_i, X_j)}{\sqrt{Var[X_i]}\sqrt{Var[X_j]}}.$$

Given a data set that includes $X_1, X_2, \ldots, X_p$ variables, the correlation matrix is symmetric and diagonal elements are equal to one, as shown below:

$$\rho(X) = \begin{bmatrix} 1 & cor(X_1, X_2) & \dots & cor(X_1, X_p) \\ cor(X_2, X_1) & 1 & \dots & cor(X_2, X_p) \\ \vdots & \vdots & & \vdots \\ cor(X_p, X_1) & cor(X_p, X_2) & \dots & 1 \end{bmatrix}$$

since

$$cor(X_i, X_i) = \frac{cov(X_i, X_i)}{\sqrt{Var[X_i]}\sqrt{Var[X_i]}} = \frac{Var[X_i]}{Var[X_i]} = 1, \forall i.$$

It should be noted that correlation is independent of the scales of variables, while covariance is not. This is the reason why the correlation matrix is used when variables have different measurements; oppositely, the covariance matrix is used when all variables in X express the same measurement unit. Note that even though the correlation matrix supposedly handles the differentiation between the units measurements of the variables, data standardization is highly recommended in the presence of extreme multicollinearity regardless of the selected matrix.

Finally, the basic elements of the decomposition process are the eigenvalues and the corresponding eigenvectors, which are related according to equation:

$$Xv = \lambda v.$$

As stated in [36] *"In this equation X is an n-by-n matrix, v is a non-zero n-by-1 vector and λ is a scalar (which may be either real or complex).Any value of λ for which this equation has a solution is known as an eigenvalue of the matrix X. It is sometimes also called the characteristic value. The vector v, which corresponds to this value is called an eigenvector".*

As mentioned above, in PCA the PCs are oriented in directions of maximal variance of data points. In other words, the method initially aims to find the direction of a unit length vector p_1 that maximizes the variance of score, i.e the values that are loaded in vector t_1. This is equivalent to maximizing the function g:

$$g = t_1^\intercal t_1$$

under the constraint $p_1^\intercal p_1 = 1$ and considering that $t_1 = Xp_1$. Hence:

$$g = t_1^\intercal t_1 = p_1^\intercal X^\intercal X p_1. \tag{3.2}$$

Using Lagrange multiplier, equation (3.2) can be transformed into:

$$g = t_1^\intercal t_1 = p_1^\intercal X^\intercal X p_1 - \lambda(p_1^\intercal p_1 - 1). \tag{3.3}$$

Taking partial derivatives:

$$\frac{\partial g}{\partial p_1} = 0 \Rightarrow \frac{\partial\{p_1^\intercal X^\intercal X p_1 - \lambda(p_1^\intercal p_1 - 1)\}}{\partial p_1} = 0 \Rightarrow$$

$$2X^\intercal X p_1 - 2\lambda p_1 = 0 \Rightarrow$$

$$(X^\intercal X - \lambda I_{p\times p})p_1 = 0 \Rightarrow$$

$$X^\intercal X p_1 = \lambda p_1. \tag{3.4}$$

From equation (3.2) and considering equation (3.4) :

$$t_1^\intercal t_1 = p_1^\intercal (X^\intercal X p_1) \Rightarrow$$

$$t_1^\intercal t_1 = p_1^\intercal \lambda p_1 = \lambda p_1^\intercal p_1 \Rightarrow$$

$$t_1^\intercal t_1 = \lambda. \tag{3.5}$$

Similarly, the rest requested directions of PCs can be computed, under the additional constraint of orthogonality among all of them.

Finally, it is concluded that the directions of PCs are the directions of eigenvectors of covariance matrix $X^\intercal X$. Therefore, loadings matrix P is formed by setting as columns the eigenvectors of $X^\intercal X$ and they are ordered according to the value of respective eigenvalue. Eigenvectors with larger eigenvalues are set first. In this way, arising PCs, which consist of the columns

in the XP product matrix, have maximum variance, because their variance is equal to the respective eigenvalue, as shown in equation (3.5).

Eventually, the mechanism of eigenvalue decomposition of a set of predictors X, where variables X_i are mean-centered, can be summarized in two steps: Creation of the covariance or correlation matrix and the computation of its eigenvectors and eigenvalues. Finally, order the eigenvalues in a declining scale and form loadings matrix P using the eigenvectors. This matrix can be used to produce scores matrix T, by setting $T = XP$.

Choosing the number of Principal Components

The major aim of PCA is dimension reduction. In other words, PCA is applied to replace the p variables-columns of an X matrix by a smaller number m of PCs, without discarding a significant amount of information [17]. Although, typically p PCs can be computed, it's meaningless to work with all of them. The crucial question is how many PCs should eventually be included in the PCA model. The answer is not straightforward, as the analyst should consider a trade-off between information loss and the insertion of noise. Next, are presented the most often approaches used to determine the appropriate number of PCs:

- **Cumulative Percentage of Total Variation**

  A direct estimate of the appropriate number of PCs can be formed by the inspection of the cumulative percentage of total variation of X, that can be explained by the inclusion of different numbers of PCs. It should be noted that percentage of variance explained by the i^{th} PC can be computed by the formula [39]:

  $$q_i = 100\frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j}$$

  where λ_i refers to the eigenvalue of the i_{th} PC and p the overall number of PCs. Usually, one selects the first m PCs, which absorb $80\% - 90\%$ of initial data variation [39]:

$$\sum_{j=1}^{m} q_j = 100 \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

- **Size of variances of PCs**

  This approach is also called Kaiser's Rule and is mainly applied in cases where PCs are generated by the analysis of correlation matrix. According to [18], PCs are included in the PCA model as long as their variance is larger than 1. However, it should be mentioned that in [16] is suggested a lower variance threshold, a value of 0.7, due to independency conditions and sampling variances. In PCA of covariance matrix, sufficient threshold is considered the average value of the eigenvalues.

- **The scree graph**

  It is a graphic way to judge the number of PCs. On scree graph, also called scree plot, the proportion of variance explained by each PC against the rank of each PC is plotted. Usually, the curve that connects the points forms an elbow-like shape. The point located at its angle indicates the last PC to be included in the PCA model [4].

There have also been proposed Cross-validation and bootstrap techniques, but they are not common due to the computational cost, especially when processing large data sets [17].

3.1.1 Principal Component Regression

The Principal Component Regression (PCR) is a linear regression method that uses PCA and a regression step to overcome weaknesses of Multiple Linear Regression (MLR). It is applied in case of multicollinearity between the predictors or/and in case the number of predictors is large compared to the number of available samples. Then, X matrix can be decomposed according to PCA and, after determining the number of PCs retaining to

the model, low-dimension matrix T is used in MLR instead of X matrix. As described in [39]:

$$\begin{aligned} Y = XB + E = \tilde{X}B + E' = (TP^{\intercal})B + E' \\ = T(P^{\intercal}B) + E' = TA + E'. \end{aligned}$$

It is obvious that PCR initially decomposes the data matrix and then replaces it with scores matrix T in a regression step. Matrix A indicates the regression coefficients to be computed. The formula for this is:

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

as computed in classic Linear Regression when Ordinary Least Squares Method is applied. It should be noted that values in A matrix refer to scores. Matrix B with the regression coefficients that refer to the original variables of X can by computed by:

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

3.1.2 Advantages and Drawbacks of PCR

It is obvious that PCR is a simple technique carrying many advantages compared to MLR, such as:

- Columns in matrix T are independent. As a result, the matrix $T^{\intercal}T$ is numerically stable and therefore the consequences of collinearity in estimating regression coefficients are eliminated.

- T columns used in the final regression step include information separated by noise.

- Of course, since dimension reduction is achieved by selecting only m out of p PCs to retain in the PCA model, the complexity of the final regression model is significantly optimized.

- PCs can be used to create informative visualizations of multivariate datasets. Also, PCA can be combined with other (supervised) methods, beyond PCR, where PCA is used to pre-process the dataset.

- Unlike MLR, the PCR can be applied even when the number of available samples is lower than the number of regressors in a dataset.

On the other hand, the main drawbacks of the method are:

- Coefficients arising from the regression step of PCR refer to new variables. Therefore, they are not interpretable.

- Transformation of original variables to PCs can cause the loss of vital information, especially for prediction purposes.

- The selection of the number of PCs to be retained in the PCR model is subjective.

3.2 Partial Least Squares Method

In PCA, analysis is applied on the X data matrix. As a result, arising PCs retained in the model may not contain information related to the dependent-response variables they are next meant to predict in the regression step. In other words, valuable for prediction purposes information may be summarized in PCs that are not included in the PCA model. Partial Least Squares (PLS) is an alternative method that copes with this deficiency of PCA and, at the same time, achieves dimension reduction. As stated in [31], *"It comprises of regression and classification tasks as well as dimension reduction techniques and modeling tools."* This Master thesis refers to the PLS version that aims to model relation among two blocks of variables, a matrix X containing predictors and a matrix Y containing responses, although other versions have been used to relate more blocks. It is used when the number of predictors is large compared to the available samples and/or multicollinearity is present.

This method shares the same main idea as PCA: it forms new variables,as linear combinations of the original, which are uncorrelated. The difference is that they retain information involved both in X and Y data matrices. So, here, the aim is to generate latent variables in the direction of maximum covariance between X and Y: as stated in [39], *"PLS explicitly aims to construct latent variables in such a way as to capture most variance in X and Y, and to maximize the correlation between these matrices".* However, the algorithm to achieve this goal is a bit more complicated, since both X and Y matrices are analyzed. New latent variables are generated through an iterative procedure, in every step of which it is computed a set of scores vectors, a set of loadings, but also a set of weight vectors. One vector of each type refers to X matrix and the other refers to Y matrix. Next, follows deflation of X and Y matrices, so as to subtract the information explained by the computed components. Deflated matrices are used in the next iteration of the algorithm to generate new components. Finally, the user selects the number of components of X matrix that will be used. Of course, their number is significantly reduced compared to the number of original variables, since valuable information is summarized in the first few PCs.

Geometrically, just like PCA, X but here also Y dataset are projected to new spaces, the coordinate systems of which are formed by new latent variables formed by linear combinations of original ones as mentioned before. Latent variables of X matrix are generated so as to be orthogonal, and as a

result uncorrelated, but this is not necessarily the case for latent variables of Y matrix.

Partial Least Squares analysis of an X matrix of size $n \times p$ and a matrix Y of size $n \times k$ is:

$$X = T_\alpha P_\alpha^\intercal$$

$$Y = U_\alpha Q_\alpha^\intercal$$

where T and U are scores matrices. Again, the i^{th} columns in T and U matrices, are the coordinates of the samples in the direction of i^{th} new latent variable, values arising from the perpendicular projection of each sample onto this direction and are measured from the origin. P and Q are loadings matrices, and the superscript $^\intercal$ indicates the transpose of a matrix. In the PLS model, loadings are vectors used in the deflation process of sequential deflated matrices generated through the algorithm. The subscript α indicates the number of latent variables that we usually compute and is equal to min(n,p,k). Since, only the first few Components are adequate for the following modeling purposes, the original matrices can be rewritten as [31]:

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

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

where m is a number smaller than α and indicates the number of latent variables retained to the model. Matrices E and F contain information not explained by the first m selected Components. Eventually, the new, low-dimension matrix can be written as:

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

where $\tilde{X}$ indicates the approximation of X, that can be used for further modeling purposes.

Steps to build a PLS model

The very first step in building a PLS model is mean-centering columns in matrices X and Y, so that each one has zero mean. Additionally, in cases of variables measured in different units, scaling should be considered. The reason and technique for mean-centering and scaling is the same as in PCA model.

Concerning the computation of scores, loadings and weight vectors, plenty of algorithms have been proposed. Some of the most known are named: The Eigenvector algorithm, by Hoeskuldsson [12], Kernel algorithm for PLS introduced by Lindgren et al [24], NIPALS algorithm introduced by H. Wold in 1975 [42], SIMPLS algorithm for PLS proposed by de Jong [6], Orthogonal Projections to Latent Structures (O-PLS) proposed by Trygg and Wold in 2002 [37]. Present Master thesis display Nonlinear Iterative Partial Least Squares algorithm (NIPALS).

NIPALS Algorithm

The NIPALS algorithm constructs the PLS model's matrices in sequential steps. In every step X-scores vector, denoted by t and results from the projection of X-data matrix on the direction of the new latent variable, is about to be constructed in X data space. Y-score vector, denoted by u, arises alike. These directions are defined by weight vectors w and c, respectively. Mathematically, this can be written as:

$$t = Xw$$

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

Directions of the w and c vectors are found so as to maximize the covariance between score vectors t and u, a value proportional to the quantity $t^{\intercal}u$. An additional constraint on unit length weight vectors is applied. To sum up, the problem to be solved can be written as [31]:

$$max\{cov(t,u)\} = max\{t^{\intercal}u\} = max\{(Xw)^{\intercal}Yc\} = max\{w^{\intercal}X^{\intercal}Yc\}. \quad (3.6)$$

In every step, the computation of weights and scores is followed by the

deflation of X and Y matrices. This process is based on p and q loadings of X and Y matrices respectively, computed as:

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

$$q = Y^\intercal u/(u^\intercal u).$$

Actually, there are several variations on how to run deflation. The choice depends on the aim of PLS modeling. Herman Wold initially proposed to deflate matrices as follows:

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

$$Y_{new} = Y_{old} - uq^\intercal.$$

This version of deflation of Y is used when the PLS model is built to reflect relations between blocks of variables and the algorithm is called PLS Mode A. In case a PLS model is built for prediction, the algorithm is called PLS1, when there is only one Y response variable to be predicted, and PLS2, when also Y matrix, as X, is multidimensional. The rest of this Master Thesis is focused on PLS1 and PLS2, namely PLS models for regression. These variations run deflation as follows:

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

$$Y_{new} = Y_{old} - btc^\intercal$$

where

$$b = u^\intercal t/(t^\intercal t). \tag{3.7}$$

When the c vector is not scaled to have a unit length (and this is most frequently the case), as shown below, b is equal to one [12]:

$$u^\intercal t = c^\intercal Y^\intercal t/(c^\intercal c) = c^\intercal (Y^\intercal t)/(c^\intercal c) = c^\intercal c(t^\intercal t)/(c^\intercal c) = t^\intercal t \tag{3.8}$$

and from equations (3.7) and (3.8)

$$b = u^\intercal t/(t^\intercal t) = t^\intercal t/(t^\intercal t) = 1.$$

In the following, the c vector is not supposed to be scaled, so b is considered equal to one. Because of this, deflation of Y matrix becomes:

$$Y_{new} = Y_{old} - tc^\intercal.$$

Once deflation is completed new matrices X_{new} and Y_{new}, also called residual matrices, are analyzed in the next step so that the next latent variable can be extracted. The procedure for PLS2 can be described by the following pseudocode:

Algorithm 1 Pseudocode for PLS2

Input: A data set consisted by a $n \times p$ matrix X and a $n \times k$ matrix Y, where each X_j and Y_j column represents a variable.

Output: Vectors t, p, w and c

Step 1: Set the vector u as the first or any other column of Y

Step 2: Compute X weight: $w = X^\intercal u/(u^\intercal u)$

Step 3: Scale w to be unit length vector, $||w|| = 1$

Step 4: Compute X scores: $t = Xw$

Step 5: Compute Y weight: $c = Y^\intercal t/(t^\intercal t)$

Step 6: Update u scores vector: $u = Yc/(c^\intercal c)$

Step 7: Test convergence of ratio $v = ||t_{old} - t_{new}||/||t_{new}||$

- If $v > \epsilon$, go to step 2 (where ϵ set to a number between $(10^{-8}, 10^{-6})$ for instance)
- If $v < \epsilon$, go to step 8

Step 8: Compute X loadings: $p = X^\intercal t/(t^\intercal t)$

Step 9: Deflation process: $X_{new} = X - tp^\intercal$ and $Y_{new} = Y - tc^\intercal$

Step 10: Set $X = X_{new}$ and $Y = Y_{new}$ and go to step 2

The way t scores are derived implies that they also contain information about Y. As a result, they are also good predictors of Y (and that is the reason that deflation of Y matrix is done by subtracting $tc^\intercal = tt^\intercal Y/(t^\intercal t)$, where $t^\intercal Y/(t^\intercal t)$ is the OLS estimate v of coefficient in regression $Y = tv$).

At this point, it should be mentioned that in [41] it is supported that

deflation of Y matrix is optional, since as it is stated there *"the results are equivalent with or without Y-deflation"*.

With this in mind, the procedure for the PLS1 version is more simple as shown below:

Algorithm 2 Pseudocode for PLS1

Input: A data set consisted by a $n \times p$ matrix X and a $n \times 1$ matrix Y, where each X_j column represents a variable.

Output: Vectors t, p, w and c

Step 1: Set the vector u as the Y column, the unique vector of response variable

Step 2: Compute X weight: $w = X^\intercal u/(u^\intercal u)$

Step 3: Scale w to be unit length vector, $||w|| = 1$

Step 4: Compute X scores: $t = Xw$

Step 5: Compute Y weight: $c = Y^\intercal t/(t^\intercal t)$

Step 6: Compute X loadings: $p = X^\intercal t/(t^\intercal t)$

Step 7: Deflation process: $X_{new} = X - tp^\intercal$

Step 8: Set $X = X_{new}$ and go to step 2

The maximal number of such components that have nonzero covariance with Y is $\min(n-1, p)$, where n the number of samples and p the number of variables in X matrix [3].

The way the weight vectors are found ensures that these give the solution to the problem formulated in equation (3.6). To prove that we can use the fact that non deflation of Y matrix does not influence the results. Further, let us

denote as $X_i^\intercal$ the residual matrix that is going to be used for the construction of the i^{th} latent variable. For this i^{th} dimension we denote as w_{n-1} the weight vector of $(n-1)^{th}$ iteration of steps 2 to 6 before the convergence and as w_n the weight vector of n^{th} iteration before the convergence. Then as stated in [12], the weight vector can be analyzed as:

$$\begin{aligned} w_n &= X_i^\intercal u_{n-1}/(u_{n-1}^\intercal u_{n-1}) = \\ &= X_i^\intercal Y c_{n-1}/(u_{n-1}^\intercal u_{n-1})(c_{n-1}^\intercal c_{n-1}) = \\ &= X_i^\intercal Y Y^\intercal t_{n-1}/(u_{n-1}^\intercal u_{n-1})(c_{n-1}^\intercal c_{n-1})(t_{n-1}^\intercal t_{n-1}) = \\ &= X_i^\intercal Y Y^\intercal X_i w_{n-1}/(u_{n-1}^\intercal u_{n-1})(c_{n-1}^\intercal c_{n-1})(t_{n-1}^\intercal t_{n-1}). \end{aligned}$$

Considering that the i^{th} latent variable takes s iterations till convergence is achieved, then we can conclude that w_s and w_{s-1} do not differ significantly. So, by previous equation vector w_s is eigenvector of matrix $X_i^\intercal Y Y^\intercal X_i$.

Accordingly, we can find that c_s is the eigenvector of matrix $Y^\intercal X_i X_i^\intercal Y$ [12]:

$$\begin{aligned} c_n &= Y^\intercal t_n/(t_n^\intercal t_n) = \\ &= Y^\intercal X_i w_n/(t_n^\intercal t_n) = \\ &= Y^\intercal X_i X_i^\intercal u_{n-1}/(t_n^\intercal t_n)(u_{n-1}^\intercal u_{n-1}) = \\ &= Y^\intercal X_i X_i^\intercal Y c_{n-1}/(t_n^\intercal t_n)(u_{n-1}^\intercal u_{n-1})(c_{n-1}^\intercal c_{n-1}). \end{aligned}$$

Eventually, w_s and c_s are the first eigenvectors that correspond to the largest eigenvalue of $X_i^\intercal Y Y^\intercal X_i$ and $Y^\intercal X_i X_i^\intercal Y$ matrices, respectively. Therefore, from SVD properties, these vectors maximize the quantity of interest in equation (3.6).

Furthermore, PLS latent variables not only explain maximum covariance between X and Y. An other useful property of them is that they are mutually orthogonal. The retroactive relation between residual matrices of the PLS

model will help to prove it [12].

$$X_j = X_{j-1} - t_{j-1}p_{j-1}^\intercal =$$

$$= X_{j-1} - X_{j-1}w_{j-1}t_{j-1}^\intercal X_{j-1}/(t_{j-1}^\intercal t_{j-1}) =$$

$$= X_{j-1}[I - w_{j-1}t_{j-1}^\intercal X_{j-1}/(t_{j-1}^\intercal t_{j-1})] =$$

$$= [X_{j-2} - t_{j-2}p_{j-2}^\intercal][I - w_{j-1}t_{j-1}^\intercal X_{j-1}/(t_{j-1}^\intercal t_{j-1})] =$$

$$= [X_{j-2} - X_{j-2}w_{j-2}t_{j-2}^\intercal X_{j-2}/(t_{j-2}^\intercal t_{j-2})][I - w_{j-1}t_{j-1}^\intercal X_{j-1}/(t_{j-1}^\intercal t_{j-1})] = ...$$

$$= [X_i - X_i w_i t_i^\intercal X_i/(t_i^\intercal t_i)]...[I - w_{j-2}t_{j-2}^\intercal/(t_{j-2}^\intercal t_{j-2})][I - w_{j-1}t_{j-1}^\intercal X_{j-1}/(t_{j-1}^\intercal t_{j-1})].$$

Next, the proof of orthogonality between scores vectors follows: Let indices i and j denote now two different directions of extracted latent variables (suppose $i < j$). From the retroactive relation above, we can write [12]:

$$X_j = [X_i - X_i w_i t_i^\intercal X_i/(t_i^\intercal t_i)] * \cdots * [I - w_{j-2}t_{j-2}^\intercal/(t_{j-2}^\intercal t_{j-2})]$$
$$[I - w_{j-1}t_{j-1}^\intercal X_{j-1}/(t_{j-1}^\intercal t_{j-1})] =$$
$$= [X_i - X_i w_i t_i^\intercal X_i/(t_i^\intercal t_i)]Z$$

where Z some matrix.
Further,

$$\begin{aligned} t_i^\intercal X_j &= t_i^\intercal [X_i - X_i w_i t_i^\intercal X_i/(t_i^\intercal t_i)]Z = \\ &= t_i^\intercal X_i - (t_i^\intercal X_i w_i t_i^\intercal X_i)/(t_i^\intercal t_i) = \\ &= t_i^\intercal X_i - (t_i^\intercal t_i)t_i^\intercal X_i/(t_i^\intercal t_i) = \\ &= t_i^\intercal X_i - t_i^\intercal X_i = 0. \end{aligned} \tag{3.9}$$

Consequently,

$$t_i^\intercal t_j = t_i^\intercal X_j w_j = 0.$$

This means that scores vectors are mutually orthogonal and as a result uncorrelated, a very significant property.

When the whole process of extracting latent variables is completed, the involved t scores vectors, p loadings vectors, w and c weight vectors are combined as column-vectors and form respectively T scores matrix, P loadings matrix and weight matrices W and C.

However, frequently, for interpretation purposes, another matrix is being computed:

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

The need for R arises because derived weight scores do not refer to the original matrix X and its original variables, but to the sequential deflated matrices X_i. On the contrary, each column vector in R matrix expresses the weights of the original variables of X at the corresponding dimension [39]. Algebraically, is the generalized inverse of matrix $P^\intercal$, which is singular, and it is [39]:

$$T = XR.$$

3.2.1 Partial Least Squares Regression

Partial Least Squares Regression (PLSR) is a linear regression method that uses PLS, as a method of extracting latent variables, and a regression step. It is used when Multiple Linear Regression is impossible to give a solution to a regression problem or its solution is not stable. Typical situations include collinear predictors and/or need for dimension reduction. Then, PLS1 or PLS2 can be applied to matrices X and Y, depending on the dimension of Y. The PLS model with the selected components is then used in regression:

$$\begin{aligned} Y = XB + E = \tilde{X}B + E' &= (TP^\intercal)B + E' \\ &= T(P^\intercal B) + E' = TA + E'. \end{aligned}$$

The above regression scheme, as presented in [39], is identical to PCR.

The difference lies in the computation of scores, which takes into account the response variable(s). Regression coefficients are again computed using the Ordinary Least Squares Method.

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

In PLS1 A and E' are column vectors, while in PLS2 they are matrices, where the number of their columns is equal to the number of response variables, exactly like MLR. Matrix B with the regression coefficients that refer to the original variables of X can be computed by the fact that the inverse matrix of $P^\intercal$ is R:

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

(*Note: In NIPALS, algorithm loading vectors p are not mutually orthogonal as scores are. However, it is interesting that in [25] it has been proposed another algorithm that generates orthogonal loadings instead, and it is shown that regression coefficients are the same as those resulted by PLSR [13].*)

3.2.2 Model Selection Criteria

In the PLSR method, we are unable to define the appropriate number of PLS components that form an efficient summarization of the two-block investigated system, before the integration of the model in a regression step. This is due to the direct relation of Y matrix with the PLS components and the purpose the PLS model will eventually serve. This means that a different number of PLS components would be appropriate for regression purposes and for a classification problem, for example. The appropriate number of components to retain in the PLS model can be judged by Wold's R criterion and Adjusted Wold's R criterion.

Alternatively, both in PCR and in PLSR, AIC and R^2_{adj} are used to define the number of retaining components. These three mentioned measures are not only used to define the complexity of PCR and PLSR models but also to select the best model out of a set of tested models that an analyst produces when analyzing data. By "best model" it is meant the evaluation of the goodness of fit and parsimony of a model. In the statistical analysis of a

complicated system it is quite often to produce a set of models that differ, not only in the grade of complexity but also in the predictors they arise from. For instance, it is common to apply variable selection before PLS and PCA analysis. Different sets of original predictors are resulted, sets that then are analyzed by PCA and PLS. In this case a researcher should define not only the appropriate number of components to retain but also the appropriate set of variables, from which arises the model that performs better.

- Wold's R criterion: A commonly used measure for evaluation of plsr models is Wold's R criterion. It involves cross-validation technique as follows: Data matrix X and Y are split into k groups. A reiterative procedure is applied k times, according to which each time one (different) group is excluded from the data, and a PLSR model is generated based on rest *(k-1)* groups of the data. Then, a PRESS value (Predicted Error Sum of Squares) is computed by testing the excluded group on the generated model.

  $$PRESS = \sum (y_i - \hat{y}_{(i)})^2,$$

  where $\hat{y}_{(i)}$ the prediction with the i-group excluded. The PRESS statistic is a measure of the predictive ability of the model. In practice, in the first iteration of the algorithm when this procedure of exclusion takes place for all k groups, the produced PRESS values are added to form the $PRESS_{(1)}$ value -this is the PRESS value that corresponds to the PLSR model when the first latent variable is included. The process is repeated, however this time two latent variables are going to be used as predictors to eventually produce $PRESS_{(2)}$. The process continues adding latent variables in the PLSR model until the $min(n, p)^{th}$ latent variable is found. It should be mentioned that the residual matrices needed when adding a latent variable in PLSR model for the computation of k models, are the residual matrices as they arise from the algorithm when applied in full data matrix X. Wold's R value is computed by:

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

where number *m+1* in the numerator and m in the denominator indicates the number of latent variables included in the k individual PLSR models. The value of ratio R is compared to 1, and if it is higher than it, only first m components are included in the model. In PLS2, the number of computed Wold's R indexes is equal to the number of response variables.

- Adjusted Wold's R criterion: It is nothing but a variation of Wold's R criterion, where ratio R is compared to values 0.95 ($R_{adj}^{0.95}$) and 0.90 ($R_{adj}^{0.90}$), due to sampling variability, as it is supported in [21]. As stated in [22] *"The adjusted R criteria states that an additional latent variable will not be included in the PLS model unless it provides significantly better predictions"*.

Note: the PRESS statistic can also be used in an alternative approach to define the number of retaining components. Data are separated into k groups and PRESS is computed as the sum of individual PRESS values. However, this time residual matrices arise from the deflation of reduced data matrices that do not contain the excluded group. Then, the following ratio is computed:

$$\frac{PRESS_{(i)}}{N-i-1}$$

where i denotes the number of components in the model. In the final model only the first j latent variables are included, where j corresponds to the model with the lowest ratio.

3.2.3 Advantages and Drawbacks of PLSR

PLSR and PCR share the main philosophy as regression methods; therefore they present similarities when compared with MLR:

- As shown, utilized scores vectors in PLSR are mutually orthogonal and thus uncorrelated. As a result, they can replace correlated predictors in regression and they can effectively lead to the estimation of regression coefficients.

- Because of the mechanism that generates T matrix, the information compressed in it is directly related to response variables. This means that PLSR can also deal with noisy data.

- Even in cases of large data without collinear variables, dimension reduction achieved by PLSR can beneficially reduce model complexity.

- Dimension reduction is also helpful when graphical representations are used to get a "big picture" of the data. They allow a good sense of their structure.

- PLSR successfully deals with the "small n large p" problem, a situation that MLR cannot overcome. This situation is very common in regression analysis of biological, chemical and other scientific problems.

- Additionally, large number of predictors is also associated with a phenomenon called over-fitting. In such situations, MLR models fit perfectly the training data set, since samples are described explicitly by the predictors. However, these models fail to perform efficiently in predictions of unseen data. On the contrary, dimension reduction conducted by PLSR eliminates the danger of over-fitting.

- In the case of full rank matrix X, the PLSR model that includes as many latent variables as columns in X, gives an identical solution as the MLR model. However, in the case of correlated original predictors, as is most commonly the case, MLR regression predictors are misleading due to multicollinearity. On the contrary, PLSR regression coefficients are shrunk estimates and thus more robust, leading to better predictions [10].

- A multidimensional Y matrix is analyzed in different ways in two methods. In MLR, a linear regression model is produced for each response variable and the estimates of the regression coefficients are different in the generated models. In PLSR, in case of correlated responses, the variation PLS2 can be applied and one regression model is produced. Hence, the regression coefficients are common for all Y variables. Besides the fact that in this way the analysis is completed very fast, the relations between the response variables play significant role in the definition of regression coefficients. However in case there is no correlation between response variables, individual PLS1 models is more appropriate choice.

Between its disadvantages are:

- Regression coefficients estimated by the regression step of PLSR need extra process, so as to refer to original variables.

- The magnitude of summarized information, eventually used for predictions may depend on the user and the interpretation of methods that help choose the retaining components.

3.3 A comparison between PCA and PLS

PCA and PLS are dimension reduction techniques based on the same main idea: the aim is to construct latent variables that summarize as much as possible data information and achieve dimension reduction by using the most informative of them. Despite their similarities, methods differ in the following:

- PCA achieves its purpose with a simple one-step algorithm and produces elements, meaning scores and loadings, which refer to the original variables. On the other hand, PLS make use of iterative procedures, so scores, weights and loadings refer to sequential deflated matrices, which impeding their direct interpretation.

- Furthermore, *"in PLS dimension reduction and regression are performed simultaneously"* as referred in [43]. In contrast, the implementation of the low-dimension matrix, resulted from PCA of an X matrix, in a regression scheme is a different step.

- Technically, they differ at the optimization problem they aim to solve in order to extract these latent variables. PCA derives variables by maximizing the information of X matrix that is explained, while PLS maximizes the covariance of X and Y matrices that is explained.

- Their main difference when occupied in regression problems is that PLS involves also information in Y to model the data (supervised method), while PCA is independent from responses (unsupervised method). As stated in [15], "*there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.*"

- The next statement is directly related to this difference: *"Because PLS components are developed as latent variables possessing a high correlation with Y, the optimum number of PLS components is usually smaller than the optimum number of PCA components in PCR"*[38]. It is expected that the PLSR model will perform better, because it includes information coming from the overall system of variables that is being modeled, and not from a part of it, meaning the part of the system

that defines matrix X. This has the advantage of fewer factors to interpret and minimization of computational cost. In [43] it has been proven that PLSR models are most parsimonious and have the higher predictive accuracy.

- When multivariate Y is about to be predicted, PCR will run multiple regression steps, one for the prediction of each response variable, using the same Principal Components for all. On the contrary, PLS has a variation, PLS2, that is appropriate for dependent response variables. In this case, the same latent variables will be used for simultaneously prediction of them. The way they are generated implies that relations between Y responses are taken into account, leading to a better illustration of the whole investigated system. However, in the case of independent variables, separate PLSR models perform better, "*a single PLSR model tends to have many components and be difficult to interpret*" [41].

- Generally, PCR and PLSR result in different regression coefficients. However, when adding components, models tend to become more similar. [3]

Chapter 4

Numerical Application

The aforesaid, Partial Least Squares method and Principal Component Analysis are techniques that attain dimension reduction of high-dimensional data sets. The PCA is widely used for this purpose. It is considered to be the cornerstone of feature extraction dimension reduction techniques, and numerous of scientific papers have been published concerning its widespread utility in a variety of fields; it is characteristic the comment of Jolliffe in [17]: *"Web of Science identifies over 2000 articles published in the two years 1999–2000 that include the phrases 'principal component analysis' or 'principal components analysis' in their titles, abstracts or keywords"*. On the other hand, PLS method was first applied in econometrics, later introduced and became very popular in chemistry, and although today is used in many fields, until recently it was rather overlooked by statisticians.

In this chapter PLS method is tested as a dimensionality reduction technique. The method is applied on two different high-dimensional datasets to test its ability/prospects on both univariate and multivariate case of response variable(s). To accomplish this purpose, initial coefficients of original variables as arise from a PLSR model, are used to determine sufficient thresholds that allow to consider/judge as insignificant and non-informative the original variables with absolute value lower than the defined threshold. These variables are discarded. Furthermore, at a second level it is attempted a comparison between PCA and PLS methods, as dimension reduction methods for regression analysis purposes. The generated models are compared according to AIC and R^2_{adj} model selection criteria, while in the case of PLSR models, Adjusted Wold's R criterion in both variations was also computed, although $R^{0.90}_{adj}$ was chosen for the final results of our applications.

The numerical applications are implemented using the R programming language. The choice of programming language was determined due to the fact that R makes it feasible to process high-dimensional data sets, as those in our case. The user is allowed to apply models demanding high computational power and have supervision of what is being calculated, in contrast to a black-box software. Additionally, it is a very popular tool for applications in numerous fields and the wide community that makes use of this language shares examples and packages.

4.1 Dimension Reduction in Univariate case with PLS Method

Initially, the PLS method is applied to data coming from the gasoline dataset found in the `pls` package. This dataset consists of a sample of gasoline observations. Each observation is described by its octane number and NIR spectrum, which consists of 401 diffuse reflectance measurements from 900 to 1700 nm [19]. The aim here is to produce a flexible, yet reliable linear regression model able to predict the octane number by NIR spectrum of unseen samples. The modeling process is going to be based on 50 samples. Given that the number of available samples in the training set is small compared to the number of predictors and the existence of multicollinearity, the application of MLR with regression coefficients estimated with the OLS method would give models of poor performance, as documented in the aforementioned chapters.

Below the application of PLSR and PCR method on an initial data set of 50 samples and 401 predictors it is presented. Table 4.1 includes the values of the information criteria for these models.

models of 401 predictors					
	PLSR			**PCR**	
No.LVs	AIC	R^2_{adj}	$R^{(0.9)}_{adj}$	AIC	R^2_{adj}
1	171.9813	0.2792		180.0716	0.1526
2	18.51905	0.9671		179.3640	0.1802
3	0.3639	0.9776	3	18.0970	0.968
4	-7.1816	0.9811		6.5083	0.9751
5	-26.4575	0.9873		8.1747	0.9747

Table 4.1: Information criteria values of models generated based on 401 predictors, where No.LVs indicates the number of Latent Variables.

In the PLSR model of Table 4.1, the reduction in AIC value from the one-component to the two-component model is remarkable . The R^2_{adj} value is also significantly increased in the second case. These indications, in combination with the fact that the two-component model explains 85.58% of the variance of X matrix and 96.85% of the variance of the response variable, lead to the conclusion that a two-component PLSR model is sufficient to predict the number of octane in new data. The adjusted Wold's R ($R^{(0.9)}_{adj}$)criterion suggests a three-component model, a reasonable choice in case we wish to increase the percentage of explained variance in X data matrix to 93%.

However, in contradiction to PLSR, the two-component PCR model can be considered insufficient for the modeling of the response variable based on either AIC or R^2_{adj}. PCR suggests the use of a three-component model, so to be considered as sufficient. It is noteworthy that the three-component model of the PCR model is as effective as the two-component PLSR model -based on both displayed information criteria, which highlights the predominance and effectiveness of PLSR against PCR when it comes to regression purposes. Note that we choose to select the three-component PLSR model not only because the Wold's R criterion, specially designed for PLSR, suggests its use but also due to further significant improvement indicated by AIC and R^2_{adj}.

Next, regression coefficients of the original variables were estimated based on the PLSR model. Then, various thresholds of their absolute values were tested, so as to decide those variables that will eventually be used in the final regression model. We tested numerous models, where the design matrix X contained:

- only the variables that the absolute values of estimated regression coefficients in all models built with up to three components were larger

than the tested thresholds 0.09, 0.25, 0.30, 0.4, 0.5, 0.7, 1

- only the variables that the absolute values of estimated regression coefficients in all models built with up to two components were larger than the tested thresholds 0.09, 0.25, 0.30, 0.4, 0.5, 0.7, 1
- all variables except those with absolute values of estimated regression coefficients larger than the thresholds in all models built with up to three components and, simultaneously these absolute values were even larger in models built with 4 and 5 components
- all variables except those with absolute values of estimated regression coefficients larger than the tested thresholds in both models built with up to two components, and simultaneously these absolute values were even larger in models built with 3,4 and 5 components
- all variables except those excluded from the first case and those with absolute values of estimated regression coefficients larger than the tested thresholds in all models built with up to three components, and simultaneously these absolute values were even larger in models built with 4 and 5 components (*or simply the combination of bullets 1 and 3*)
- all variables except those excluded from the second case and those with absolute values of estimated regression coefficients larger than the tested thresholds in both models built with up to two components, and simultaneously these absolute values were even larger in models built with 3,4 and 5 components (*or simply the combination of bullets 2 and 4*).

Table 4.2 presents the information criteria values of the models that appear to have the best performance between the tested models constructed with the mentioned constraints. The specific PLSR and PCR models are based on the X matrix that included only the variables that the absolute values of the estimated regression coefficients were higher than 0.30 in all models built with up to three components. In this way, 139 variables were excluded from the predictors dataset; that is, the 34% of the initial set. The remaining models were rejected due to poor performance.

Looking at Table 4.2 we infer that in the case of PLSR all three criteria, AIC, R^2_{adj} and adjusted Wold's R demonstrate the good performance of this model and the nomination of the three-component model as most sufficient.

models of 262 predictors					
	PLSR			**PCR**	
No.LVs	AIC	R^2_{adj}	$R^{(0.9)}_{adj}$	AIC	R^2_{adj}
1	171.5231	0.2858		180.0834	0.1524
2	21.3171	0.9653		149.5523	0.5484
3	-1.7011	0.9785	3	10.1766	0.9727
4	-12.0819	0.9828		5.8030	0.9754
5	-25.3679	0.9871		6.9709	0.9753

Table 4.2: Information Criteria values of models generated based on 262 predictors.

We can see that the difference in AIC values between the one-component and the two-component models clearly shows the significant improvement in performance, something that is validated by the huge increase of the R^2_{adj} value. We choose to include the third component in the final model, retaining the advantage of data visualization in three dimensions, while further optimization in AIC and R^2_{adj} values is achieved. $R^{0.90}_{adj}$, a specialized information criteria for PLSR models, also suggests the three-component model, which explains 93.76% of the variance in the reduced X matrix and 97.98% of the variance in the response variable. It should be noted that AIC values tend to decrease as more components are added to the model. In the present case, the significant decrease between the one and the three-component model, leads to the conclusion that the last performs better. Finally, the results of the model constructed from the exclusion of the variables with absolute values of estimated regression coefficients in all models built with up to three components less than one, were very similar to those presented. However, the severe removal of the 86% of the original set of predictors, might lead to model underestimation and we decided on its rejection.

As far as the comparison with the PCR models, we can see that the PLSR models perform better in all respective cases. Additionally, in PCR method the inclusion of the third component is absolutely necessary for a sufficient performance.

4.2 Dimension Reduction in Multivariate case with PLS Method

Now, let us see how PLS method works in the multivariate case with an application on data coming from the corn data set [5]. Each sample is described by the moisture, oil, protein and starch values and measurements on an NIR spectrometer, where the wavelength range in $[1100-2498]_{nm}$, at 2nm intervals. To sum up, there are four response variables and 700 predictors. The multicollinearity phenomenon is also present here, since almost half the predictors have extreme multicollinearity. Additionally, the response variables are also correlated, as shown in Table 4.3. Hence, the specific dataset is appropriate for the application of the `PLSR2` algorithm, which in the modeling process takes into account these relations in the response variables.

	Y1	Y2	Y3	Y4
Y1	1	-0.3457	-0.3176	-0.0656
Y2	-0.3457	1	0.2853	0.0253
Y3	-0.3176	0.2853	1	-0.7983
Y4	-0.0656	0.0253	-0.7983	1

Table 4.3: Correlation between response variables

Initially, we made use of a training set consisting of 60 samples and applied PLSR2 method to the initial set of predictors. We computed the Adjusted Wold's R values, R^2_{adj} and based on them we determined the significant components for each response individually. We came to the conclusion that Y1 , Y2 , Y3 and Y4 respectively, require 5, 21, 7 and 8 components. Then we applied constraints to find the final subset of predictors, which will be used in the final regression step. For this purpose, initially we defined the following four subsets:

- In the first subset the informative variables for Y1 are included. They are the variables with absolute values of regression coefficients higher than the tested threshold in all models built with up to five components.
- In the second subset the informative variables for Y2 are included. They are the variables with absolute values of regression coefficients higher than the tested threshold in all models built with up to twenty one components.

- In the third subset the informative variables for Y3 are included. They are the variables with absolute values of regression coefficients higher than the tested threshold in all models built with up to seven components.
- In the fourth subset the informative variables for Y4 are included. They are the variables with absolute values of regression coefficients higher than the tested threshold in all models built with up to eight components.

Then, the intersection of these subsets consisted of the final reduced data set used for the regression model. The tested thresholds in this case were higher than the previous univariate case -they were 2, 2.25 and 2.50, and the excluded variables were 44, 69 and 99, respectively. The optimum threshold is found to be 2.50 and the results of the PLSR and PCR method are displayed in Tables 4.4 and 4.5.

PLSR METHOD						
No.LVs		1	2	3	4	5
AIC	Y1	0.8066	1.8514	-1.9315	-23.9521	-50.1671
	Y2	-62.6191	-62.1212	-66.1768	-68.6419	-68.0212
	Y3	76.1178	65.6200	4.9934	-30.4190	-34.8127
	Y4	153.0169	144.8910	115.1701	85.9975	63.4061
R^2_{adj}	Y1	0.5149	0.5186	0.5551	0.6964	0.8068
	Y2	0.2704	0.2759	0.3338	0.3703	0.3732
	Y3	0.1744	0.3179	0.7555	0.8666	0.8778
	Y4	-0.01514	0.1275	0.4766	0.683	0.7857
$R^{0.90}_{adj}$	Y1	0.49094	0.48029	0.51808	0.65369	0.78243
	Y2	0.2253	0.2193	0.27650	0.24688	0.24197
	Y3	0.1383	0.26207	0.66346	0.82373	0.84728
	Y4	-0.0666	0.04530	0.32932	0.58041	0.72499
No.LVs		6	7	8	21	
AIC	Y1	-140.2870	-141.7700	-166.3753	-366.0695	
	Y2	-66.8455	-87.2085	-154.2896	-214.2232	
	Y3	-54.8042	-56.3919	-60.0893	-181.0950	
	Y4	38.6846	27.8286	6.7515	-131.0660	
R^2_{adj}	Y1	0.9576	0.9592	0.9733	0.9992	
	Y2	0.3701	0.5577	0.8574	0.9543	
	Y3	0.9137	0.9172	0.9232	0.9911	
	Y4	0.8601	0.8849	0.9201	0.993	
$R^{0.90}_{adj}$	Y1	0.93693	0.94998	0.96795	0.99860	
	Y2	0.21055	0.44174	0.81050	0.8986	
	Y3	0.88788	0.89302	0.90071	0.9624	
	Y4	0.80716	0.83255	0.87608	0.9737	

Table 4.4: Information criteria values of PLSR2 models based on 601 predictors.

In the PLSR method, the regression model must include every component which is significant for at least one Y response. Thus, in our case, twenty one first components should be included in the final PLSR regression model, according to the selected information criteria. In this case, almost all variance of the reduced X matrix is used, while the percentages of the explained variance of Y responses are 99.95% for Y1, 97.06% for Y2, 99.43% for Y3 and 99.55% for Y4. It is concluded that it achieved a sizable dimensionality

PCR METHOD						
No.LVs		1	2	3	4	5
AIC	Y1	0.8279	0.6323	-8.7237	-20.9680	-115.7999
	Y2	-62.6076	-61.0262	-66.6877	-69.4438	-68.5784
	Y3	76.13027	73.8200	71.59042	49.1969	22.35033
	Y4	153.0161	150.9798	151.3416	145.2836	120.132
R^2_{adj}	Y1	0.5192	0.5283	0.6027	0.681	0.9353
	Y2	0.2703	0.2626	0.3394	0.3787	0.379
	Y3	0.1742	0.218	0.2582	0.497	0.6833
	Y4	-0.01513	0.03427	0.0435	0.1485	0.4484
No.LVs		6	7	8	21	
AIC	Y1	-114.1676	-153.1206	-151.1345	-371.4285	
	Y2	-97.8725	-116.0777	-124.317	-158.1575	
	Y3	3.2084	-24.3550	-56.3728	-112.8488	
	Y4	106.6219	76.7425	51.8404	-46.76802	
R^2_{adj}	Y1	0.9345	0.9663	0.9656	0.9992	
	Y2	0.6244	0.7267	0.765	0.8837	
	Y3	0.7731	0.8587	0.9183	0.9723	
	Y4	0.566	0.74	0.8307	0.9715	

Table 4.5: Information criteria values of PCR models based on 601 predictors

reduction, with a negligible loss of information. However, it should be mentioned that if the theoretical frame was less strict, we would choose the eight components model for even less complexity.

Further, PCR model also needs twenty one components to achieve a similar performance and to explain 99.95%, 92.51% , 98.21% , 98.17% of the variance in Y1, Y2, Y3 and Y4 correspondingly. However, the PLSR model provides better results in reference to the variability explained for all four response variables.

4.3 Final conclusions and Future work

The present Thesis approached Partial Least Squares method as a regression analysis method. However, in a more general frame, Partial Least Squares is a method that can be implemented in numerous applications and there are many variations that serve different purposes. To name some of them, it can be used for time series analysis, discriminant analysis, non-linear modelling and hierarchical modeling, for univariate and multivariate binary classification and also for survival analysis [41][3]. It effectively handles different types of data and situations that classic methods cannot overcome or lead to non-reliable results: such cases are high-dimensional, multicollinear, noisy or incomplete data.

In regression analysis problems the method counts the relationships between a set of Y variables and allows us, by using adjusted scaling, to use the prior knowledge of investigating systems in order to *"focus the model on more important Y-variables, and use experience to increase the weights of more informative X-variables"*[41]. As a result, we can investigate more complex problems with a more realistic and holistic view. This is further supported by the less strict assumptions of this method compared to classical regression, as far as noise, errors and multicollinearity. Additionally, PLSR 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 are tested.

Such an experimentation conducted in the final chapter of this Master Thesis. PLSR was used as a dimension reduction technique on two levels. Firstly, it successfully operated as a variable selection technique, as through this we removed up to 34% and 14% of the initial predictors in the final selected models in the univariate and multivariate cases, respectively. Additionally, it operated as a feature extraction method and its results were compared to those found through PCR.

In the univariate case, the final selected model is based on only 262 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 and R^2_{adj}, demonstrates that PLSR model gave more sufficient results.

In the multivariate case, the problem appears to be more complicated. Initially PLSR2 was implemented on the data out of necessity, due to the fact that correlations were observed between the response variables. We estimated the 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 procedure could be considered as a combined approach of PLS2 and PLS1, for variable selection. 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 responses. 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 of a PLSR model is optimum also in the multivariate case.

Concerning possible future expansion of this work, an application of the Elastic Net Regularization along with Partial Least Squares Method is planned to be implemented. We aim to investigate the cooperative effects of these two techniques on high-dimensional multicollinear data in order to make a projection on a low-dimensional space and thus to construct less simplex and more interpretable linear regression models of high predictive accuracy with a penalized set of predictors.

Bibliography

[1] Aggarwal, C.C. (Ed.). (2014). *Data Classification: Algorithms and Applications*. (1st ed.). Chapman and Hall/CRC.

[2] Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann.Inst. Stat. Math.* , 21, 243 – 247.

[3] Boulesteix, A.-L., and Strimmer, K. (2006). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8(1), 32–44.

[4] Cattell, R. (1966). The scree test for the number of factors, *Multivariate Behavioral Research*,1 (2), 245-76.

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

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

[7] Dunn, K. *Process Improvement Using Data*. Retrieved October 1, 2021 from: `https://learnche.org/pid/` .

[8] Fan, J., Han, F., and Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, 1(2), 293-314.

[9] 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`

[10] Frank, I. E., and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, 35(2), 109–135.

[11] Gujarati, D.N. and Porter, D.C. (2009). *Basic Econometrics*, McGraw-Hill Education (India) Pvt Limited.

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

[13] Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in Statistics - Simulation and Computation*, 17(2), 581–607.

[14] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417–441.

[15] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

[16] Jolliffe, I. T. (1972). Discarding Variables in a Principal Component Analysis. I: Artificial Data. *Journal of the Royal Statistical Society*. Series C (Applied Statistics), 21(2), 160–173.

[17] Jolliffe, I. (1986). *Principal Component Analysis*. Springer Verlag.

[18] Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, 20, 141-151

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

[20] Kim J. H. (2019). Multicollinearity and misleading statistical results. *Korean journal of anesthesiology*, 72(6), 558–569.

[21] Krzanowski, W.J.(1987). Cross-validation in principal component analysis. *Biometrics*, 43, 575 – 584.

[22] 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.

[23] Li, G.-Z., and Zeng, X.-Q. (2009). Feature Selection for Partial Least Square Based Dimension Reduction. *Foundations of Computational Intelligence*, 5, 3–37.

[24] Lindgren, F., Geladi, P., and Wold, S. (1993). The kernel algorithm for PLS. *Journal of Chemometrics*, 7(1), 45–59.

[25] Martens H, Naes T. (1989). *Multivariate Calibration*. New York: Wiley.

[26] Masaeli,M., Dy,J.G. and Fung,G. (2010). From transformation-based dimensionality reduction to feature selection. In *Proceedings of the 27th International Conference on Machine Learning*, 751–758.

[27] Ntotsis, K., Kalligeris, E. N. and Karagrigoriou, A. (2020). A Comparative Study of Multivariate Analysis Techniques for Highly Correlated Variable Identification and Management, *International Journal of Mathematical, Engineering and Management Sciences*, 5(1), 45-55.

[28] Ntotsis, K., Papamichail, M., Hatzopoulos, P. and Karagrigoriou, A. (2020). On the Modeling of Pension Expenditures in Europe, *Comm. in Stat. Case Studies-Data Analysis and Applications*, 6 (1), 50-68

[29] 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*, 187-202, iSTE Wiley and Sons.

[30] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559-572.

[31] Rosipal R., Krämer N. (2006). Overview and Recent Advances in Partial Least Squares. In: Saunders C., Grobelnik M., Gunn S., Shawe-Taylor J. (eds)*Subspace, Latent Structure and Feature Selection*. SLSFS 2005. Lecture Notes in Computer Science, vol 3940. Springer, Berlin, Heidelberg.

[32] See Wikipedia, Dimensionality Reduction, `https://en.wikipedia.org/wiki/Dimensionality_reduction` (Jan 4, 2022, 13:37, GMT)

[33] See Wikipedia, Big Data, `https://en.wikipedia.org/wiki/Big_data` (Jan 8, 2022, 16:58, GMT)

[34] See Wikipedia, Latent variable, `https://en.wikipedia.org/wiki/Latent_variable` (Jan 14, 2022, 12:53, GMT)

[35] See Wikipedia, Covariance,
`https://en.wikipedia.org/wiki/Covariance` (Jan 11, 2022, 14:28, GMT)

[36] Swarthmore College, Department of Engineering,
`http://lpsa.swarthmore.edu/MtrxVibe/EigMat/MatrixEigen.html`

[37] Trygg, J., and Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3), 119–128.

[38] Varmuza, K. and Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics* (1st ed.). CRC Press.

[39] Wehrens, H.R.M.J. (2011). *Chemometrics with R: multivariate data analysis in the natural sciences and life sciences*. New York, NY: Springer

[40] Wold, S., Martens, H., and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. *Lecture Notes in Mathematics*, 286–293.

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

[42] Wold, H. (1975). Path models with latent variables: The NIPALS approach. In: H.M. Blalock et al., editor, Quantitative Sociology: International perspectives on mathematical and statistical model building, pages 307–357. Academic Press.

[43] Yeniay, Ö. and Goktas, A. (2002). A comparison of partial least squares regression with other prediction methods. *Hacettepe Journal of Mathematics and Statistics*. 31.