UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

TRABAJO FIN DE MÁSTER

PLS Regression for Multivariate and Functional Data

Máster Universitario en Ciencia de Datos

Autor: David del Val Moncholí

Codirector: Alberto Suárez González
Departamento de Informática

Codirector: José Ramón Berrendero Díaz
Departamento de Matemáticas

Agosto, 2024

PLS Regression for Multivariate and Functional Data

Author: David del Val Moncholí
Co-advisor: Alberto Suárez González
Co-advisor: José Ramón Berrendero Díaz

Departamento de Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid

August 2024

Resumen

Mínimos cuadrados parciales (PLS) es una familia de técnicas de reducción de dimensionalidad formuladas en el contexto de problemas de regresión, que aprovechan las dependencias lineales entre las variables predictoras y las variables a predecir. En concreto, en PLS, los componentes se determinan proyectando en direcciones a lo largo de las cuales se maximiza la covarianza cruzada entre proyecciones dentro de cada uno de los espacios correspondientes a ambos grupos de variables. De este modo, PLS combina el criterio de optimización del análisis de componentes principales (PCA), que consiste en maximizar la varianza a lo largo de las direcciones en el espacio de las variables predictoras, y maximización de la correlación entre tales proyecciones y combinaciones lineales de las variables a predecir. Estos componentes extraídos por PLS pueden ser utilizados para formular modelos más simples y, en algunos casos, más precisos que los que se construyen a partir de las observaciones originales.

En este trabajo, se presenta una formulación general de PLS aplicado a problemas de regresión con respuesta escalar, suponiendo únicamente que las variables predictoras son elementos de un espacio de Hilbert. Además del producto interno estándar (euclídeo), este espacio está dotado de un producto interno generalizado: el producto conjugado, definido bajo la métrica inducida por la inversa del operador de covarianza de las variables predictoras. PLS es un proceso iterativo cuyo objetivo es identificar una secuencia de subespacios anidados de dimensión creciente. Estos subespacios están generados por un conjunto de elementos del espacio de Hilbert, que forman una base no necesariamente ortogonal. En cada iteración, la base de PLS se amplía incorporando el elemento del espacio de Hilbert que maximiza la covarianza con las variables a predecir, bajo ciertas restricciones. Dependiendo del tipo de restricciones consideradas, se pueden identificar diferentes bases de PLS que generan el mismo subespacio. Si se impone ortogonalidad con los anteriores elementos de la base, se obtiene la base ortogonal de PLS, que es la que se construye en el algoritmo NIPALS. La base conjugada se obtiene imponiendo una relación de conjugación definida en términos del producto interno generalizado. Esta base puede construirse utilizando el algoritmo de gradientes conjugados. Tanto la base ortogonal como la conjugada generan una secuencia de subespacios de Krylov definidos en términos del operator de covarianza de las variables predictoras y la covarianza cruzada entre las variables predictoras y las variables a predecir. Esto permite identificar una tercera base de PLS: la base de Krylov, que contiene los elementos obtenidos al aplicar repetidamente el operador de covarianza de los regresores sobre la covarianza cruzada. La generalidad de esta formulación permite aplicar PLS no solo a datos multivariados y funcionales, que residen naturalmente en espacios euclídeos, sino también a objetos matemáticos más complejos, como grafos o textos, estableciendo una cor-

respondencia entre dichos objetos (por ejemplo, a través de *kernel embeddings*) y elementos de un espacio de Hilbert.

A partir de la conexión con gradientes conjugados, es posible analizar la convergencia de PLS a mínimos cuadrados ordinarios (OLS) en problemas de regresión multilineal con predictores multivariantes y respuesta escalar. En concreto, es posible derivar un límite superior para las diferencias entre los coeficientes de regresión calculados mediante PLS y mediante OLS en función del número de componentes considerados en PLS. Este límite depende únicamente de la distribución de los autovalores de la matriz de covarianza de las variables predictoras. Cuando el número de componentes es igual al número de autovalores distintos de esta matriz de covarianza, el coeficiente de regresión de PLS coincide con el calculado usando OLS. En la práctica, si los valores propios están agrupados en clústeres, PLS proporciona una aproximación precisa al coeficiente de regresión de OLS cuando el número de componentes considerados es igual al número de clústeres presentes en el espectro de la matriz de covarianza de los regresores.

Finalmente, se llevan a cabo una serie de experimentos en conjuntos de datos reales para evaluar el rendimiento de PLS como método de reducción de dimensionalidad, especialmente en comparación con PCA. Se consideran tanto conjuntos de datos multivariantes como funcionales. En los problemas analizados, asumiendo un modelo de regresión lineal, PLS es más eficaz que PCA cuando se utilizan pocos componentes. Las diferencias disminuyen a medida que aumenta el número de componentes considerados. Asimismo, se han realizado experimentos adicionales en los que se PCA y PLS se emplean como preprocesamiento previo a la aplicación de predictores más generales. En concreto, los primeros componentes que resultan del análisis se usan como variables de entrada de regresores no lineales como, por ejemplo, máquinas de vector soporte, redes neuronales y bosques aleatorios. Los resultados de esta evaluación empírica muestran que PLS puede ser un método eficaz de reducción de dimensionalidad en problemas del mundo real, incluso cuando las dependencias entre las variables predictoras y las variables a predecir son no lineales.

Palabras clave

Análisis de datos funcionales, reducción de dimensionalidad, regresión, mínimos cuadrados parciales

Abstract

Partial least squares (PLS) is a family of dimensionality reduction techniques formulated in the context of regression problems that take advantage of linear dependencies between the predictor and target variables. Specifically, the PLS components are determined by projecting onto directions along which the cross-covariance between projections within the spaces of the predictor and of the target variables is maximized. In doing so, PLS combines the optimization criterion of principal component analysis (PCA), which consists in maximizing the variance along directions within the space of predictor variables, and the maximization of the correlation of these projections with linear combinations of the target variables. The components extracted by PLS can then be utilized to formulate models that are simpler and, in some cases, more accurate than those based on the original observations.

In this work, a general formulation of PLS is made for regresssion problems with scalar response, assuming that the predictor variables are elements of a Hilbert space. Besides the standard (Euclidean) inner product, this space is endowed with a generalized, conjugate inner product defined under the metric induced by the inverse of the covariance operator of the predictor variables. PLS is an iterative process whose goal is to identify a sequence of subspaces of increasing dimension. These subspaces are the linear span of a set of elements in the Hilbert space that form a basis, which is not necessarily orthogonal. At each iteration, the PLS basis is enlarged by incorporating the element of the Hilbert space for which the covariance with the target variable is maximized, subject to some constraints. Depending on the types of constraints considered, different PLS bases that span the same subspace can be identified. If orthogonality with the previous basis elements is enforced, one obtains the orthogonal PLS basis computed in the NIPALS algorithm. The conjugate basis is obtained by imposing a conjugacy relation defined in terms of the generalized inner product. This basis can be constructed using the conjugate gradients algorithm. It is shown that both the orthogonal and the conjugate bases span a sequence of Krylov subspaces defined in terms of the covariances of the predictor variables and the covariances between the predictor and target variables. This allows the identification of a third PLS basis: the Krylov basis, which contains the elements obtained by repeatedly applying the regressor covariance operator onto the cross-covariance. The generality of the formulation makes it possible to apply PLS not only to multivariate and functional data, which naturally reside in Euclidean spaces, but also to more complex mathematical objects, such as graphs or texts, by mapping them (e.g., through kernel embeddings) onto elements of a Hilbert space.

Based on the connection with conjugate gradients, it is possible to analyze the convergence of PLS to ordinary least squares (OLS) in multilinear regression problems with

multivariate predictors and scalar response. In particular, it is possible to derive an upper bound on the difference between the PLS and OLS regression coefficients as a function of the number of components considered in PLS. This bound depends only on the distribution of the eigenvalues of the covariance matrix of the predictor variables. When the number of components is equal to the number of distinct eigenvalues of this covariance matrix, the PLS regression coefficient coincides with the one computed using OLS. In practice, if the eigenvalues are grouped in clusters, PLS provides an accurate approximation to the OLS regression coefficient when the number of components considered equals the number of clusters in the spectrum of the regressors' covariance matrix.

Finally, a series of experiments on real-world datasets are carried out to assess the performance of PLS as a dimensionality reduction method, especially in comparison with PCA. Both multivariate and functional datasets are considered. In the problems analyzed, assuming a linear regression model, PLS is more effective than PCA when few components are used, while the differences become smaller as the number of components considered increases. Additional experiments are carried out in which PCA and PLS are used as a preprocessing step in combination with more general predictors. Specifically, the first components that result from the analysis are used as inputs of non-linear regressors, such as support vector machines, neural networks and random forests. The results of this empirical evaluation show that PLS can be an effective dimensionality reduction method in real-world problems even when the dependencies between the predictor and the target variables are non-linear.

Keywords

Functional data analysis, dimensionality reduction, regression, partial least squares.

Contents

1 Introduction **1**

2 Application of PLS to regression problems **3**

2.1 Principal component regression 6

2.2 Partial least squares 7

2.2.1 The orthogonal PLS basis 7

2.2.2 PLS as a constrained least squares problem 11

2.2.3 The conjugate PLS basis 15

2.3 Numerical algorithms for PLS regression 22

2.3.1 NIPALS 22

2.3.2 Conjugate Gradients 28

3 Multivariate Regression: Relationship between PLS and OLS regression **33**

3.1 Partial least squares on a sample 33

3.2 Partial least squares regression on a sample 37

3.3 Relation between partial least squares and ordinary least squares . . . 40

3.4 Empirical study 46

3.4.1 Synthetic data 46

3.4.2 The Californian Housing dataset 51

4 Empirical comparison of PCA and PLS Regression **53**

4.1 Multiple regression 53

4.1.1 Results 57

4.2 Functional regression with scalar response 65

4.2.1 Results 67

5 Conclusions and future work **73**

References **75**

Chapter 1

Introduction

Partial least squares (PLS) is a family of dimensionality reduction methods introduced in the field of chemometrics (Noonan & Wold, 1977; Wold, Ruhe, Wold, & Dunn, 1984; Helland, 1990; Wold, Sjöström, & Eriksson, 2001; Abdi, 2010), where it is extensively used. Its success in this discipline has led to its adoption in other scientific areas, such as medicine (Nguyen & Rocke, 2002; Zhang, Han, & Deng, 2017), ecology (Burnett et al., 2021), oceanography (Okwuashi, Ndehedehe, & Attai, 2020), and neuroscience (Krishnan, Williams, McIntosh, & Abdi, 2011; Nakua et al., 2024), among others (Mehmood & Ahmed, 2016). PLS was originally formulated as a dimensionality reduction method for multivariate regression, assuming a linear relation between the predictor and the response variables, in a sample of independent observations (Cook, Forzani, & Liu, 2023). Since its introduction, it has been extended to deal with classification problems (Ståhle & Wold, 1987; Barker & Rayens, 2003; Moindjié, Dabo-Niang, & Preda, 2023), and with dependent observations (Wang, Gu, Wang, & Saporta, 2019). Moreover, even if it was originally formulated in the context of linear regression, it has been shown to be effective for dimensionality reduction also when nonlinear relations are present (Cook & Forzani, 2021). PLS is most effective in problems when many predictor variables contribute information about the response (Cook & Forzani, 2018, 2019).

One of the earliest sources for PLS is Noonan and Wold (1977). In that work, the PLS components are defined computationally as the result of applying the NIPALS (nonlinear iterative partial least squares) algorithm. For scalar Y, the components identified by NIPALS are the solution of a constrained optimization problem. This problem consists in finding orthogonal linear combinations of the coordinates of X that maximize the covariance with the response variable (de Jong, 1993). Utilizing this optimization problem, it is also possible to show that the conjugate gradient method (Wold et al., 1984) and the Lanczos bidiagonzalization algorithm (Eldén, 2004) can be used as alternatives to NIPALS for PLS regression.

The extension of PLS to functional data was introduced in Preda and Saporta (2005). In this work, NIPALS was adapted to consider the case in which the predictors are functions that depend on a continuous parameter, such as time or space. Then, in Delaigle and Hall (2012), functional PLS was presented as a constrained optimization problem, analogous to multivariate PLS. More recent advances in the field have compared PLS and principal com-

ponent regression (PCR) with functional regressors (Febrero-Bande, Galeano, & González-Manteiga, 2017), and presented a formulation of functional PLS based on the conjugate gradient method (Babii, Carrasco, & Tsafack, 2024).

The main goal of this work is to provide a general formulation that provides a framework to understand the relation of these variants of PLS. This formulation assumes only that the regressor variables can be characterized as elements of a Hilbert space. Regression with multivariate and functional predictors are particular cases of this formulation. In the multivariate setting, the space of regressor variables is typically a subspace of R^D, where D is the number of regressors. In the functional case, the regressors are functions in a subspace of L^2, the space of square-integrable functions. Chapter 2 introduces PLS as an iterative process whose goal is to identify a sequence of nested subspaces of increasing dimensions. In turn, these subspaces are generated by bases obtained iteratively by PLS. At each iteration, PLS includes in the basis the direction in the Hilbert space that maximizes the covariance with the target, subject to some constraints. Different constraints lead to the construction of different bases that span the same spaces. The orthogonal basis calculated in NIPALS is obtained when orthogonality with respect to the usual inner product is imposed. The conjugate basis built in the conjugate gradient method is the result of enforcing orthogonality with respect to the conjugate inner product associated to the inverse of the covariance operator of the regressors.

Chapter 3 focuses on the analysis of PLS in multiple regression. By applying general properties of the conjugate gradient (Hestenes & Stiefel, 1952; Nocedal & Wright, 1999), PLS can be proven to be equivalent to a polynomial fitting problem, as shown in Blazère, Gamboa, and Loubes (2014). From this reformulation, it is possible to derive an upper bound for the distance between the PLS and ordinary least squares (OLS) approximations to the regression coefficient that depends only on the spectrum of the regressor covariance operator. In light of this analysis, we explore the relation between these estimations in terms of the characteristics of the distribution of eigenvalues. In particular, if these eigenvalues are grouped into k tight clusters, PLS with k components provides a good approximation to OLS.

Finally, in Chapter 4, the results of an empirical evaluation of PLS in real-world regression problems are presented, analyzing both multilinear and functional problems. In this study, PLS and PCA are employed for dimensionality reduction. The components extracted by these methods are then used as inputs to both linear and non-linear predictors, such as support vector machines, random forests, and neural networks. From the analysis of these results, one observes that the predictive capacity of the first PLS components is larger than that of the PCA components. This result is to be expected since PLS components not only incorporate information about the variance of the regressor variables, as in PCA, but also consider the correlation between the regressor and target variables.

Chapter 2

Application of PLS to regression problems

This chapter is devoted to the application of partial least squares to regression problems with scalar response. To enable the unified treatment of multivariate and functional data, the sole assumption is that the regressors are elements of a Hilbert space $\mathcal{X}$. Specifically, when multivariate regressors are considered, $\mathcal{X}$ typically is a subspace of $\mathbb{R}^D$. For functional regressors, a common assumption is that $\mathcal{X}$ is a subspace of $L^2[0,T]$, the space of square-integrable real-valued functions defined in $[0,T]$. In either case, the inner product in $\mathcal{X}$ will be denoted as $\langle\cdot,\cdot\rangle$. Besides random vectors and random functions, this formulation of PLS is valid also when the predictor variables are more complex mathematical objects, such as graph or texts. In such cases, to apply PLS, it is sufficient to define a correspondence between these types of objects and elements of a Hilbert space using, for example, kernel embeddings.

Given the predictor $X \in \mathcal{X}$ and the response variable Y in $\mathbb{R}$, the linear regression model is of the form

$$Y = a^* + \langle \beta^*, X \rangle + \epsilon, \tag{2.1}$$

where $a^* \in \mathbb{R}$ is the intercept, $\beta^* \in \mathcal{X}$ is the regression coefficient, and the noise term $\epsilon \in \mathbb{R}$ is a random variable such that $\mathbb{E}(\epsilon|X) = 0$. This implies $\mathbb{E}(\epsilon) = 0$ and $\mathbb{E}(X\epsilon) = 0$, which means that the noise is uncorrelated with the predictor.

Assuming that a^*, β^* are known, the optimal prediction for the response variable is the regression function

$$\hat{Y} = \mathbb{E}\left[Y|X\right] = a^* + \langle \beta^*, X \rangle .$$

Additionally, from (2.1) and applying $\mathbb{E}\left[X\epsilon\right] = 0$, one can obtain an equation for the regression coefficient:

$$\mathrm{Cov}\left(X,Y\right) = \mathrm{Cov}\left(X, a^* + \langle \beta^*, X \rangle + \epsilon\right) = \mathrm{Cov}\left(X, \langle \beta^*, X \rangle\right) = \mathcal{K}\beta^* \implies \gamma = \mathcal{K}\beta^*, \tag{2.2}$$

where $\mathcal{K} = \mathrm{Cov}(X,X)$ and $\gamma = \mathrm{Cov}\left(X,Y\right)$ are the regressor covariance operator and the cross-covariance, respectively.

Since $\mathcal{K}$ is a covariance operator, it is symmetric and positive definite. Furthermore, we will assume that it is continuous. Under these circumstances, Mercer's theorem (Mercer & Forsyth, 1997; Ghojogh, Ghodsi, Karray, & Crowley, 2021), states that there is an orthonormal basis of $\mathcal{X}$ consisting of eigenvectors (or eigenfunctions) of $\mathcal{K}$.

However, if some of the eigenvalues are zero, the null space of the operator will be greater than zero. As a result, one can find some $\beta_0 \in \mathcal{X}$, $\beta_0 \neq 0$ such that $\mathcal{K}\beta_0 = 0$. In this case, (2.2) does not uniquely identify β^*, since $\mathcal{K}(\beta^* + \beta_0) = \mathcal{K}\beta^* + 0 = \gamma$, while $(\beta^* + \beta_0) \neq \beta^*$. This issue can be dealt with by restricting the search of β^*, to the space generated by eigenvectors (or eigenfunctions) associated to non-zero eigenvalues.

In the cases considered in this work, the regressor variables are either random vectors or random functions. If X is a random vector in $\mathcal{X}$, $X = (X_1, \ldots, X_D)^\top$, $\mathcal{K} \in \mathbb{R}^{D\times D}$ is the covariance matrix, $\gamma \in \mathbb{R}^D$ is the vector $(\mathrm{Cov}\,(X_1 Y), \ldots, \mathrm{Cov}\,(X_D Y))^\top$, $\beta^* = (\beta_1^*, \ldots, \beta_D^*)^\top \in \mathbb{R}^D$, and the inner product is given by

$$\langle \beta^*, X \rangle = \sum_{d=1}^{D} \beta_d^* X_d.$$

If X is a random function in $L^2[0,T]$, $\mathcal{K}$ is the covariance operator associated to the covariance function $k(s,t) = \mathrm{Cov}(X(t), X(s))$:

$$\begin{aligned} \mathcal{K} : L^2[0,T] &\longrightarrow L^2[0,T] \\ f &\longrightarrow (\mathcal{K}f)(t) = \int_0^T k(t,s) f(s) ds, \end{aligned}$$

γ is the function $\gamma(t) = \mathrm{Cov}\,(X(t)Y)$, $t \in [0,T]$, and the inner product is given by

$$\langle \beta^*, X \rangle = \int_0^T \beta^*(t) X(t) dt. \tag{2.3}$$

More details on the properties of these quantities in the functional case can be found in Preda and Saporta (2005). Additionally, a summary of the described quantities in both cases is included in Table 2.1.

	Multivariate	**Functional**
$\mathcal{X}$	$\mathbb{R}^D$	$L^2[0,T]$
$\langle a, b \rangle$	$a^\top b$	$\int_0^T a(s)b(s)ds$
$\mathcal{K}$	$\mathrm{Cov}(XX^\top)$	$\mathcal{K}f(t) = \int_0^T k(s,t)f(s)ds$
γ	$\mathrm{Cov}(XY)$	$\gamma(t) = \mathrm{Cov}(X(t), Y)$

Table 2.1: Multivariate and functional notation

Another aspect to keep in mind when functional regressors are considered, is that each element in L^2 is not a unique function, but an equivalence class of functions: Two functions in L^2 are equivalent if they differ only in a zero-measure set. In particular, due to the integral in (2.3), changes to β^* in a zero-measure set do not alter the result of the inner product.

Note that, in most cases, functional data is observed in a grid of M points $\mathbf{s} = (s_1, \ldots, s_M)$. We will assume that the grid is fine enough for the functional characteristics of X to be apparent. By evaluating in the grid, one gets a vector $X(\mathbf{s}) = (X(s_1), X(s_2), \ldots, X(s_M))^\top$. Moreover, if the coefficient is discretized in the same grid, the inner product is calculated as

$$\langle \beta^*, X \rangle = \int_0^T \beta^*(s) X(s) ds = \sum_{m=1}^{M} a_m \beta^*(s_m) X(s_m) = X(\mathbf{s})^\top \mathbf{A} \beta^*(\mathbf{s}),$$

where $\mathbf{A} = \operatorname{diag}(a_1, \ldots, a_M)$ contains the integration weights. Furthermore, since X is usually assumed to be smooth, the variables in $X(\mathbf{s})$ are not independent, and collinearity issues can arise. As a result, considering the multivariate regression problem on the discretized data can be troublesome, and the functional nature of the data must be taken into account.

Without loss of generality, in what follows, we assume that both the predictor and the response variables are centered with the mean, which implies that $a^* = 0$. Under this assumption, the regression model becomes

$$Y = \langle \beta^*, X \rangle + \epsilon,$$

with $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, and $\mathbb{E}[\epsilon | X] = 0$.

One approach to compute β^* is by applying dimensionality reduction techniques. These methods consider projections of the original data onto a low-dimensionality space. As a result, they can improve computational efficiency, reduce the noise in the input data, and lead to more interpretable models. Dimensionality reduction is particularly useful to address collinearity issues, making it an essential tool in the functional setting. Additionally, when dealing with functional data the covariance operator $\mathcal{K}$ is not invertible, as zero is an accumulation point of its eigenvalues (Cuevas, 2014), complicating the direct estimation of β^*.

The projection onto low dimensionality spaces could be performed by using a standard basis (e.g., polynomial or Fourier basis). However, there is no guarantee that the projections onto those bases will preserve the relevant information. To retain the meaningful aspects of the original data, a finite basis $U = \{u_1, \ldots, u_L\}$, $u_\ell \in \mathcal{X}$, and $L \leq D$, the dimensionality of $\mathcal{X}$, can be defined following some criterion. For instance, maximizing some measure that is expected to be relevant for prediction. Then, a simplified regression model can be considered: $Y = \sum_{\ell=1}^{L} \langle X, u_\ell \rangle \beta_\ell + \epsilon_L$, where we seek to predict Y based only on the projections of the original data onto the subspace spanned by this basis. Note that this restricted model is equivalent to assuming that the regression coefficient must be contained in the space generated by the basis. Therefore, the reconstruction of Y using the projections onto U can be defined as

$$Y^{(U)} = \sum_{\ell=1}^{L} \langle X, u_\ell \rangle \beta_\ell = \left\langle X, \beta^{(U)} \right\rangle, \quad \beta^{(U)} = \sum_{\ell=1}^{L} u_\ell \beta_\ell, \quad \beta^{(U)} \in \operatorname{span}\{u_1, \ldots, u_L\}. \tag{2.4}$$

It is worth noting that, with this formulation, the previous expressions hold both in the multivariate, and the functional setting, using the correspondences detailed in Table 2.1. In the remaining of this chapter, we will utilize this notation extensively, to introduce PLS both

for multivariate and functional regressors. As a first step, in the next section, we describe principal component analysis, which will be used as a reference point in the discussion of PLS that follows.

2.1 Principal component regression

Principal component analysis (PCA) is a widely-used dimensionality reduction method in multivariate analysis (e.g. Abdi & Williams, 2010). Moreover, it has been extensively applied in functional data analysis (Cardot, Ferraty, & Sarda, 1999; Hall & Horowitz, 2007; Ramsay & Silverman, 2013). The principal component basis $\{\psi\}_{\ell\geq 1}$ is defined in terms of the solutions of the eigenvalue equation

$$\mathcal{K}\psi_\ell = \lambda_\ell \psi_\ell, \quad \ell \geq 1,$$

subject to $\langle \psi_i, \psi_j \rangle = \delta_{ij}$, with $i, j \geq 1$, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$.

Equivalently, the elements of the basis can be computed iteratively by identifying a sequence of orthogonal directions along which the variance is maximized. Specifically, the ℓ-th element in the basis can be obtained by solving the following optimization problem

$$\psi_\ell = \underset{\psi\in\mathcal{X}}{\operatorname{argmax}} \ \operatorname{Var}\left(\langle \psi, X\rangle\right) \quad \text{subject to} \quad \begin{aligned} &\langle \psi, \psi\rangle = 1; \\ &\langle \psi, \psi_i\rangle = 0 \quad i = 1, \ldots, \ell - 1. \end{aligned}$$

In principal component regression (PCR), the regression coefficient is represented as an expansion in the PCA basis

$$\beta^* = \sum_{\ell\geq 1} b_\ell^{(\mathrm{PCR})} \psi_\ell, \quad b_\ell^{(\mathrm{PCR})} = \frac{1}{\lambda_\ell} \langle \psi_\ell, \gamma \rangle . \tag{2.5}$$

Therefore, a natural approximation of β^* can be obtained by truncating this series. If only the first L components are taken into account, we obtain the approximation

$$\beta_L^{(\mathrm{PCR})} = \sum_{\ell=0}^{L} b_\ell^{(\mathrm{PCR})} \psi_\ell. \tag{2.6}$$

Thus, by rewriting (2.4) for the particular case of the basis composed of the first L principal components, we obtain

$$Y_L^{(\mathrm{PCR})} = \left\langle \beta_L^{(\mathrm{PCR})}, X \right\rangle = \sum_{\ell=1}^{L} b_j^{(\mathrm{PCR})} \langle \psi_\ell, X \rangle . \tag{2.7}$$

Note that the PCA basis is defined solely in terms of the covariance structure of X, without taking into account the relation between the predictor and the target variables. Therefore, the first L elements of the PCA basis, which capture the largest part of the

variance of X, need not be the ones that have the largest predictive capacity. A possible way to address this shortcoming is to include in the linear model the principal components whose correlation with the dependent variable is largest Aguilera, Ocaña, and Valderrama (1997); Febrero-Bande et al. (2017). As an alternative, the components can be identified by optimizing a combination of the variance within the space of predictor variables and the correlation between predictor and target variables. This is the strategy adopted in partial least squares (PLS), which is described in the next section.

2.2 Partial least squares

The goal of partial least squares (PLS) is to identify directions along which the covariance between X and Y is maximized. As discussed in Rosipal and Krämer (2005), the PLS maximization criteria can be better understood by rewriting the square of the covariance in terms of the square of the correlation and the variances of each of the variables

$$\operatorname{Cov}^2\left(\langle \phi, X\rangle, Y\right) = \operatorname{Var}\left(\langle \phi, X\rangle\right) \operatorname{Corr}^2\left(\langle \phi, X\rangle, Y\right) \operatorname{Var}\left(Y\right).$$

From this expansion, it is apparent that in PLS one attempts to simultaneously maximize the variance captured by the projection (as in PCA) and the correlation with the target variable. Therefore, the PLS basis not only reflects the covariance structure of X, but also considers the predictive capacity of the components identified.

A PLS basis is built iteratively: The first element in a PLS basis is such that the covariance between the projection of X along the corresponding direction and Y is maximized. The ℓ-th element in the basis is obtained by maximizing the covariance of the projection of X along the corresponding direction and $Y - Y_{\ell-1}^{(\mathrm{PLS})}$. Following the same conventions as in PCR, $Y_{\ell-1}^{(\mathrm{PLS})}$ represents the approximation of the target variable Y based on the projections onto the first $\ell - 1$ basis elements identified.

However, different variants of PLS can be found in the literature (see e.g., de Jong (1993), Wegelin (2000), and Ergon (2009)). In the scalar response setting, these variants are equivalent as they all can be proved to be equivalent to formulating an optimization problem restricted to a sequence of Krylov subspaces of increasing order, which we will describe in Section 2.2.2. The main differences between these formulations reside in the restrictions placed on the basis. The orthogonal basis, described in Section 2.2.1, is obtained by enforcing pair-wise orthogonality of the basis elements with respect to the inner product of the Hilbert space $\mathcal{X}$. The conjugate basis, introduced in Section 2.2.3, is obtained if one enforces orthogonality with respect to the generalized (conjugate) inner product defined in terms of the inverse of the covariance operator of the regressor variables.

2.2.1 The orthogonal PLS basis

In this section, the PLS basis of pair-wise orthogonal elements is presented, and its properties are explored. As discussed earlier, the PLS components are optimized by maximizing the covariance of a linear combination of the regressor variables and the target variable.

Therefore, to obtain a PLS basis with L elements, one can consider the following sequence of optimization problems:

$$\phi_\ell = \underset{\phi\in\mathcal{X}}{\operatorname{argmax}} \operatorname{Cov}\left(\langle X,\phi\rangle, Y - Y_{\ell-1}^{(\phi)}\right) \quad \text{subject to} \quad \langle\phi,\phi\rangle = 1, \tag{2.8}$$

for $\ell = 1,\dots,L$; where $Y_{\ell-1}^{(\phi)}$ is the optimal prediction of Y from the projections of X onto the first $\ell-1$ elements of the basis according to the least squares criterion. In terms of the elements of this basis, the optimal prediction is

$$Y_\ell^{(\phi)} = \sum_{i=1}^{\ell} b_i^{(\phi,\ell)} \langle X,\phi_i\rangle, \tag{2.9}$$

where the coefficients $\{b_i^{(\phi,\ell)}\}_{i=1}^{\ell}$ minimize the expected value of $\left(Y - Y_{\ell-1}^{(\phi)}\right)^2$:

$$(b_1^{(\phi,\ell)},\dots,b_\ell^{(\phi,\ell)}) = \underset{(b_1,\dots,b_\ell)\in\mathbb{R}^\ell}{\operatorname{argmin}} \mathbb{E}\left[\left(Y - \sum_{i=1}^{\ell} b_i \langle X,\phi_i\rangle\right)^2\right]. \tag{2.10}$$

The iterative process halts when no additional information can be extracted from the covariance. The maximum number of basis elements $L_{\max}$ is defined by the following conditions:

$$\max_{\phi\in\mathcal{X}} \operatorname{Cov}\left(\langle X,\phi\rangle, Y - Y_\ell^{(\phi)}\right) \neq 0,\ \ell < L_{\max} \quad \text{and} \quad \max_{\phi\in\mathcal{X}} \operatorname{Cov}\left(\langle X,\phi_\ell\rangle, Y - Y_{L_{\max}}^{(\phi)}\right) = 0. \tag{2.11}$$

In the following, we will always assume that $L \leq L_{\max}$. With this consideration in mind, we can seek an expression for the elements of the basis defined by (2.8).

Proposition 2.2.1. *The solution of the optimization problem* (2.8) *is*

$$\phi_\ell = \frac{1}{\lambda^{(\ell)}}\left(\gamma - \sum_{i=1}^{\ell-1} b_i^{(\phi,\ell-1)} \mathcal{K}\phi_i\right), \tag{2.12}$$

where $\lambda^{(\ell)}$ *is a normalization constant and* $b_i^{(\phi,\ell-1)}$ *are defined in* (2.10).

Proof. The quantity maximized in (2.8) can be rewritten as

$$\begin{aligned}\operatorname{Cov}\left(\langle X,\phi\rangle, Y - Y_{\ell-1}^{(\phi)}\right) &= \mathbb{E}\left(Y\langle X,\phi\rangle\right) - \mathbb{E}\left(\langle X,\phi\rangle Y_{\ell-1}^{(\phi)}\right) = \\ &= \langle \mathbb{E}(XY),\phi\rangle - \mathbb{E}\left(\langle X,\phi\rangle \sum_{i=1}^{\ell-1} b_i^{(\phi,\ell-1)}\langle X,\phi_i\rangle\right) = \\ &= \langle\gamma,\phi\rangle - \sum_{i=1}^{\ell-1} b_i^{(\phi,\ell-1)}\langle\phi,\mathcal{K}\phi_i\rangle.\end{aligned}$$

To find the element $\phi\in\mathcal{X}$ that maximizes this quantity, we consider the Lagrangian:

$$\mathcal{L}(\phi) = \langle\gamma,\phi\rangle - \sum_{i=1}^{\ell-1} b_i^{(\phi,\ell-1)}\langle\phi,\mathcal{K}\phi_i\rangle - \lambda(\langle\phi,\phi\rangle - 1).$$

Finally, the maximum is obtained by finding the zeroes of the derivative:

$$\frac{\partial \mathcal{L}}{\partial \phi} = \gamma - \sum_{i=1}^{\ell-1} b_i^{(\phi,\ell-1)} \mathcal{K}\phi_i - \lambda\phi = 0 \implies \phi_\ell = \frac{1}{\lambda^{(\ell)}} \left(\gamma - \sum_{i=1}^{\ell-1} b_i^{(\phi,\ell-1)} \mathcal{K}\phi_i \right),$$

where $\lambda^{(\ell)}$ is the normalization constant, and can be calculated as

$$\lambda^{(\ell)} = \left\| \gamma - \sum_{i=1}^{\ell-1} b_i^{(\phi,\ell-1)} \mathcal{K}\phi_i \right\|^2,$$

where $\|\cdot\|$ is the norm induced by the inner product of $\mathcal{X}$. □

Even though in the formulation of PLS given by (2.8) an orthogonality constrain is not considered explicitly, the following proposition shows how this constraint is implicitly enforced. As evidenced in the proof, the orthogonality is a result of the choice of the normalization, along with the requirements that the coefficients in (2.10) satisfy a least squares problem, and that the optimal prediction based on the components identified in previous steps is subtracted from the target variable.

Proposition 2.2.2. *The basis functions $\{\phi_\ell\}_{\ell=1}^{L}$ are pair-wise orthogonal.*

Proof. To simplify the notation in this proof, consider the following definitions:

$$J_\ell(\phi) = \operatorname{Cov}\left(\langle X, \phi \rangle, Y - Y_{\ell-1}^{(\phi)} \right) \quad \text{and} \quad \mathcal{F}_{\ell-1} = \operatorname{span}\left\{ \phi_1, \ldots, \phi_{\ell-1} \right\}.$$

Let us recall that the expected value defines an inner product as $\langle A, B \rangle_{\mathbb{E}} = \mathbb{E}(AB)$, where A, B are scalar random variables. With this property in mind, we will prove that ϕ_ℓ is orthogonal to the previous components.

By the properties of the least squares fit of (2.10), $Y - Y_{\ell-1}^{(\phi)}$ is orthogonal to $\operatorname{span}\left\{ \langle X, \phi_1 \rangle, \ldots, \langle X, \phi_{\ell-1} \rangle \right\}$ with respect to $\langle \cdot, \cdot \rangle_{\mathbb{E}}$. Therefore, $J_\ell(\phi) = 0$ for all $\phi \in \mathcal{F}_{\ell-1}$.

The space of predictor variables admits the decomposition

$$\mathcal{X} = \mathcal{F}_{\ell-1} \oplus (\mathcal{F}_{\ell-1})^\perp,$$

where $\mathcal{F}_{\ell-1}^\perp$ is the orthogonal complement of $\mathcal{F}_{\ell-1}$.

Thus, the solution of the optimization problem in the ℓ-th iteration can be written as $\phi_\ell = \phi_\ell^{\parallel} + \phi_\ell^{\perp}$, where $\phi_\ell^{\parallel} \in \mathcal{F}_{\ell-1}$ and $\phi_\ell^{\perp} \in (\mathcal{F}_{\ell-1})^\perp$. Due to the orthogonality of the decomposition, and since J_ℓ is a linear functional, one has

$$\langle \phi, \phi \rangle = \left\langle \phi_\ell^{\parallel}, \phi_\ell^{\parallel} \right\rangle + \left\langle \phi_\ell^{\perp}, \phi_\ell^{\perp} \right\rangle, \quad \text{and} \quad J_\ell(\phi_\ell) = J_\ell(\phi_\ell^{\perp}) + J_\ell(\phi_\ell^{\parallel}) = J_\ell(\phi_\ell^{\perp}).$$

Note that the restriction in (2.11) implies $J_\ell(\phi_\ell^{\perp}) = J_\ell(\phi_\ell) \neq 0$ and, therefore, $\phi_\ell^{\perp} \neq 0$ and $\|\phi_\ell^{\perp}\| \neq 0$. Furthermore, it is possible to show that $J_\ell(\phi_\ell) > 0$ by contradiction. If

$J_\ell(\phi_\ell) < 0$, then $J_\ell(-\phi_\ell) = -J_\ell(\phi_\ell) > 0 > J_\ell(\phi_\ell)$ and, thus, ϕ_ℓ does not maximize J_ℓ. Therefore, necessarily, $J_\ell(\phi_\ell) > 0$.

Using all these properties, we can prove that $\phi_\ell^{\|} = 0$. We proceed by contradiction, assuming that $\phi_\ell^{\|} \neq 0$. Under that hypothesis, we can find $g \neq \phi_\ell$ such that $J_\ell(g) > J_\ell(\phi_\ell)$:

$$g = \frac{\|\phi_\ell^\perp\|^2 + \|\phi_\ell^{\|}\|^2}{\|\phi_\ell^\perp\|^2}\phi_\ell^\perp \implies J_\ell(g) = \frac{\|\phi_\ell^\perp\|^2 + \|\phi_\ell^{\|}\|^2}{\|\phi_\ell^\perp\|^2}J_\ell(\phi_\ell) > J_\ell(\phi_\ell),$$

Therefore, necessarily, $\phi_\ell^{\|} = 0$, which implies that ϕ_ℓ is perpendicular to all the previous basis elements. □

Remark. *Since this basis is orthogonal, the optimization problem in* (2.8) *is equivalent to the following formulation, which includes explicitly the orthogonality constraints, in spite of the fact that they are redundant:*

$$\phi_\ell = \underset{\phi\in\mathcal{X}}{\operatorname{argmax}} \ \operatorname{Cov}\left(\langle\phi, X\rangle, Y - Y_{\ell-1}^{(\phi)}\right) \quad \textit{subject to} \quad \langle\phi,\phi\rangle = 1; \\ \langle\phi,\phi_i\rangle = 0 \quad i = 1,\dots,\ell-1,$$

The importance of the expression (2.12) is that it can be used to show that the subspace generated by the orthogonal basis is a Krylov space. A Krylov subspace is defined as follows:

Definition 1. *The order L Krylov space generated by an operator $A : \mathcal{X} \to \mathcal{X}$ and an element $b \in \mathcal{X}$ is defined as*

$$\operatorname{Kry}_L(A, b) = \operatorname{span}\left\{b, Ab, \dots A^{L-1}b\right\}.$$

We can now formally prove that the space generated by the basis functions corresponds to the Krylov subspace.

Proposition 2.2.3. *The basis $\{\phi_\ell\}_{\ell=1}^L$ spans the Krylov subspace $\operatorname{Kry}_L(\mathcal{K}, \gamma)$, where $\mathcal{K}$ is the covariance operator of the regressors and γ is the cross-covariance.*

Proof. This property can be proved by induction. If $L = 1$,

$$\phi_1 \propto \gamma \quad \text{and} \quad \operatorname{Kry}_1(\mathcal{K}, \gamma) = \operatorname{span}\{\gamma\} \implies \operatorname{span}\{\phi_1\} = \operatorname{span}\{\gamma\} = \operatorname{Kry}_1(\mathcal{K}, \gamma).$$

Then, we can assume that the property holds for $\ell \leq k-1$. Therefore, due to the properties of the Krylov subspaces,

$$\phi_1, \dots, \phi_{k-1} \in \operatorname{Kry}_{k-1}(\mathcal{K}, \gamma) \implies \mathcal{K}\phi_1, \dots, \mathcal{K}\phi_{k-1} \in \operatorname{Kry}_k(\mathcal{K}, \gamma).$$

From (2.12), ϕ_k is a linear combination of these elements, and $\gamma \in \operatorname{Kry}_\ell(\mathcal{K}, \gamma)$ for all $\ell \geq 1$. Therefore, necessarily $\phi_k \in \operatorname{Kry}_k(\mathcal{K}, \gamma)$, and $\operatorname{span}\{\phi_1, \dots, \phi_k\} \subset \operatorname{Kry}_k(\mathcal{K}, \gamma)$. Fi-

nally, since the basis functions are pair-wise orthogonal, the dimension of span $\{\phi_1, \ldots, \phi_k\}$ is k. As a result, the previous inclusion must be an equality. □

To conclude the exploration of this first form of PLS, we introduce the PLS approximation of β^*, which we will denote $\beta_L^{(\phi)}$. This approximation appears naturally by rewriting the approximation of Y based on the PLS basis in (2.9) so that we obtain an expression analogous to (2.4).

$$Y_L^{(\phi)} = \sum_{\ell=1}^{L} b_\ell^{(\phi,L)} \langle X, \phi_\ell \rangle = \left\langle X, \sum_{\ell=1}^{L} b_\ell^{(\phi,L)} \phi_\ell \right\rangle = \left\langle X, \beta_L^{(\phi)} \right\rangle, \quad \beta_L^{(\phi)} = \sum_{\ell=1}^{L} b_\ell^{(\phi,L)} \phi_\ell.$$

This expression also shows that $\beta_L^{(\phi)}$ is an element of the Krylov subspace generated by the PLS basis. However, note that the coefficients of $\beta_L^{(\phi)}$ are different for different values of L. That is to say, for example, $b_1^{(\phi,1)}$ is not necessarily equal to $b_1^{(\phi,2)}$. This is a major drawback of this basis, since the PLS estimations cannot be obtained by truncating expansion on a basis with more components, unlike in PCR (see (2.5) and (2.6)). In Section 2.2.3, we will introduce the conjugate PLS basis, which does fulfill that property.

2.2.2 PLS as a constrained least squares problem

In this section, we explore a different interpretation of PLS. We forego the explicit definition of the basis, and define the PLS approximation as the solution of a least squares problem constrained to a Krylov subspace. With this goal in mind, we manipulate the optimization problem formulated in (2.10) to obtain a characterization of the PLS approximation to β^* as the result of an optimization problem:

$$\begin{aligned}\min_{(b_1,\ldots,b_\ell)\in\mathbb{R}^\ell} \mathbb{E}\left[\left(Y - \sum_{i=1}^{\ell} b_i \langle X, \phi_i \rangle\right)^2\right] &= \min_{(b_1,\ldots,b_\ell)\in\mathbb{R}^\ell} \mathbb{E}\left[\left(Y - \left\langle X, \sum_{i=1}^{\ell} b_i \phi_i \right\rangle\right)^2\right] = \\ &= \min_{\beta\in \mathrm{Kry}_\ell(\mathcal{K},\gamma)} \mathbb{E}\left[(Y - \langle X, \beta \rangle)^2\right],\end{aligned}$$

where in the last step we have used that the basis obtained spans the Krylov space of order L (Proposition 2.2.3). Therefore, one can define the PLS approximation of β^* as

$$\beta_L^{(\mathrm{PLS})} = \operatorname*{argmin}_{\beta\in \mathrm{Kry}_L(\mathcal{K},\gamma)} \mathbb{E}\left[(Y - \langle X, \beta \rangle)^2\right] = \beta_L^{(\phi)} = \sum_{\ell=1}^{L} b_\ell^{(\phi,L)} \phi_\ell. \tag{2.13}$$

Note that we have dropped the dependency on the particular PLS basis defined in Proposition 2.2.1. Moreover, we can also define the PLS approximation of Y without specifying the basis as

$$Y_L^{(\mathrm{PLS})} = \left\langle X, \beta_L^{(\mathrm{PLS})} \right\rangle.$$

Additionally, the least squares approximation can also be interpreted as result of minimizing distance to β^*, while staying in the Krylov subspace, if we consider the distance defined by the $\mathcal{K}$-product. This product is defined as

$$\langle u, v \rangle_{\mathcal{K}} = \langle u, \mathcal{K} v \rangle, \qquad u, v \in \mathcal{X}.$$

As usual, from this inner product, a norm: $\|u\|_{\mathcal{K}} = \langle u, u\rangle_{\mathcal{K}}$ ($\mathcal{K}$-norm), and thus a distance are derived. As the following results illustrate, this inner product has some desirable properties. Moreover, in Chapter 3, the properties of the sample version of this inner product are explored in further detail.

Proposition 2.2.4. *The PLS approximation of β^* minimizes the $\mathcal{K}$-norm of its difference with β^*. That is to say,*

$$\beta_L^{(\mathrm{PLS})} = \underset{\beta\in \mathrm{Kry}_L(\mathcal{K},\gamma)}{\operatorname{argmin}} \langle \beta-\beta^*, \beta-\beta^*\rangle_{\mathcal{K}}\,.$$

Proof. This is a consequence of manipulating (2.13):

$$\begin{aligned}
\beta_L^{(\mathrm{PLS})} &= \underset{\beta\in \mathrm{Kry}_L(\mathcal{K},\gamma)}{\operatorname{argmin}} \mathbb{E}\left[(Y-\langle X,\beta\rangle)^2\right] = \\
&= \underset{\beta\in \mathrm{Kry}_L(\mathcal{K},\gamma)}{\operatorname{argmin}} \mathbb{E}\left[Y^2\right] - \mathbb{E}\left[2Y\langle X,\beta\rangle\right] + \mathbb{E}\left[\langle X,\beta\rangle^2\right] = \\
&= \underset{\beta\in \mathrm{Kry}_L(\mathcal{K},\gamma)}{\operatorname{argmin}} -2\langle\gamma,\beta\rangle + \langle\beta,\beta\rangle_{\mathcal{K}} = \\
&= \underset{\beta\in \mathrm{Kry}_L(\mathcal{K},\gamma)}{\operatorname{argmin}} -2\langle\mathcal{K}\beta^*,\beta\rangle + \langle\beta,\beta\rangle_{\mathcal{K}} + \langle\beta^*,\beta^*\rangle_{\mathcal{K}} = \\
&= \underset{\beta\in \mathrm{Kry}_L(\mathcal{K},\gamma)}{\operatorname{argmin}} \langle\beta-\beta^*,\beta-\beta^*\rangle_{\mathcal{K}}\,,
\end{aligned}$$

where we have applied that $\mathcal{K}\beta^* = \gamma$. □

Since the PLS approximation can be obtained as the least squares approximation restricted to the Krylov subspace, $\beta_L^{(\mathrm{PLS})}$ can also be expressed in terms of projections onto the Krylov subspace. To obtain this characterization, we begin by defining and enumerating the properties of the $\mathcal{K}$-conjugate projection onto the Krylov subspace.

Definition 2. *The $\mathcal{K}$-conjugate projection onto $\mathrm{Kry}_L(\mathcal{K},\gamma)$ is the linear surjective transformation $\pi^{(L)}: \mathcal{X} \to \mathrm{Kry}_L(\mathcal{K},\gamma)$ that is idempotent and self-adjoint with respect to the $\mathcal{K}$ inner product. That is to say, it fulfills*

$$\pi^{(L)}\left(\pi^{(L)}(u)\right) = \pi^{(L)}(u) \quad \textit{and} \quad \left\langle \pi^{(L)}(u), v\right\rangle_{\mathcal{K}} = \left\langle u, \pi^{(L)}(v)\right\rangle_{\mathcal{K}},$$

for any $u, v \in \mathcal{X}$.

Proposition 2.2.5. *The $\mathcal{K}$-conjugate projection fulfills*

$$u - \pi^{(L)}(u) \in \left(\mathrm{Kry}_L(\mathcal{K},\gamma)\right)^{\perp_{\mathcal{K}}},$$

for any $u \in \mathcal{X}$, where

$$\left(\mathrm{Kry}_L(\mathcal{K},\gamma)\right)^{\perp_{\mathcal{K}}} = \left\{v \in \mathcal{X} : \langle v, s\rangle_{\mathcal{K}} = 0, \quad \forall s \in \mathrm{Kry}_L(\mathcal{K},\gamma)\right\}.$$

Additionally, for any $u \in \mathrm{Kry}_L(\mathcal{K},\gamma)$, $\pi^{(L)}(u) = u$.

Proof. This is a consequence of the idempotency and self-adjoint properties. For all $v \in \mathrm{Kry}_L(\mathcal{K}, \gamma)$,

$$\left\langle u - \pi^{(L)}(u), v \right\rangle_{\mathcal{K}} = \left\langle u - \pi^{(L)}(u), \pi^{(L)}(w) \right\rangle_{\mathcal{K}} = \left\langle \left(\pi^{(L)} - \left(\pi^{(L)}\right)^2 \right)(u), w \right\rangle_{\mathcal{K}} = 0,$$

where $w \in \mathcal{X}$ is such that $\pi^{(L)}(w) = v$. The existence of w is given by the surjective property, while the second step is a consequence of the self-adjoint property, and the third is due to the idempotency.

The last remark in the proposition is a consequence of the previous property. If $u \in \mathrm{Kry}_L(\mathcal{K}, \gamma)$, $u - \pi^{(L)}(u) \in \mathrm{Kry}_L(\mathcal{K}, \gamma)$. However, we have proved that $u - \pi^{(L)}(u) \in (\mathrm{Kry}_L(\mathcal{K}, \gamma))^{\perp_{\mathcal{K}}}$. Therefore, $u - \pi^{(L)}(u)$ must be in the intersection, which only contains zero. As a result, $u = \pi^{(L)}(u)$. □

Using the projection operator, the PLS approximation of β^* can be expressed as its $\mathcal{K}$-conjugate projection onto the Krylov subspace.

Proposition 2.2.6. *The PLS estimation of β^* with L components can be expressed as*

$$\beta_L^{(\mathrm{PLS})} = \pi^{(L)}(\beta^*),$$

where $\pi^{(L)}$ is the $\mathcal{K}$-conjugate projection onto the Krylov subspace, as defined in Definition 2.

Proof. The starting point is the result of Proposition 2.2.4. We start by expanding the expression to minimize, adding and subtracting $\pi^{(L)}\beta^*$:

$$\begin{aligned}\langle \beta - \beta^*, \beta - \beta^* \rangle_{\mathcal{K}} &= \left\langle \left(\beta - \pi^{(L)}\beta^*\right) + \left(\pi^{(L)}\beta^* - \beta^*\right), \left(\beta - \pi^{(L)}\beta^*\right) + \left(\pi^{(L)}\beta^* - \beta^*\right) \right\rangle_{\mathcal{K}} = \\ &= \left\langle \beta - \pi^{(L)}\beta^*, \beta - \pi^{(L)}\beta^* \right\rangle_{\mathcal{K}} \\ &+ 2 \left\langle \beta - \pi^{(L)}\beta^*, \pi^{(L)}\beta^* - \beta^* \right\rangle_{\mathcal{K}} \\ &+ \left\langle \pi^{(L)}\beta^* - \beta^*, \pi^{(L)}\beta^* - \beta^* \right\rangle_{\mathcal{K}}.\end{aligned}$$

We can now consider the middle term. Since $\beta \in \mathrm{Kry}_L(\mathcal{K}, \gamma)$, and $\pi^{(L)}(\beta^*) \in \mathrm{Kry}_L(\mathcal{K}, \gamma)$, $\beta - \pi^{(L)}\beta^* \in \mathrm{Kry}_L(\mathcal{K}, \gamma)$. However, from Proposition 2.2.5, $\beta^* - \pi^{(L)}\beta^* \in (\mathrm{Kry}_L(\mathcal{K}, \gamma))^{\perp_{\mathcal{K}}}$. Therefore, this conjugate product is zero. Moreover, the third term obtained does not depend on β. Therefore, the optimization problem can be simplified to

$$\beta_L^{(\mathrm{PLS})} = \underset{\beta \in \mathrm{Kry}_L(\mathcal{K}, \gamma)}{\operatorname{argmin}} \left\langle \beta - \pi^{(L)}\beta^*, \beta - \pi^{(L)}\beta^* \right\rangle_{\mathcal{K}}.$$

Now, since $\pi^{(L)}\beta^* \in \mathrm{Kry}_L(\mathcal{K}, \gamma)$, this quantity is zero when $\beta = \pi^{(L)}\beta^*$. Due to the positive-definiteness of the conjugate product, this is the optimum value.

□

The characterization of PLS as a constrained least squares approximation can be utilized to compare PLS with other regression methods. In particular, to close this section, we can show that, if PCA regression provides a perfect prediction, PLS also obtains a perfect prediction, needing at most the same number of components as PCA and, in some cases, less.

Theorem 2.2.1. *Assume that the PCA prediction with L components is equal to the target variable, except for some error e independent of the regression. That is to say:*

$$Y = Y_L^{(\mathrm{PCR})} + e,$$

where $\mathbb{E}(eX) = 0$. Then, $\beta_{L-M}^{(\mathrm{PLS})} = \beta_L^{(\mathrm{PCR})}$ and, thus, $Y = Y_{L-M}^{(\mathrm{PLS})} + e$, where M is the number of repeated eigenvalues among the first L eigenvalues of the covariance operator.

Proof. The cross covariance $\gamma = \mathbb{E}(XY)$ can be expressed as

$$\gamma = \mathbb{E}(XY) = \mathbb{E}(XY_L^{(\mathrm{PCR})}) + \mathbb{E}(Xe) = \mathbb{E}(XY_L^{(\mathrm{PCR})}).$$

Now we can apply (2.7):

$$\gamma = \mathbb{E}\left(X\left\langle X, \sum_{\ell=1}^{L} b_\ell^{(\mathrm{PCR})}\psi_\ell\right\rangle\right) = \mathcal{K}\left(\sum_{\ell=1}^{L} b_\ell^{(\mathrm{PCR})}\psi_\ell\right) = \sum_{\ell=1}^{L}\lambda_\ell b_\ell^{(\mathrm{PCR})}\psi_\ell,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_L$ are the eigenvalues of $\mathcal{K}$, and $\{\psi_\ell\}_{\ell=1}^{L}$ are the corresponding eigenvectors.

In this section we saw that PLS can be characterized as a least squares minimization in a Kyrlov subspace generated by the covariance operator $\mathcal{K}$ and the cross covariance. Since the cross covariance is a combination of the eigenvectors of the covariance operator, the generated Krylov space is

$$\mathrm{Kry}_L(\mathcal{K}, \gamma) = \mathrm{span}\left\{\sum_{\ell=1}^{L}\lambda_\ell b_\ell^{(\mathrm{PCR})}\psi_\ell, \ldots, \sum_{\ell=1}^{L}\lambda_\ell^L b_\ell^{(\mathrm{PCR})}\psi_\ell\right\}.$$

Now, to show that PLS and PCR obtain the same coefficients, we need only show that the result obtained by PCR is contained in the Krylov subspace. That is to say, that

$$\begin{aligned}\sum_{\ell=1}^{L} b_\ell^{(\mathrm{PCR})}\psi_\ell &= a_1\left(\sum_{\ell=1}^{L}\lambda_\ell b_\ell^{(\mathrm{PCR})}\psi_\ell\right) + \cdots + a_L\left(\sum_{\ell=1}^{L}\lambda_\ell^L b_\ell^{(\mathrm{PCR})}\psi_\ell\right) = \\ &= \sum_{\ell=1}^{L}\left(\sum_{i=1}^{L} a_i\lambda_\ell^i\right) b_\ell^{(\mathrm{PCR})}\psi_\ell,\end{aligned} \tag{2.14}$$

for some coefficients $\{a_\ell\}_{\ell=1}^{L}$. Therefore, it suffices to show that

$$\underbrace{\begin{pmatrix} \lambda_1 & \dots & \lambda_1^L \\ \vdots & \ddots & \vdots \\ \lambda_L & \dots & \lambda_L^L \end{pmatrix}}_{\mathbf{V}} \begin{pmatrix} a_1 \\ \vdots \\ a_L \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{2.15}$$

If $\lambda_1 \neq \lambda_2 \neq \ldots \neq \lambda_L$, $\mathbf{V}$ is a Vandermonde matrix. Therefore, it is invertible, and we can find a combination of coefficients that solves (2.14). If $\lambda_i = \lambda_k$, $i \neq k$, since the independent term in all equations in (2.15) is the same, we can discard the i-th equation, set the last coefficient of $\mathbf{a}$ to zero, and consider the system containing the rest of the equations. This process can be repeated until all repeated eigenvalues have been removed from $\mathbf{V}$. Since, at each step we also discarded its last column, $\mathbf{V}$ will be a square Vandermonde matrix, and we can invert it to find the values of the remaining coefficients.

This process can be applied as many times as repeated eigenvalues. Therefore, the last M coefficients of $(a_1, \ldots, a_L)$ are zero, where M is the number of repeated eigenvalues among the first L eigenvalues. Therefore, $\beta_L^{(\mathrm{PLS})}$ will also be contained in the Krylov space of dimension $L - M$, which implies that PLS will converge after $L - M$ iterations.

□

As shown in this section, the $\mathcal{K}$-product appears naturally when attempting to express the least squares solution as a projection onto the underlying space. Therefore, it is natural to consider the construction of a PLS basis where the basis elements are pair-wise orthogonal with respect to this inner product. This basis is the focus of the next chapter.

2.2.3 The conjugate PLS basis

In the previous sections, PLS was introduced as an iterative process that identifies a sequence of subspaces of increasing dimension in a Hilbert space. Naturally, these subspaces can be characterized by the sequence of elements that generate them, and, in particular, by a basis of the space, to which an element is added each iteration. Further, we showed that these spaces are Krylov spaces generated by the covariance of the regressors and the cross covariance, defined as

$$\mathrm{Kry}_L(\mathcal{K}, \gamma) = \mathrm{span}\left\{\gamma, \mathcal{K}\gamma, \ldots, \mathcal{K}^{L-1}\gamma\right\}.$$

Therefore, a natural basis to consider is $\{\gamma, \mathcal{K}\gamma, \ldots, \mathcal{K}^{L-1}\gamma\}$. We will refere to this basis as the Krylov basis. In principle, this set of elements need not be linearly independent. However, as long as L is less than the maximum dimension of the Krylov subspace, they are linearly independent. The maximum dimension of a Krylov subspace generated by $\mathcal{K}$ and γ, $d_{\max}$, is the smallest value that fulfills $\mathrm{Kry}_r(\mathcal{K}, \gamma) = \mathrm{Kry}_{d_{\max}}(\mathcal{K}, \gamma)$ for all $r > d_{\max}$. Consider the Krylov subspace generated by the first L elements. As long as $L < d_{\max}$,

$\mathcal{K}^L\gamma$ must be linearly independent of the previous L elements. We can prove this fact by contradiction. If $\mathcal{K}^L\gamma$ can be expressed as a linear combination of the previous L elements, $\mathcal{K}^L\gamma = \sum_{\ell=1}^{L} a_\ell \mathcal{K}^{\ell-1}\gamma$, for some coefficients $\{a_\ell\}_{\ell=1}^{L}$. However, then:

$$\begin{aligned}\operatorname{Kry}_{L+1}(\mathcal{K},\,\gamma) &= \operatorname{span}\left\{\gamma, \mathcal{K}\gamma, \ldots, \mathcal{K}^{L-1}\gamma, \mathcal{K}^L\gamma\right\} = \operatorname{span}\left\{\gamma, \mathcal{K}\gamma, \ldots, \mathcal{K}^{L-1}\gamma, \sum_{\ell=1}^{L} a_\ell \mathcal{K}^{\ell-1}\gamma\right\} \\ &= \operatorname{span}\left\{\gamma, \mathcal{K}\gamma, \ldots, \mathcal{K}^{L-1}\gamma\right\} = \operatorname{Kry}_{L+1}(\mathcal{K},\,\gamma),\end{aligned}$$

and thus L fulfills the condition for the maximum dimension. Therefore, $d_{\max} \leq L$, which concludes this proof and shows that the Krylov basis is indeed a basis as long as $L < d_{\max}$.

In the literature, this basis is utilized due to its convenience to prove theoretical properties of the PLS method (Delaigle & Hall, 2012). However, the Krylov basis tends to become almost linearly dependent, as repeated exponentiation of the operator scales the eigenvectors exponentially depending on their eigenvalues. As a result, it is not usually employed in numerical algorithms.

Another alternative is the orthogonal basis introduced in Section 2.2.1. However, as we already discussed, the PLS approximation with L components cannot be calculated by truncating its expansion on this basis, as the coefficients depend on dimension of the basis.

In this section, we introduce yet another basis, in which the coefficients do not depend on the total number of components considered: the conjugate basis. The components of this basis are not orthogonal with respect to the usual inner product, instead they are orthogonal with respect to the $\mathcal{K}$-product introduced in the last section. This basis can be defined by adding conjugacy constraints to (2.8) as follows

$$\begin{aligned}\varphi_\ell = \underset{\varphi\in\mathcal{X}}{\operatorname{argmax}}\ \operatorname{Cov}\left(\langle\varphi, X\rangle, Y - Y_{\ell-1}^{(\varphi)}\right) \quad \text{subject to} \quad & \langle\varphi,\varphi\rangle = 1; \\ & \langle\varphi,\varphi_i\rangle_{\mathcal{K}} = 0 \quad i = 1,\ldots,\ell-1,\end{aligned} \tag{2.16}$$

where $Y_L^{(\varphi)}$ is defined as

$$Y_\ell^{(\varphi)} = \sum_{i=1}^{\ell} b_i^{(\varphi,\ell)} \langle\varphi_i, X\rangle,$$

and $\{b_i^{(\varphi,\ell)}\}_{i=1}^{\ell}$ are fitted by least squares:

$$(b_1^{(\varphi,\ell)},\ldots,b_\ell^{(\varphi,\ell)}) = \underset{(b_1,\ldots,b_\ell)\in\mathbb{R}^\ell}{\operatorname{argmin}}\ \mathbb{E}\left[\left(Y - \sum_{i=1}^{\ell} b_i \langle X,\varphi_i\rangle\right)^2\right]. \tag{2.17}$$

However, in this case, the function to optimize can be simplified due to the conjugacy constraints.

Lemma 2.2.2. *Under the conjugacy constraints in* (2.16)*,*

$$\operatorname{Cov}\left(\langle\varphi, X\rangle, Y - Y_{\ell-1}^{(\varphi)}\right) = \operatorname{Cov}\left(\langle\varphi, X\rangle, Y\right).$$

Proof. This can be proved by expanding the LHS

$$\begin{aligned}\operatorname{Cov}\left(\langle\varphi, X\rangle, Y - Y_{\ell-1}^{(\varphi)}\right) &= \mathbb{E}\left[Y\langle\varphi, X\rangle - Y_{\ell-1}^{(\varphi)}\langle\varphi, X\rangle\right] = \\ &= \mathbb{E}\left[Y\langle\varphi, X\rangle\right] - \mathbb{E}\left[Y_{\ell-1}^{(\varphi)}\langle\varphi, X\rangle\right] = \\ &= \mathbb{E}\left[Y\langle\varphi, X\rangle\right] - \sum_{i=1}^{\ell-1} b_i^{(\varphi,\ell)}\mathbb{E}\left[\langle\varphi_i, X\rangle\langle\varphi, X\rangle\right] = \\ &= \mathbb{E}\left[Y\langle\varphi, X\rangle\right] - \sum_{i=1}^{\ell-1} b_i^{(\varphi,\ell)}\langle\varphi, \varphi_i\rangle_{\mathcal{K}} = \\ &= \mathbb{E}\left[Y\langle\varphi, X\rangle\right] = \\ &= \operatorname{Cov}\left(\langle\varphi, X\rangle, Y\right).\end{aligned}$$

□

As in the analysis of the orthogonal basis, we now seek to find an expression for the basis functions.

Proposition 2.2.7. *The solution of the optimization problem* (2.16) *is*

$$\varphi_\ell = \frac{1}{\lambda_0^{(\ell)}}\left(\gamma - \sum_{i=1}^{\ell-1}\lambda_i^{(\ell)}\mathcal{K}\varphi_i\right), \tag{2.18}$$

where $\{\lambda_i^{(\ell)}\}_{i=1}^{\ell-1}$ *are constants determined by the conjugacy constraints and* $\lambda_0^{(\ell)}$ *is the normalization constant.*

Proof. Applying the simplification of Lemma 2.2.2, we obtain the Lagrangian

$$\mathcal{L}(\varphi) = \langle\varphi, \gamma\rangle - \sum_{i=1}^{\ell-1}\lambda_i(\langle\varphi, \mathcal{K}\varphi_i\rangle - 1) - \lambda_0(\langle\varphi, \varphi\rangle - 1).$$

To find the minimum, we seek the points at which the gradient is zero:

$$\frac{\partial\mathcal{L}}{\partial\varphi} = \gamma - \sum_{i=1}^{\ell-1}\lambda_i\mathcal{K}\varphi_i - \lambda_0\varphi = 0 \implies \varphi_\ell = \frac{1}{\lambda_0}\left(\gamma - \sum_{i=1}^{\ell-1}\lambda_i\mathcal{K}\varphi_i\right).$$

□

As with the orthogonal basis, the space spanned by this basis is the Krylov subspace. This is a consequence of expression (2.18) and the conjugacy constraints. Therefore, once more, $\beta_L^{(\mathrm{PLS})}$ can be expressed in this basis as

$$\beta_L^{(\mathrm{PLS})} = \sum_{\ell=1}^{L} b_\ell^{(\varphi,L)}\varphi_\ell.$$

In this case, solving the optimization problem directly did not provide a closed expression for the basis elements. In order to obtain the concrete values, it is required to find the values of the multipliers $\{\lambda_i^{(\ell)}\}_{i=1}^{\ell}$. To do so, it is possible to solve the following linear equation system

$$\langle \varphi_\ell, \varphi_j \rangle_{\mathcal{K}} = 0 \implies \langle \gamma, \varphi_j \rangle_{\mathcal{K}} - \sum_{i=1}^{\ell-1} \lambda_i^{(\ell)} \langle \mathcal{K}\varphi_i, \varphi_j \rangle_{\mathcal{K}} = 0, \qquad j = 1, \dots, \ell - 1.$$

In Section 2.3.2, the conjugate gradient method will be explored, which provides an algorithm to calculate a conjugate basis of the Krylov space without explicitly solving these equations. In the remainder of this section, we cover additional properties, which are applied in the conjugate gradient method. Our first goal is to show that we can drop the number of components in the superindex of the coefficients defined in (2.17).

Proposition 2.2.8. *The coefficients of $\beta_L^{(\mathrm{PLS})}$, the PLS approximation of β^* in this basis, do not depend on L, the dimension of the explored Krylov subspace. That is to say, we can express the PLS approximation of β using this basis as:*

$$\beta_L^{(\mathrm{PLS})} = \sum_{\ell=1}^{L} b_\ell^{(\varphi)} \varphi_\ell, \quad \textit{where} \quad b_\ell^{(\varphi)} = \frac{\left\langle \beta_L^{(\mathrm{PLS})}, \varphi_i \right\rangle_{\mathcal{K}}}{\langle \varphi_i, \varphi_i \rangle_{\mathcal{K}}}. \tag{2.19}$$

Proof. Let us consider the results of applying PLS with two different numbers of components L and K. WLOG, we will assume that $K > L$. In principle, all we know is that the PLS approximations are contained in the Krylov space. Thus, they can be expressed as:

$$\beta_L^{(\varphi)} = \sum_{i=1}^{L} b_i^{(\varphi,L)} \varphi_i \quad \text{and} \quad \beta_K^{(\varphi)} = \sum_{i=1}^{K} b_i^{(\varphi,K)} \varphi_i.$$

Our goal now is to prove that $b_i^{(\varphi,K)} = b_i^{(\varphi,L)}$ for any $i = 1, \dots, L$. To begin, we notice that the conjugancy of the basis provides a closed-form expression for these coefficients:

$$b_i^{(\varphi,L)} = \frac{\left\langle \beta_L^{(\varphi)}, \varphi_i \right\rangle_{\mathcal{K}}}{\langle \varphi_i, \varphi_i \rangle_{\mathcal{K}}} \quad \text{and} \quad b_i^{(\varphi,K)} = \frac{\left\langle \beta_K^{(\varphi)}, \varphi_i \right\rangle_{\mathcal{K}}}{\langle \varphi_i, \varphi_i \rangle_{\mathcal{K}}}. \tag{2.20}$$

Therefore, the difference between the coefficients can be expressed as

$$b_i^{(\varphi,K)} - b_i^{(\varphi,L)} = \frac{\left\langle \beta_K^{(\varphi)} - \beta_L^{(\varphi)}, \varphi_i \right\rangle_{\mathcal{K}}}{\langle \varphi_i, \varphi_i \rangle_{\mathcal{K}}}. \tag{2.21}$$

In order to show that this quantity is zero, notice that we can rewrite $\beta_L^{(\varphi)}$ as a projection of $\beta_K^{(\varphi)}$ in the following manner:

$$\beta_L^{(\varphi)} - \pi^{(L)}(\beta_K^{(\varphi)}) = \pi^{(L)}(\beta^* - \beta_K^{(\varphi)}) = \pi^{(L)}(\beta^* - \pi^{(K)}\beta^*) = 0, \tag{2.22}$$

where we have used the characterization of $\beta_K^{(\varphi)}$ as the $\mathcal{K}$-orthogonal projection onto the Krylov space and the last equality holds because

$$\beta^* - \pi^{(K)}\beta^* \in (\mathrm{Kry}_K(\mathcal{K}, \gamma))^{\perp_\mathcal{K}} \subset (\mathrm{Kry}_L(\mathcal{K}, \gamma))^{\perp_\mathcal{K}} .$$

From (2.22) we know that $\beta_L^{(\varphi)}$ can be expressed as $\pi^{(L)}(\beta_K^{(\varphi)})$. If we substitute this into (2.21), we would obtain the expression $\beta_K^{(\varphi)} - \pi^{(L)}\beta_K^{(\varphi)}$ as part of the inner product. The result of this operation is $\mathcal{K}$-orthogonal to the Krylov space of order L and, therefore, to φ_i. As a result, the inner product in the numerator of (2.21) is zero and the coefficients must be equal.

Finally, the expression for $b_\ell^{(\varphi)}$ in (2.19) is the result of dropping the L or K from the formulas in (2.20). □

Using this result, we can express the PLS approximation of Y as a series where we add a new term for every PLS component added:

$$Y_L^{(\mathrm{PLS})} = \left\langle X, \beta_L^{(\mathrm{PLS})} \right\rangle = \sum_{\ell=i}^{L} b_\ell^{(\varphi)} \langle X, \varphi_\ell \rangle = Y_{L-1}^{(\mathrm{PLS})} + b_L^{(\varphi)} \langle X, \varphi_L \rangle .$$

Additionally, by dropping the additional dependency of the coefficients on the total number of components, we obtain coefficients that have similar properties to those of PCR. This last expression is comparable to the definition of $Y_L^{(\mathrm{PCR})}$ in (2.7).

However, Proposition 2.2.8 does not provide a way of calculating $\{b_\ell^{(\varphi)}\}_{\ell=1}^L$. As described at the beginning of this section, these coefficients are defined as the result of an L-dimensional least squares fit. However, the problem can be greatly simplified. In particular, the following proposition provides an explicit expression for each coefficient.

Proposition 2.2.9. *The coefficients of the PLS estimation using the conjugate basis can be calculated explicitly as*

$$b_\ell^{(\varphi)} = \frac{\left\langle \gamma - \mathcal{K}\beta_{\ell-1}^{(\varphi)}, \varphi_\ell \right\rangle}{\langle \varphi_\ell, \varphi_\ell \rangle_\mathcal{K}} = \frac{\langle \gamma, \varphi_\ell \rangle}{\langle \varphi_\ell, \varphi_\ell \rangle_\mathcal{K}}.$$

Proof. Applying (2.10) to this basis, we obtain that

$$(b_1^{(\varphi)}, \ldots, b_\ell^{(\varphi)}) = \underset{(b_1,\ldots,b_\ell)\in\mathbb{R}^\ell}{\operatorname{argmin}} \mathbb{E}\left[\left(Y - \sum_{i=1}^{\ell} b_i \langle X, \varphi_i \rangle \right)^2 \right].$$

However, from Proposition 2.2.8, we know that all coefficients but the last are shared with the previous approximation. Therefore, we can solve this problem incrementally,

searching in a one-dimensional space at a time. To do so, we rewrite the problem as

$$\begin{aligned} b_\ell^{(\varphi)} &= \underset{b_\ell \in \mathbb{R}}{\operatorname{argmin}}\ \mathbb{E}\left[\left(Y - Y_{\ell-1}^{(\varphi)} - b_\ell \langle X, \varphi_\ell \rangle\right)^2\right] = \\ &= \underset{b_\ell \in \mathbb{R}}{\operatorname{argmin}} \left(\mathbb{E}\left[\left(Y - Y_{\ell-1}^{(\varphi)}\right)^2\right] - 2\mathbb{E}\left[(Y - Y_{\ell-1}^{(\varphi)}) b_\ell \langle X, \varphi_\ell \rangle\right] + \mathbb{E}\left[(b_\ell \langle X, \varphi_\ell \rangle)^2\right]\right) = \\ &= \underset{b_\ell \in \mathbb{R}}{\operatorname{argmin}} \left(-2b_\ell \left\langle \mathbb{E}\left[\left(Y - Y_{\ell-1}^{(\varphi)}\right) X\right], \varphi_\ell \right\rangle + b_\ell^2 \langle \varphi_\ell, \mathcal{K}\varphi_\ell \rangle\right). \end{aligned}$$

To find the solution of the optimization problem, it suffices to find the zeros of the derivative. In this case, the only zero is

$$\begin{aligned} b_\ell^{(\varphi)} &= \frac{\left\langle \mathbb{E}\left[\left(Y - Y_{\ell-1}^{(\varphi)}\right) X\right], \varphi_\ell \right\rangle}{\langle \varphi_\ell, \varphi_\ell \rangle_{\mathcal{K}}} = \\ &= \frac{\left\langle \mathbb{E}\left[\left(Y - \left\langle X, \beta_{\ell-1}^{(\varphi)} \right\rangle\right) X\right], \varphi_\ell \right\rangle}{\langle \varphi_\ell, \varphi_\ell \rangle_{\mathcal{K}}} = \\ &= \frac{\left\langle \gamma - \mathcal{K}\beta_{\ell-1}^{(\varphi)}, \varphi_\ell \right\rangle}{\langle \varphi_\ell, \varphi_\ell \rangle_{\mathcal{K}}}. \end{aligned} \tag{2.23}$$

However, we can still simplify this expression. In particular, the quantity $\left\langle \mathcal{K}\beta_{\ell-1}^{(\varphi)}, \varphi_\ell \right\rangle$ is always zero:

$$\left\langle \mathcal{K}\beta_{\ell-1}^{(\varphi)}, \varphi_\ell \right\rangle = \sum_{i=1}^{\ell-1} b_i^{(\varphi)} \langle \mathcal{K}\varphi_i, \varphi_\ell \rangle = 0,$$

where the last step is true due to the conjugacy property. Substituting this result into (2.23), we obtain

$$b_\ell^{(\varphi)} = \frac{\left\langle \gamma - \mathcal{K}\beta_{\ell-1}^{(\varphi)}, \varphi_\ell \right\rangle}{\langle \varphi_\ell, \varphi_\ell \rangle_{\mathcal{K}}} = \frac{\langle \gamma, \varphi_\ell \rangle}{\langle \varphi_\ell, \varphi_\ell \rangle_{\mathcal{K}}} - \frac{\left\langle \mathcal{K}\beta_{\ell-1}^{(\varphi)}, \varphi_\ell \right\rangle}{\langle \varphi_\ell, \varphi_\ell \rangle_{\mathcal{K}}} = \frac{\langle \gamma, \varphi_\ell \rangle}{\langle \varphi_\ell, \varphi_\ell \rangle_{\mathcal{K}}}.$$

□

This last proposition enables us to calculate the PLS approximations to β^* iteratively, calculating only two inner products for each iteration. However, the conjugate basis $\{\varphi_\ell\}_{\ell=1}^L$ is needed for these calculations. This basis can be obtained in different ways. For once, it could be derived by orthogonalizing any basis of the Krylov subspace with respect to the $\mathcal{K}$-inner product. However, a more efficient approach is provided by the conjugate gradient method, described in Section 2.3.2.

The projections of X onto a conjugate PLS basis $T_\ell = \langle X, \varphi_\ell \rangle$ are typically called scores (Rosipal & Krämer, 2005). As a result of the conjugacy of $\{\varphi_\ell\}_{\ell=1}^L$, the scores are orthogonal. Even though our focus so far has been on the projection directions that constitute the PLS basis, many sources first introduce the components, and then present PLS regression as linear regression between the components and the response variable (Wold et al., 2001; de Jong, 1993). This is equivalent to the formulation in (2.10).

Once the components are known, they can be used to reconstruct the original data. The *loadings* are defined with that goal in mind. Consider the model

$$X = \sum_{\ell=1}^{L} p_\ell T_\ell + \varepsilon,$$

where p_ℓ are the parameter to fit. We assume that $\mathbb{E}(X\epsilon) = 0$. Therefore, $\mathbb{E}(T_\ell \epsilon) = 0$ for all $\ell = 1, \dots, L$ and we can obtain an expression for p_ℓ:

$$\mathbb{E}(XT_k) = \mathbb{E}\left(\sum_{\ell=1}^{L} p_\ell T_\ell T_k\right) + \mathbb{E}(\varepsilon T_k) \implies \mathbb{E}(XT_k) = p_k \mathbb{E}(T_k T_k) + 0 \implies$$
$$\implies p_k = \frac{1}{\mathbb{E}(T_l^2)} \mathbb{E}(XT_l).$$

With these loadings, the original data can be reconstructed from the scores as

$$X_L = \sum_{\ell=1}^{L} p_\ell T_\ell.$$

Furthermore, the norm of this reconstruction is minimal, as the expression obtained for $\{p_\ell\}_{\ell=1}^{L}$ minimizes the following quantity:

$$\min_{p_1,\dots,p_k \in \mathcal{X}} \mathbb{E}\left[\left\| X - \sum_{\ell=1}^{L} p_\ell T_\ell \right\|^2\right] =$$
$$\min_{p_1,\dots,p_k \in \mathcal{X}} \mathbb{E}\left(\langle X, X\rangle\right) - 2\sum_{\ell=1}^{L} \mathbb{E}\left(T_\ell \langle X, p_\ell\rangle\right) + \sum_{\ell=1}^{L} \langle p_\ell, p_\ell\rangle \mathbb{E}\left(T_\ell^2\right) =$$
$$\mathbb{E}\left(\langle X, X\rangle\right) + \min_{p_1,\dots,p_k \in \mathcal{X}} \sum_{\ell=1}^{L} \langle p_\ell, p_\ell\rangle \mathbb{E}\left(T_\ell^2\right) - 2\sum_{\ell=1}^{L} \langle \mathbb{E}\left(T_\ell X\right), p_\ell\rangle =$$
$$\mathbb{E}\left(\langle X, X\rangle\right) + \min_{p_1,\dots,p_k \in \mathcal{X}} \sum_{\ell=1}^{L} \langle p_\ell, p_\ell\rangle \mathbb{E}\left(T_\ell^2\right) - 2\left\langle \mathbb{E}\left(T_\ell X\right), p_\ell\right\rangle,$$

where we have used that the scores are uncorrelated. To continue, we can change the order of the summation and the minimization:

$$\sum_{\ell=1}^{L} \min_{p_1,\dots,p_k \in \mathcal{X}} \langle p_\ell, p_\ell\rangle \mathbb{E}\left(T_\ell^2\right) - \langle 2\mathbb{E}\left(T_\ell X\right), p_\ell\rangle =$$
$$\sum_{\ell=1}^{L} \min_{p_1,\dots,p_k \in \mathcal{X}} \langle p_\ell, p_\ell\rangle - \left\langle 2\frac{\mathbb{E}\left(T_\ell X\right)}{\mathbb{E}\left(T_\ell^2\right)}, p_\ell\right\rangle,$$

where we can see that the minimum is reached for the value of p_ℓ in (2.2.3).

During this analysis we have seen how PLS can be understood as an iterative process that explores a Hilbert space, identifying a sequence of subspaces of increasing dimension. Since these subspaces are generated by some basis, PLS can be understood as the process of

obtaining this basis from the original data. Once the basis is known, the PLS components can be easily calculated by projecting onto the direction of the basis elements. Furthermore, different basis that span the same space can be obtained, depending on the constraints enforced. Orthogonality with respect to the usual inner product yields the orthogonal basis, while orthogonality with respect to the inner product induced by the covariance operator produces the conjugate basis. Additionally, we also showed how the expansion of the PLS approximation in the conjugate basis has the advantage of sharing the coefficients obtained in previous iterations.

2.3 Numerical algorithms for PLS regression

In this section, two algorithms that can be utilized to perform PLS regression are described. The first of them, NIPALS (Nonlinear Iterative Partial Least Squares) constructs an orthogonal base of the Krylov subspace, while calculating the PLS scores and loading. In contrast, the conjugate gradient method exploits the properties of the conjugate basis of the Krylov subspace to build directly the PLS approximation in each iteration.

2.3.1 NIPALS

NIPALS is an iterative algorithm that calculates the PLS scores, the loadings and an orthogonal basis of the Krylov subspace $\mathrm{Kry}_L(\mathcal{K}, \gamma)$. One of the first versions of this algorithm can be found in Noonan and Wold (1977), which has been the object of successive refinements (Wegelin, 2000). After each iteration, the regressor X gets deflated (see line 7 of Algorithm 1). This step ensures that the next component is orthogonal to the previous ones. This deflation can be understood as subtracting the information already present in the extracted components.

Algorithm 1 NIPALS

Input: X, Y and L
Output: $\{\phi_\ell\}_{\ell=1}^L$, $\{T_\ell\}_{\ell=1}^L$ and $\{p_\ell\}_{\ell=1}^L$

1: $X_0 \leftarrow X$
2: **for** $\ell = 1, \ldots, L$ **do**
3: $\quad \phi_\ell \leftarrow \frac{1}{\|\mathbb{E}(X_{\ell-1}Y)\|}\mathbb{E}(X_{\ell-1}Y)$ ▷ Basis
4: $\quad T_\ell \leftarrow \langle \phi_\ell, X_{\ell-1} \rangle$ ▷ Projection directions
5: $\quad \mathcal{K}_{\ell-1} f := \mathbb{E}(X_{\ell-1} \langle X_{\ell-1}, f \rangle)$ ▷ Deflated covariance operator
6: $\quad p_\ell \leftarrow \frac{1}{\langle \phi_\ell, \mathcal{K}_{\ell-1}\phi_\ell \rangle}\mathcal{K}_{\ell-1}\phi_\ell$ ▷ Loadings
7: $\quad X_\ell \leftarrow X_{\ell-1} - T_\ell p_\ell$ ▷ Deflate regressor
8: **end for**

In the remainder of this section, we show that the NIPALS calculates the orthogonal basis obtained in Proposition 2.2.1, alongside the PLS scores and loadings. To get started, the following proposition provides an alternate expression for the deflated regressor.

Proposition 2.3.1. *An alternative expression for X_ℓ is*

$$X_\ell = X_{\ell-1} - T_\ell \mathbb{E}(X_{\ell-1}T_\ell)/\mathbb{E}(T_\ell^2).$$

Proof. We can apply the identities in the algorithm until we reach the desired expression:

$$\begin{aligned} X_\ell = X_{\ell-1} - T_\ell p_\ell &= X_{\ell-1} - T_\ell \left(\frac{1}{\langle \phi_\ell, \mathcal{K}_{\ell-1}\phi_\ell \rangle} \mathcal{K}_{\ell-1}\phi_\ell \right) = \\ &= X_{\ell-1} - T_\ell \left(\frac{1}{\langle \phi_\ell, \mathcal{K}_{\ell-1}\phi_\ell \rangle} \mathbb{E}(X_{\ell-1} \langle X_{\ell-1}, \phi_\ell \rangle) \right) = \\ &= X_{\ell-1} - T_\ell \left(\frac{1}{E(T_\ell^2)} \mathbb{E}(X_{\ell-1}T_\ell) \right) = \\ &= X_{\ell-1} - T_\ell \mathbb{E}(X_{\ell-1}T_\ell) / \left(E(T_\ell^2) \right), \end{aligned}$$

where we have used that

$$\begin{aligned} \langle \phi_\ell, \mathcal{K}_{\ell-1}\phi_\ell \rangle = \langle \phi_\ell, \mathbb{E}(X_\ell \langle X_{\ell-1}, \phi_\ell \rangle) \rangle &= \langle \phi_\ell, \mathbb{E}(X_\ell T_\ell) \rangle = \\ &= \mathbb{E}(\langle \phi_\ell, X_\ell \rangle T_\ell) = \mathbb{E}(T_\ell^2). \end{aligned}$$

□

The following result shows that the deflation of NIPALS produces orthogonal scores.

Theorem 2.3.1. *The scores $\{T_\ell\}_{\ell=1}^L$ obtained by NIPALS are uncorrelated. That is to say, $\mathbb{E}(T_i T_j) = 0$ if $1 \leq i, j \leq L$, $i \neq j$.*

Proof. We proceed by induction. If $L = 1$, the statement is trivial. To complete the proof, we will assume that the scores $\{T_\ell\}_{\ell=1}^{k-1}$ are orthogonal and prove that, then, the scores $\{T_\ell\}_{\ell=1}^{k}$ are orthogonal.

Therefore, we focus on T_k. For any $1 \leq i < k$ we need to prove that $\mathbb{E}(T_i T_k) = 0$. Before starting, we note that our goal will be proven if $\mathbb{E}(T_i X_{k-1}) = 0$ since

$$\mathbb{E}(T_i T_k) = \mathbb{E}(T_i \langle \phi_k, X_{k-1} \rangle) = \langle \phi_k, \mathbb{E}(T_i X_{k-1}) \rangle .$$

If $i < k-1$, from line 7 of NIPALS,

$$\begin{aligned} \mathbb{E}(T_i X_{k-1}) &= \mathbb{E}(T_i X_{k-2} - T_i T_{k-1} p_{k-1}) = \\ &= \mathbb{E}(T_i X_{k-2}) - \mathbb{E}(T_i T_{k-1} p_{k-1}) = \\ &= \mathbb{E}(T_i X_{k-2}) = \cdots = \mathbb{E}(T_i X_i). \end{aligned}$$

Finally, either if $i = k-1$, or following the previous steps, we can substitute the

expression for X_i in Proposition 2.3.1,

$$\begin{aligned}\mathbb{E}(T_i X_i) &= \mathbb{E}(T_i X_{i-1} - T_i T_i \mathbb{E}(X_{i-1} T_i)/\mathbb{E}(T_i^2)) = \\ &= \mathbb{E}(T_i X_{i-1}) - \mathbb{E}(T_i T_i)\mathbb{E}(X_{i-1} T_i))/\mathbb{E}(T_i^2) = \\ &= \mathbb{E}(T_i X_{i-1}) - \mathbb{E}(X_{i-1} T_i)) = 0.\end{aligned}$$

□

The next result is a direct consequence of the deflation in NIPALS. These three claims will be used extensively to prove the following theorems.

Lemma 2.3.2. *The following equalities hold*

$$\begin{aligned}&1.\ \mathbb{E}(X_i T_\ell) = \mathbb{E}(X_j T_\ell) \quad \text{if} \quad 0 \leq j < i < \ell \leq L; \\ &2.\ \langle \phi_i, p_j \rangle = 0 \quad \text{if} \quad 1 \leq i < j \leq L; \\ &3.\ \langle \phi_i, p_i \rangle = 1 \quad \text{if} \quad 1 \leq i \leq L.\end{aligned} \tag{2.24}$$

Proof. We prove them sequentially:

1. $\mathbb{E}(X_i T_\ell) = \mathbb{E}(X_j T_\ell) \quad \text{if} \quad j < i < l.$

   It is a consequence of the deflation and the orthogonality of the PLS scores

   $$\begin{aligned}\mathbb{E}(X_i T_\ell) &= \mathbb{E}\left(X_j - \sum_{m=j+1}^{i} T_m p_m\right) T_\ell = \\ &= \mathbb{E}(X_j T_j) - \sum_{m=j+1}^{i} \mathbb{E}(T_m T_\ell) p_m = \\ &= \mathbb{E}(X_j T_\ell),\end{aligned}$$

   where the last step holds due to the orthogonality of the scores.

2. $\langle \phi_i, p_j \rangle = 0 \quad \text{if} \quad i < j.$

   It is a consequence of the definition of p_j and the previous property.

   $$\begin{aligned}\langle \phi_i, p_j \rangle &= \langle \phi_i, \mathbb{E}(X_{j-1} \langle X_{j-1}, \phi_j \rangle) \rangle = \langle \phi_i, \mathbb{E}(X_{j-1} T_j) \rangle = \langle \phi_i, \mathbb{E}(X_i T_j) \rangle = \\ &= \mathbb{E}(\langle \phi_i, X_i T_j \rangle) = \mathbb{E}(T_j \langle \phi_i, X_i \rangle) = \mathbb{E}(T_j T_i) = 0.\end{aligned}$$

3. $\langle \phi_i, p_i \rangle = 1.$

   It is a direct consequence of the definition of p_i in NIPALS.

   $$\langle \phi_i, p_i \rangle = \frac{1}{\langle \phi_\ell, \mathcal{K}_{\ell-1} \phi_\ell \rangle} \langle \phi_\ell, \mathcal{K}_{\ell-1} \phi_\ell \rangle = 1.$$

□

The following theorem shows that NIPALS explores Krylov spaces of increasing order. This is the first step in showing that the basis produced is an orthogonal basis of the Krylov space

Theorem 2.3.3. *The loadings obtained by NIPALS fulfill $p_\ell \in \mathrm{Kry}_\ell(\mathcal{K}, \mathcal{K}\gamma)$, while the basis elements fulfill $\phi_\ell \in \mathrm{Kry}_\ell(\mathcal{K}, \gamma)$.*

Proof. This proof is performed by induction.

In the first iteration of NIPALS, $\phi_1 \propto \gamma$. Therefore, $\phi_1 \in \mathrm{Kry}_1(\mathcal{K}, \gamma)$. Additionally, since $\mathcal{K}_0 f = \mathcal{K} f$, $p_1 \propto \mathcal{K}\gamma$. Knowing this, we proceed inductively. That is to say, we assume that

$$p_m \in \mathrm{Kry}_m(\mathcal{K}, \mathcal{K}\gamma), \quad \phi_m \in \mathrm{Kry}_m(\mathcal{K}, \gamma), \qquad m < \ell.$$

We will prove each of the two statements separately.

$\underline{\phi_\ell \in \mathrm{Kry}_\ell(\mathcal{K}, \gamma)}$

By expanding ϕ_ℓ, we obtain

$$\begin{aligned}\phi_\ell \propto \mathbb{E}(X_{\ell-1}Y) = \mathbb{E}\left((X_{\ell-2} - T_{\ell-1}p_{\ell-1})Y\right) = a\phi_{\ell-1} - \mathbb{E}\left(t_{\ell-1}p_{\ell-1}Y\right) = \\ = a\phi_{\ell-1} - p_{\ell-1}\mathbb{E}\left(T_{\ell-1}Y\right) = a\phi_{\ell-1} - p_{\ell-1}b,\end{aligned}$$

where a and b are constants. Since $p_{\ell-1} \in \mathrm{Kry}_{\ell-1}(\mathcal{K}, \mathcal{K}\gamma) \subset \mathrm{Kry}_\ell(\mathcal{K}, \gamma)$, this result implies $\phi_\ell \in \mathrm{Kry}_\ell(\mathcal{K}, \gamma)$.

$\underline{p_\ell \in \mathrm{Kry}_\ell(\mathcal{K}, \mathcal{K}\gamma)}$

Finally, we expand p_ℓ. In order to do so, we first apply its definition in the algorithm and, then, the second property of Lemma 2.3.2.

$$\begin{aligned}p_\ell \propto \mathcal{K}_{\ell-1}\phi_\ell = \mathbb{E}(X_{\ell-1}T_\ell) = \mathbb{E}(XT_\ell) = \mathbb{E}(X\left\langle X_{\ell-1}, \phi_\ell\right\rangle) = \\ = \mathbb{E}\left(X\left\langle\left(X - \sum_{i=1}^{\ell-1} T_i p_i\right), \phi_\ell\right\rangle\right) = \\ = \mathbb{E}\left(X\left\langle X, \phi_\ell\right\rangle - \sum_{i=1}^{\ell-1} XT_i\left\langle p_i, \phi_\ell\right\rangle\right) = \\ = \mathbb{E}(X\left\langle X, \phi_\ell\right\rangle) - \sum_{i=1}^{\ell-1}\left\langle p_i, \phi_\ell\right\rangle\mathbb{E}(XT_i) = \\ = \mathbb{E}(X\left\langle X, \phi_\ell\right\rangle) - \sum_{i=1}^{\ell-1}\left\langle p_i, \phi_\ell\right\rangle\mathbb{E}(X_{i-1}T_i) = \\ = \mathbb{E}(X\left\langle X, \phi_\ell\right\rangle) - \sum_{i=1}^{\ell-1}\left\langle p_i, \phi_\ell\right\rangle\left\langle \phi_i, \mathcal{K}_{i-1}\phi_i\right\rangle p_i = \\ = a\mathcal{K}\phi_\ell - \sum_{i=1}^{\ell-1} b_i p_i,\end{aligned}$$

where a and $\{b_i\}_{i=1}^{\ell-1}$ are unimportant constants. Since $\phi_\ell \in \mathrm{Kry}_\ell(\mathcal{K}, \gamma)$, $\mathcal{K}\phi_\ell \in \mathrm{Kry}_\ell(\mathcal{K}, \mathcal{K}\gamma)$. Therefore, $p_\ell \in \mathrm{Kry}_\ell(\mathcal{K}, \mathcal{K}\gamma)$.

□

Finally, we can show that the basis extracted by NIPALS form an orthogonal basis of the Krylov subspace.

Theorem 2.3.4. *NIPALS obtains an orthonormal basis* $\{\phi_\ell\}_{\ell=1}^{L}$ *that spans the Krylov subspace* $\mathrm{Kry}_L(\mathcal{K}, \gamma)$.

Proof. First, note that, by construction (line 3 of Algorithm 1), the basis vectors have norm one. Considering this and Theorem 2.3.3, we only need to show that $\{\phi_\ell\}$ are orthogonal. Let us denote $c_i = \|\mathbb{E}(X_{i-1}Y)\|^{-1}$. WLOG, we assume $i < j$ and expand the product:

$$\begin{aligned}
\langle \phi_i, \phi_j \rangle &= \langle \phi_i, c_j \mathbb{E}(X_{j-1}Y) \rangle = \\
&= \mathbb{E} \langle \phi_i, c_j X_{j-1} Y \rangle = \\
&= \mathbb{E} \left\langle \phi_i, c_j \left(X_{i-1} - \sum_{k=i}^{j-1} T_k p_k \right) Y \right\rangle \\
&= \mathbb{E} \langle \phi_i, c_j X_{i-1} Y \rangle - \sum_{k=i}^{j-1} \mathbb{E} \langle \phi_i, c_j T_k p_k Y \rangle = \\
&= c_j c_i^{-1} \langle \phi_i, \phi_i \rangle - \sum_{k=i}^{j-1} c_j \mathbb{E} \left(T_k Y \langle \phi_i, p_k \rangle \right) = \\
&= c_j c_i^{-1} - c_j \mathbb{E} \left(T_i Y \langle \phi_i, p_i \rangle \right) = \\
&= c_j c_i^{-1} - c_j \mathbb{E} \left(T_i Y \right) = \\
&= c_j c_i^{-1} - c_j \mathbb{E} \left(\langle \phi_i, X_{i-1} \rangle Y \right) = \\
&= c_j c_i^{-1} - c_j c_i^{-1} \mathbb{E} \left(\langle \phi_i, \phi_i \rangle \right) = 0.
\end{aligned}$$

□

Using this orthogonal basis, it is already possible to calculate the PLS approximation by applying the formulas derived in Section 2.2.1. However, for completeness, the following theorem shows that the scores obtained by NIPALS are the PLS scores, as defined in Section 2.2.3.

Theorem 2.3.5. *The scores obtained by NIPALS can be expressed as projections of* X *onto a conjugate set of directions that span the Krylov space* $\mathrm{Kry}_\ell(\mathcal{K}, \gamma)$ *with respect to* $\mathcal{K}$.

Proof. Substituting repeatedly the deflation step of NIPALS, we can express T_ℓ as

$$T_\ell = \langle X_{\ell-1}, \phi_\ell \rangle = \left\langle X - \sum_{i=1}^{\ell-1} T_i p_i, \phi_\ell \right\rangle = \langle X, \phi_\ell \rangle - \sum_{i=1}^{\ell-1} T_i \langle p_i, \phi_\ell \rangle . \tag{2.25}$$

We now proceed by induction. From the algorithm, it is immediate that $T_1 = \langle X, \phi_1 \rangle$,

and we have already proven that $\phi_1 \in \mathrm{Kry}_1(\mathcal{K}, \gamma)$. We assume that the result holds true for all scores up to $T_{\ell-1}$. That is to say

$$T_i = \langle X, v_i \rangle, \quad v_i \in \mathrm{Kry}_i(\mathcal{K}, \gamma), \qquad i = 1, \ldots, \ell - 1.$$

By substituting the inductive hypothesis into (2.25), one gets

$$T_\ell = \langle X, \phi_\ell \rangle - \sum_{i=1}^{\ell-1} \langle X, v_i \rangle \langle p_i, \phi_\ell \rangle .$$

Finally, if we reorder the terms and group them inside the inner product, we get

$$T_\ell = \left\langle X, \phi_\ell - \sum_{i=1}^{\ell-1} v_i \langle p_i, \phi_\ell \rangle \right\rangle .$$

Therefore, $T_\ell = \langle X, v_\ell \rangle$, where $v_\ell = \phi_\ell - \sum_{i=1}^{\ell-1} v_i \langle p_i, \phi_\ell \rangle$. This proves that the scores can be expressed as projections of X onto the Krylov subspace. Moreover, since the scores are uncorrelated, these projection directions must be conjugate with respect to the covariance operator. □

The following corollary contains a formula to calculate the conjugate directions that define the scores.

Corollary 2.3.1. *The directions v_ℓ that fulfill $T_\ell = \langle X, v_\ell \rangle$ can be calculated as:*

1. $v_\ell = \phi_\ell - \sum_{i=1}^{\ell-1} v_i \langle p_i, \phi_\ell \rangle$, *and* $v_1 = \phi_1$.
2. $(v_1, \ldots, v_\ell) = (\phi_1, \ldots, \phi_\ell) \left((\langle p_i, \phi_j \rangle)_{1<i,j<L} \right)^{-1}$.

Proof. The first expression is a direct consequence of the previous proof. To prove the second one, consider the matrix $(\langle p_i, \phi_j \rangle)_{1<i,j<L}$:

$$\begin{pmatrix} \langle p_1, \phi_1 \rangle & \langle p_1, \phi_2 \rangle & \ldots & \langle p_1, \phi_\ell \rangle \\ \langle p_2, \phi_1 \rangle & \langle p_2, \phi_2 \rangle & \ldots & \langle p_2, \phi_\ell \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle p_\ell, \phi_1 \rangle & \langle p_L, \phi_2 \rangle & \ldots & \langle p_L, \phi_\ell \rangle \end{pmatrix} = \begin{pmatrix} 1 & \langle p_1, \phi_2 \rangle & \ldots & \langle p_1, \phi_\ell \rangle \\ 0 & 1 & \ldots & \langle p_2, \phi_\ell \rangle \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix},$$

where we have applied (2.24) and the definition of p_ℓ in NIPALS.

We can now multiply this matrix by $(v_1, \ldots, v_\ell)$:

$$\begin{aligned}
&\begin{pmatrix} v_1 & v_2 & \ldots & v_\ell \end{pmatrix} \begin{pmatrix} 1 & \langle p_1, \phi_2 \rangle & \ldots & \langle p_1, \phi_\ell \rangle \\ 0 & 1 & \ldots & \langle p_2, \phi_\ell \rangle \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} = \\
&= \begin{pmatrix} v_1 & v_2 + v_1 \langle p_1, \phi_2 \rangle & \ldots & v_\ell + \sum_{i=1}^{\ell-1} v_i \langle p_i, \phi_\ell \rangle \end{pmatrix} = \\
&= \begin{pmatrix} \phi_1 & \phi_2 & \ldots & \phi_\ell \end{pmatrix},
\end{aligned}$$

where the last step holds due to the expression for v_ℓ. Therefore, by inverting the $(\langle p_i, \phi_j \rangle)_{1<i,j<L}$ matrix, we obtain the desired result. Moreover, we know that the matrix is invertible since it is upper-triangular and its diagonal does not contain any zeros. □

During this section, we have explored the properties of NIPALS. This algorithm calculates an orthogonal basis of the Krylov subspace, alongside the PLS scores and loadings in an iterative fashion. However, it does not calculate the PLS approximation or the conjugate projection directions for the PLS scores directly.

Therefore, to obtain the PLS approximation, either a least squares fit must be performed (as described in Section 2.2.1) or the conjugate directions must be obtained (as described in Corollary 2.3.1). In both cases, an additional matrix inversion of an $L \times L$ matrix is required. Usually, L takes a relatively low value and, thus this is not an issue. Nevertheless, if we want to calculate all PLS approximations up to a specified number of components, we will need L matrix inversions. This process will be at least of $\mathcal{O}(L^{3.373})$ (Williams, 2014), possibly reaching $\mathcal{O}(L^4)$ depending on the matrix inversion method being used. Therefore, it can become a bottleneck when dealing with a high number of components. In the following section, we cover the conjugate gradient method, which does not present this drawback.

2.3.2 Conjugate Gradients

The conjugate gradient method is an iterative algorithm that minimizes a quadratic form by exploring Krylov spaces of increasing order. This method was introduced in Hestenes and Stiefel (1952), while Nocedal and Wright (1999) provides a summary of its properties. Since the PLS estimation can be characterized as a least squares approximation restricted to a Krylov subspace (2.13), $\beta_L^{(\mathrm{PLS})}$ can be calculated in L iterations of the conjugate gradient algorithm.

This method extracts a conjugate basis of the Krylov subspace, while obtaining both the PLS approximation to β^* and an additional orthogonal basis. By utilizing the properties described in Section 2.2.3, the PLS approximation to β^* for each number of components is calculated directly in each iteration.

Algorithm 2 Conjugate Gradient Algorithm

Input: X,Y and L
Output: $\{\beta_\ell\}_{l=1}^{L}$ and $\{\varphi_\ell\}_{l=1}^{L}$

$g_0 \leftarrow -\gamma$
$\varphi_0 \leftarrow \gamma$
$\beta_0 \leftarrow 0$
for $l = 1, \ldots, L$ **do**
  $\alpha_\ell \leftarrow -\frac{\langle g_{\ell-1}, \varphi_{\ell-1}\rangle}{\langle \varphi_{\ell-1}, \mathcal{K}\varphi_{\ell-1}\rangle}$ ▷ Calculate step size
  $\beta_l \leftarrow \beta_{\ell-1} + \alpha_\ell \varphi_{\ell-1}$ ▷ Calculate next coefficient
  $g_\ell \leftarrow \mathcal{K}\beta_\ell - \gamma$ ▷ Compute the gradient
  $\gamma_\ell \leftarrow \frac{\langle g_\ell, \mathcal{K}\varphi_{\ell-1}\rangle}{\langle \varphi_{\ell-1}, \mathcal{K}\varphi_{\ell-1}\rangle}$ ▷ Step size for the conjugate direction update
  $\varphi_\ell \leftarrow -g_\ell + \gamma_\ell \varphi_{\ell-1}$ ▷ Next conjugate direction
end for

Similarly to NIPALS, our first goal is to show that the directions obtained constitute a conjugate basis of the Krylov subspace. However, before starting, we need to state a technical lemma, to be used in the subsequent proofs.

Lemma 2.3.6. *From Algorithm 2, we can deduce that* $g_\ell = g_{\ell-1} + \alpha_\ell \mathcal{K}\varphi_{\ell-1}$.

Proof. Expanding the expression for g_ℓ in the algorithm:

$$g_\ell = \mathcal{K}\beta_\ell - \gamma = \mathcal{K}(\beta_{\ell-1} + \alpha_\ell\varphi_{\ell-1}) - \gamma = \mathcal{K}\beta_{\ell-1} - \gamma + \mathcal{K}\alpha_\ell\varphi_{\ell-1} = \\ = g_{\ell-1} + \mathcal{K}\alpha_\ell\varphi_{\ell-1}.$$

□

Using the previous result, we can now prove the main properties of the conjugate gradient method.

Theorem 2.3.7. *Assuming* $\beta_\ell \neq \beta_{\ell+1}$, *the following conditions hold:*

1. g_ℓ *is orthogonal to* $\operatorname{span}\{\varphi_0, \ldots, \varphi_{\ell-1}\}$.
2. $\langle \varphi_\ell, \mathcal{K}\varphi_i\rangle = 0$ *for* $i < \ell$.
3. $\operatorname{span}\{\varphi_0, \ldots, \varphi_\ell\} = \operatorname{Kry}_{\ell+1}(\mathcal{K}, \gamma)$.
4. $\operatorname{span}\{g_0, \ldots, g_\ell\} = \operatorname{Kry}_{\ell+1}(\mathcal{K}, \gamma)$.

Proof. We will prove these statements at the same time by induction. For $\ell = 0$, they are trivial. Therefore, we assume the statements hold for $\ell - 1$ and prove them for ℓ.

Statement 1

Let $i < \ell - 1$, then

$$\langle \varphi_i, g_\ell \rangle = \langle \varphi_i, g_{\ell-1} + \alpha_\ell \mathcal{K} \varphi_{\ell-1} \rangle = \langle \varphi_i, g_{\ell-1} \rangle + \alpha_\ell \langle \varphi_i, \mathcal{K} \varphi_{\ell-1} \rangle = 0,$$

where both terms are 0 due to the inductive hypothesis. Otherwise, if $i = \ell - 1$, then

$$\begin{aligned} \langle \varphi_{\ell-1}, g_l \rangle &= \langle \varphi_{\ell-1}, g_{\ell-1} \rangle + \alpha_\ell \langle \varphi_{\ell-1}, \mathcal{K} \varphi_{\ell-1} \rangle = \\ &= \langle \varphi_{\ell-1}, g_{\ell-1} \rangle - \frac{\langle g_{\ell-1}, \varphi_{\ell-1} \rangle}{\langle \varphi_{\ell-1}, \mathcal{K} \varphi_{\ell-1} \rangle} \langle \varphi_{\ell-1}, \mathcal{K} \varphi_{\ell-1} \rangle = \\ &= 0. \end{aligned}$$

Statement 2

To prove this statement, we expand the inner product as

$$\langle \varphi_\ell, \mathcal{K} \varphi_i \rangle = \langle -g_\ell, \mathcal{K} \varphi_i \rangle + \gamma_\ell \langle \varphi_{\ell-1}, \mathcal{K} \varphi_i \rangle \tag{2.26}$$

If $i < \ell - 1$, $\varphi_i \in \mathrm{Kry}_{i+1}(\mathcal{K}, \gamma)$. Therefore,

$$\mathcal{K} \varphi_i \in \mathrm{Kry}_{i+2}(\mathcal{K}, \gamma) = \mathrm{span}\{\varphi_0, \ldots, \varphi_{i+1}\} \subset \mathrm{span}\{\varphi_0, \ldots, \varphi_{\ell-1}\} \perp g_\ell,$$

and the first term is 0. The second term is also zero due to the second inductive hypothesis.

If $i = \ell - 1$, (2.26) is 0 due to the definition of γ_ℓ

$$\langle \varphi_\ell, \mathcal{K} \varphi_{\ell-1} \rangle = \langle -g_\ell, \mathcal{K} \varphi_{\ell-1} \rangle + \frac{\langle g_\ell, \mathcal{K} \varphi_{\ell-1} \rangle}{\langle \varphi_{\ell-1}, \mathcal{K} \varphi_{\ell-1} \rangle} \langle \varphi_{\ell-1}, \mathcal{K} \varphi_{\ell-1} \rangle = 0.$$

Statement 4

We can now prove the fourth statement. This is a direct consequence of the expression in Proposition 2.3.6: $g_\ell = g_{\ell-1} + \alpha_\ell \mathcal{K} \varphi_{\ell-1}$. Due to the inductive hypothesis for the fourth statement, $\varphi_{\ell-1} \in \mathrm{Kry}_\ell(\mathcal{K}, \gamma)$. Therefore, $\mathcal{K} \varphi_{\ell-1} \in \mathrm{Kry}_{\ell+1}(\mathcal{K}, \gamma)$. As a result, we have that $\mathrm{span}\{g_0, \ldots g_\ell\} \subseteq \mathrm{Kry}_{\ell+1}(\mathcal{K}, \gamma)$. But, since g_ℓ is orthogonal to $\mathrm{span}\{g_0, \ldots, g_{\ell-1}\} = \mathrm{Kry}_\ell(\mathcal{K}, \gamma)$, the equality must hold.

Statement 3

Finally, we can prove the third statement. This is a direct consequence of the fourth and second statement. Since $\varphi_\ell = -g_\ell + \gamma_\ell \varphi_{\ell-1}$, $\mathrm{span}\{\varphi_0, \ldots \varphi_\ell\} \subset \mathrm{Kry}_{\ell+1}(\mathcal{K}, \gamma)$. Moreover, since $\{\varphi_i\}_{i=0}^{\ell}$ are conjugate directions with respect to a positive definite operator, the dimension of the space must be $\ell + 1$ and the equality must hold □

The previous theorem implies the following result regarding the bases of the Krylov subspace obtained by the conjugate gradient algorithm.

Corollary 2.3.2. *The conjugate gradient algorithm provides two bases of the Krylov subspace* $\mathrm{Kry}_\ell(\mathcal{K}, \gamma)$*:*

- *A* $\mathcal{K}$*-conjugate basis* $\{\varphi_\ell\}_{l=0}^{\ell-1}$.
- *An orthogonal basis* $\{g_\ell\}_{l=0}^{\ell-1}$.

Proof. The first claim is a direct consequence of statements 2 and 3 of the previous Theorem. To prove the second, it suffices to consider statements 1, 3 and 4, which imply that g_ℓ is orthogonal to the previous gradients. Thus, it is an orthogonal base. □

The last property that we want to present is that this algorithm calculates the PLS approximation directly, unlike NIPALS, where $\beta_L^{(\mathrm{PLS})}$ has to be computed separately. Moreover, all the PLS approximations, for each number of components up to the specified number of iterations (L), are obtained in a single execution of the algorithm.

Theorem 2.3.8. *The sequence* $\{\beta_\ell\}_{l=1}^{L}$ *coincides with the PLS approximation of* β^* *for each number of components.*

Proof. Line 6 of the Algorithm 2 builds the PLS approximation iteratively, therefore, as an expansion on the conjugate basis. Therefore, it suffices to show that the coefficients $\{\alpha_\ell\}_{\ell=1}^{L}$ correspond to $\{b_\ell^{(\varphi)}\}_{\ell=1}^{L}$ from (2.17). However, this is immediate once we compare the expression obtained for $b_\ell^{(\varphi)}$ in Proposition 2.2.9 with lines 5 and 7 of the conjugate gradient algorithm. □

In this section, we have explored the properties of the conjugate gradient method, which centers around the construction of a conjugate basis of the Krylov subspace. This approach has several advantages with respect to the orthogonal basis in the previous section.

In particular, the PLS approximations are calculated in the algorithm itself, without the need of any additional steps. This is possible since the coordinates of $\beta_L^{(\mathrm{PLS})}$ on the conjugate basis do not have to be completely recalculated at each step (see Proposition 2.2.8).

Chapter 3

Multivariate Regression: Relationship between PLS and OLS regression

In this chapter, we focus on the application of PLS to multiple linear regression problems. We already showed that the PLS approximation can be defined as the solution of a least squares problem restricted to a Krylov subspace in Section 2.2.2. Therefore, there is a close relationship between PLS and ordinary least squares (OLS), as the only difference between them is the restriction of the coefficient search space. In this chapter, we compare the coefficients estimated by both methods, studying the difference between them as the number of components considered by PLS increases.

3.1 Partial least squares on a sample

During this analysis, we will consider a sample of N observations of the regressor variables and the response variable: $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$. As usual, the observations will be grouped row-wise into a matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top \in \mathbb{R}^{N \times D}$, and a vector $\mathbf{y} = (y, \ldots, y_N)^\top \in \mathbb{R}^N$.

Additionally, we will utilize the usual notation in most of the multivariate PLS literature, such as Rosipal and Krämer (2005) or de Jong (1993). In particular, the orthogonal PLS basis is denoted $\{\mathbf{w}_\ell\}_{\ell=1}^L$, and, the conjugate PLS basis, $\{\mathbf{r}_\ell\}_{\ell=1}^L$. Since we are working on a sample space, each PLS component is also a vector (with N observations): $\mathbf{t}_\ell \in \mathbb{R}^N$ for $\ell = 1, \ldots, L$. Finally, the covariance operator $\mathcal{K}$ and the cross covariance γ are now replaced by their sample estimators: $\mathbf{\Sigma}_{\text{XX}} = \mathbf{X}^\top\mathbf{X}$ and $\mathbf{\Sigma}_{\text{XY}} = \mathbf{X}^\top\mathbf{y}$.

Considering these changes, the PLS optimization problem in (2.16) can be rewritten as

$$\mathbf{t}_\ell = \underset{\mathbf{t}}{\operatorname{argmax}}\ \mathbf{t}^\top\mathbf{y} \quad \text{subject to} \quad \mathbf{t} = \mathbf{X}\mathbf{r}, \quad \mathbf{r} \in \mathbb{R}^D, \quad \|\mathbf{r}\| = 1;$$
$$\mathbf{t}^\top\mathbf{t}_i = 0 \quad i \in \{1, \ldots, \ell - 1\}.$$

Associated with the components, the weight vectors $\{\mathbf{r}_\ell\}_{\ell=1}^L$ are defined so that $\|\mathbf{r}_\ell\| = 1$ and $\mathbf{t}_\ell = \mathbf{X}\mathbf{r}_\ell$, for $\ell \in \{1, \ldots, L\}$. These correspond to the conjugate PLS basis described in the

previous chapter.

The PLS properties and identities presented in the previous chapter can be easily extended to apply them to the sampled values, instead of the random variables (see Table 3.1). However, the multivariate nature of the problems considered in this chapter allows us to simplify some of the results.

Regressor covariance	$\mathcal{K}$	$\mathbf{\Sigma}_{\mathrm{XX}} = \mathbf{X}^\top \mathbf{X}$
Cross covariance	γ	$\mathbf{\Sigma}_{\mathrm{XY}} = \mathbf{X}^\top \mathbf{y}$
Orthogonal basis	$\{\phi_\ell\}_{\ell=1}$	$\{\mathbf{w}_\ell\}_{\ell=1}^L$
Conjugate basis	$\{\varphi_\ell\}_{\ell=1}$	$\{\mathbf{r}_\ell\}_{\ell=1}^L$

Table 3.1: Population quantities, along with their sample estimators.

For once, it is possible to rewrite the NIPALS algorithm, as included in Algorithm 3 to take $\mathbf{X}$ and $\mathbf{y}$ as inputs. This version of the algorithm is the most popular in PLS literature (Höskuldsson, 1988; Rosipal & Krämer, 2005), but we have omitted the calculation of the projection directions, scores, and loadings for the target variable, since they are not needed for the scalar response case.

Algorithm 3 Sample NIPALS for PLS regression with scalar response

Input: $\mathbf{X}$: the regressor variable data matrix.
$\mathbf{y}$: the response variable data vector.
L: the number of components to extract.

Output: $\{\mathbf{w}_\ell\}_{\ell=1}^L$: projection weights.
$\{\mathbf{t}_\ell\}_{\ell=1}^L$: components.
$\{\mathbf{p}_\ell\}_{\ell=1}^L$: loadings.

1: $\mathbf{X}_0 \leftarrow \mathbf{X}$
2: $l \leftarrow 1$
3: **while** $\ell < L$ **do**
4: $\quad \mathbf{w}_\ell \leftarrow \mathbf{X}_{\ell-1}^\top \mathbf{y} / \|\mathbf{X}_{\ell-1}^\top \mathbf{y}\|$ ▷ *Weights* calculation
5: $\quad \mathbf{t}_\ell \leftarrow \mathbf{X}_{\ell-1} \mathbf{w}_\ell$ ▷ *Scores* calculation
6: $\quad \mathbf{p}_\ell \leftarrow \mathbf{X}_{\ell-1}^\top \mathbf{t}_\ell / (\mathbf{t}_\ell^\top \mathbf{t}_\ell)$ ▷ *Loadings* calculation
7: $\quad \mathbf{X}_\ell \leftarrow \mathbf{X}_{\ell-1} - \mathbf{t}_\ell \mathbf{p}_\ell^\top$ ▷ Deflate $\mathbf{X}$
8: $\quad l \leftarrow l + 1$
9: **end while**

At the end of each iteration, $\mathbf{X}$, the data matrix, is modified by removing the projection on the component computed in that iteration (line 7 of Algorithm 3). As a result, a sequence of projections of the data matrix can be considered: $\{\mathbf{X}_\ell\}_{\ell=1}^L$. This deflation step ensures that subsequent components computed by the algorithm are orthogonal to the ones extracted up to that point. In particular, lines 6 and 7 of Algorithm 3 show that the deflation can be interpreted as an orthogonal projection:

$$\mathbf{X}_\ell = \left(\mathbf{I} - \frac{\mathbf{t}_\ell \mathbf{t}_\ell^\top}{\mathbf{t}_\ell^\top \mathbf{t}_\ell} \right) \mathbf{X}_{\ell-1} \quad \ell \in \{1, \ldots, L\}.$$

Aside from projections of the original data (the scores), in the algorithm, the weights $\{\mathbf{w}_\ell\}_{\ell=1}^L$ are calculated. These vectors correspond to the orthogonal PLS basis defined in

Section 2.2.1. Note that these vectors are different from $\{\mathbf{r}_\ell\}_{\ell=1}^L$, but both sets of vectors can be utilized to calculate the PLS components. In particular, The ℓ-th component can be computed as $\mathbf{t}_\ell = \mathbf{X}\mathbf{r}_\ell$, the projection of the original data onto the ℓ-th element of the conjugate basis. Alternatively, it is $\mathbf{t}_\ell = \mathbf{X}_{\ell-1}\mathbf{w}_\ell$, where $\mathbf{w}_\ell$ is the ℓ-th weight vector extracted by NIPALS. Additionally, the regressor and response loadings can be calculated from the quantities obtained in NIPALS as $\mathbf{p}_\ell = \mathbf{X}_{\ell-1}^\top \mathbf{t}_\ell / \|\mathbf{t}_\ell\|^2$ and $q_\ell = \mathbf{y}^\top \mathbf{t}_\ell / \|\mathbf{t}_\ell\|^2$, for $\ell \in \{1, \dots, L\}$, respectively.

The following propositions summarize the properties of the quantities obtained in NIPALS that are relevant for the analysis presented in the following sections

Proposition 3.1.1. *From the NIPALS algorithm, the following properties can be derived:*

1. *In terms of the PLS components, the original data can be expressed as*

$$\mathbf{X} = \mathbf{T}_\ell \mathbf{P}_L^\top + \mathbf{X}_L, \qquad \mathbf{y} = \mathbf{T}_\ell \mathbf{Q}_L^\top + \mathbf{y}_L, \tag{3.1}$$

   where $\mathbf{X}_L \in \mathbb{R}^{N\times D}$ and $\mathbf{y}_L \in \mathbb{R}^N$ are defined as

$$\mathbf{X}_L = \prod_{i=1}^{L} \left(\mathbf{I} - \frac{\mathbf{t}_i \mathbf{t}_i^\top}{\mathbf{t}_i^\top \mathbf{t}_i} \right) \mathbf{X}, \qquad \mathbf{y}_L = \prod_{i=1}^{L} \left(\mathbf{I} - \frac{\mathbf{t}_i \mathbf{t}_i^\top}{\mathbf{t}_i^\top \mathbf{t}_i} \right) \mathbf{y}. \tag{3.2}$$

   Additionally, $\mathbf{T}_L$, $\mathbf{P}_L$ and $\mathbf{Q}_L$ are defined as

$$\mathbf{T}_L = (\mathbf{t}_1, \dots \mathbf{t}_L) \in \mathbb{R}^{N\times L}, \quad \mathbf{P}_L = (\mathbf{p}_1, \dots, \mathbf{p}_L) \in \mathbb{R}^{D\times L}, \quad \mathbf{Q}_L = (q_1, \dots, q_L) \in \mathbb{R}^{1\times L}.$$

2. *The Frobenius norms of $\mathbf{X}_\ell$ and $\mathbf{y}_L$ decrease as L increases.*

3. *After L iterations, $\mathbf{X}_L$ is orthogonal to the weights: $\mathbf{X}_L \mathbf{W}_L = \mathbf{0}$, where*

$$\mathbf{W}_L = (\mathbf{w}_1, \dots \mathbf{w}_L) \in \mathbb{R}^{M\times L}.$$

4. *The loading matrices $\mathbf{P}_L$ and $\mathbf{Q}_L$ can be expressed in terms of the components and the original data as $\mathbf{P}_L = \mathbf{X}^\top \mathbf{T}_L \mathbf{D}_L^{-2}$ and $\mathbf{Q}_L = \mathbf{y}^\top \mathbf{T}_L \mathbf{D}_\ell^{-2}$, with*

$$\mathbf{D}_L = \operatorname{diag}(\|\mathbf{t}_1\|, \dots, \|\mathbf{t}_L\|) \in \mathbb{R}^{L\times L}.$$

Proof. We prove the properties in order:

1. The identity for $\mathbf{X}_\ell$ in (3.2) is a direct consequence of substituting line 6 into line 7 of Algorithm 3. The corresponding identity for $\mathbf{y}$ can be derived similarly once one notices that adding a deflation step for $\mathbf{y}$ at the end of each iteration would not affect the results of NIPALS. The deflation step for $\mathbf{y}$ would be $\mathbf{y}_\ell = \mathbf{y}_{\ell-1} - \mathbf{t}_\ell q_\ell$. Therefore, the deflated $\mathbf{y}$ at step ℓ could be calculated as

$$\mathbf{y}_\ell = \mathbf{y} - \sum_{i=1}^{\ell} \mathbf{t}_i q_i. \tag{3.3}$$

   Since $\mathbf{y}$ is only used in the calculation of $\mathbf{w}_\ell$, it suffices to check that adding a deflation step would not alter this calculation. After adding the deflation step,

$\mathbf{w}_\ell$ would be calculated as $\mathbf{w}_\ell = \mathbf{X}_{\ell-1}^\top \mathbf{y}_{\ell-1} / \|\mathbf{X}_{\ell-1}^\top \mathbf{y}_{\ell-1}\|$. However, if we substitute (3.3):

$$\mathbf{X}_{\ell-1}^\top \mathbf{y}_{\ell-1} = \mathbf{X}_{\ell-1}^\top \left(\mathbf{y} - \sum_{i=1}^{\ell-1} \mathbf{t}_i q_i \right) = \mathbf{X}_{\ell-1}^\top \mathbf{y} - \sum_{i=1}^{\ell-1} \mathbf{X}_{\ell-1}^\top \mathbf{t}_i q_i,$$

and $\mathbf{X}_{\ell-1}^\top \mathbf{t}_i = \mathbf{0}$ as long as $i < \ell$. Therefore, the second term is zero, and we have shown that adding a deflation step for $\mathbf{y}$ would not alter the results of the algorithm.

2. The decrease of the Frobenius norm is a consequence of the expressions for $\mathbf{X}_\ell$ and $\mathbf{y}_\ell$ in (3.2). We will prove the result for $\mathbf{X}_\ell$. From (3.2), one obtains

$$\mathbf{X}_\ell = \underbrace{\left(\mathbf{I} - \frac{\mathbf{t}_\ell \mathbf{t}_\ell^\top}{\mathbf{t}_\ell^\top \mathbf{t}_\ell} \right)}_{\mathbf{\Pi}_\ell} \mathbf{X}_{\ell-1}. \tag{3.4}$$

Then, to show the decrement of the norms, we need only show that $\|\mathbf{X}_\ell\|_F \leq \|\mathbf{X}_{\ell-1}\|_F$ for $1 \leq \ell < L$. Using (3.4),

$$\begin{aligned} \|\mathbf{X}_\ell\|_F = \|\mathbf{\Pi}_\ell \mathbf{X}_{\ell-1}\|_F = \|\mathbf{U}_\ell \mathbf{S}_\ell \mathbf{U}_\ell^\top \mathbf{X}_{\ell-1}\|_F = \|\mathbf{S}_\ell \mathbf{U}_\ell^\top \mathbf{X}_{\ell-1}\|_F \leq \\ \leq \|\mathbf{U}_\ell^\top \mathbf{X}_{\ell-1}\|_F = \|\mathbf{X}_{\ell-1}\|_F, \end{aligned}$$

where $\mathbf{\Pi}_\ell = \mathbf{U}_\ell \mathbf{S}_\ell \mathbf{U}_\ell^\top$ is the eigenvector decomposition of $\mathbf{\Pi}_\ell$. Since $\mathbf{\Pi}_\ell$ is a real symmetric matrix, $\mathbf{U}_\ell$ is a unitary matrix, and we can apply that the Frobenius norm is invariant under unitary operations. Additionally, since $\mathbf{\Pi}_\ell$ is positive-definite and idempotent, its eigenvalues are either 0 or 1. Therefore, $\mathbf{S}_\ell$ has only 0s or 1s in the diagonal. As a result, multiplying by it can only reduce the Frobenius norm.

3. The orthogonality between $\mathbf{X}_\ell$ and $\mathbf{W}$ can be proven showing that $\mathbf{X}_\ell \mathbf{w}_\ell = \mathbf{0}$ if $\ell \leq L$. From (3.2),

$$\begin{aligned} \mathbf{X}_\ell \mathbf{w}_\ell = \left(\mathbf{I} - \frac{\mathbf{t}_\ell \mathbf{t}_L^\top}{\mathbf{t}_\ell^\top \mathbf{t}_L} \right) \dots \left(\mathbf{I} - \frac{\mathbf{t}_\ell \mathbf{t}_\ell^\top}{\mathbf{t}_\ell^\top \mathbf{t}_\ell} \right) \mathbf{X}_{\ell-1} \mathbf{w}_\ell = \\ = \left(\mathbf{I} - \frac{\mathbf{t}_\ell \mathbf{t}_L^\top}{\mathbf{t}_\ell^\top \mathbf{t}_L} \right) \dots \left(\mathbf{I} - \frac{\mathbf{t}_\ell \mathbf{t}_\ell^\top}{\mathbf{t}_\ell^\top \mathbf{t}_\ell} \right) \mathbf{t}_\ell = \mathbf{0}. \end{aligned}$$

4. Regarding the expressions for the loadings, both identities can be proven in the same way. We will prove the identity for $\mathbf{P}$, the X loadings, showing the equality for each column of both sides of the equation. This equality is, in turn, a consequence

of the expression for $\mathbf{X}$ in (3.2).

$$\begin{aligned}\mathbf{X}^\top \mathbf{t}_\ell \|\mathbf{t}_\ell\|^{-2} &= (\mathbf{T}_{\ell-1}\mathbf{P}_{\ell-1})^\top \mathbf{t}_\ell \|\mathbf{t}_\ell\|^{-2} + \mathbf{X}_{\ell-1}^\top \mathbf{t}_\ell \|\mathbf{t}_\ell\|^{-2} = \\ &= \mathbf{P}_{\ell-1}^\top \mathbf{T}_{\ell-1}^\top \mathbf{t}_\ell \|\mathbf{t}_\ell\|^{-2} + \mathbf{X}_{\ell-1}^\top \mathbf{t}_\ell \|\mathbf{t}_\ell\|^{-2} = \\ &= \mathbf{p}_\ell,\end{aligned}$$

where $\mathbf{T}_{\ell-1}^\top \mathbf{t}_\ell = \mathbf{0}$ because the extracted components are orthogonal.

□

NIPALS calculates both the components and the weights needed to express the components as projections of the deflated $\mathbf{X}_\ell$ data matrices. However, if one wants to calculate the PLS components of new data utilizing the weights $\{\mathbf{w}_\ell\}_{\ell=1}^L$, the deflation process would have to be repeated. A better approach is to obtain the conjugate basis $\{\mathbf{r}_\ell\}_{\ell=1}^L$ first, and then calculate the PLS components of new data by projecting onto this basis. The following proposition provides an expression for the matrix $\mathbf{R}_L$, whose columns contain the elements of the conjugate basis.

Proposition 3.1.2. *The matrix* $\mathbf{R}_L \in \mathbb{R}^{D\times L}$ *that fulfills* $\mathbf{T}_L = \mathbf{X}\mathbf{R}_L$ *is* $\mathbf{R}_L = \mathbf{W}_L(\mathbf{P}_L^\top \mathbf{W}_L)^{-1}$.

Proof. From Proposition 3.1.1, $\mathbf{X}_L\mathbf{W}_L = \mathbf{0}$. Applying this to the decomposition for $\mathbf{X}$ in (3.1), we obtain:

$$\begin{aligned}\mathbf{X}\mathbf{R}_L &= (\mathbf{T}_L\mathbf{P}_L^\top + \mathbf{X}_L)(\mathbf{W}_L(\mathbf{P}_L^\top\mathbf{W}_L)^{-1}) \\ &= \mathbf{T}_L\mathbf{P}_L^\top\mathbf{W}_L(\mathbf{P}_L^\top\mathbf{W}_L)^{-1} + \mathbf{X}_L\mathbf{W}_L(\mathbf{P}_L^\top\mathbf{W}_L)^{-1} = \\ &= \mathbf{T}_L.\end{aligned}$$

□

3.2 Partial least squares regression on a sample

As in Chapter 2, we consider the linear regression model $Y = \boldsymbol{\beta}^\top X + \epsilon$. In this model, $X \in \mathbb{R}^D$ is the regressor vector, $\boldsymbol{\beta} \in \mathbb{R}^D$ is the vector of coefficient (which needs to be estimated), ϵ is random noise independent of X, and Y is the scalar response. For the sake of simplicity, and without loss of generality, both X and Y are assumed to have zero mean. To fit this model, N independent observations drawn from this model are available: $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$. We further assume that $\{\epsilon_i\}_{i=1}^N$ are iid with variance σ^2. In this setting, we seek to estimate a vector of coefficients $\boldsymbol{\beta}$ such that $y_i = \boldsymbol{\beta}^\top\mathbf{x}_i + \epsilon_i$, where $i \in \{1,\dots,N\}$. These equations can be grouped row-wise into the matrix equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.5}$$

where $\mathbf{y} = (y_1,\dots,y_N)^\top \in \mathbb{R}^N$, $\mathbf{X} = (\mathbf{x}_1,\dots,\mathbf{x}_N)^\top \in \mathbb{R}^{N\times D}$ and $\boldsymbol{\epsilon} = (\epsilon_1,\dots,\epsilon_N)^\top \in \mathbb{R}^N$.

One possible approximation for $\boldsymbol{\beta}$ is the ordinary least squares estimator (OLS)

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^D}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = {\boldsymbol{\Sigma}_{\text{XX}}}^{-1}\boldsymbol{\Sigma}_{\text{XY}}, \tag{3.6}$$

where $\|\cdot\|$ is the euclidean norm, $\mathbf{\Sigma}_{\text{XX}} = \mathbf{X}^\top\mathbf{X}$ is the empirical estimate of the covariance matrix of X scaled by the number of observations, and $\mathbf{\Sigma}_{\text{XY}} = \mathbf{X}^\top\mathbf{y}$ is the empirical estimate of the covariance matrix of X and Y scaled by the number of observations as well.

A different estimation of $\boldsymbol{\beta}$ is obtained using PLS regression. The first step is to extract L PLS components as described in the previous section. Then, a linear prediction is made in terms of these components: $\sum_{\ell=1}^{L} \hat{\gamma}_\ell^{(L)}\mathbf{t}_\ell = \mathbf{T}_\ell\hat{\boldsymbol{\gamma}}^{(L)}$, with $\hat{\boldsymbol{\gamma}}^{(L)} = \left(\hat{\gamma}_1^{(L)}, \ldots, \hat{\gamma}_\ell^{(L)}\right)^\top$ determined by least squares as

$$\hat{\boldsymbol{\gamma}}^{(L)} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^L}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{T}_\ell\boldsymbol{\gamma}\|^2 = (\mathbf{T}_\ell^\top\mathbf{T}_L)^{-1}\mathbf{T}_L^\top\mathbf{y} = \mathbf{D}_\ell^{-2}\mathbf{T}_L^\top\mathbf{y}.$$

PLS estimates the regression coefficient by expressing this linear predictor in terms of the original variables $\mathbf{T}_\ell\hat{\boldsymbol{\gamma}}^{(L)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{PLS}}^{(L)}$. Using the definition of $\mathbf{R}_\ell$ from Proposition 3.1.2: $\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{PLS}}^{(L)} = \mathbf{T}_\ell\hat{\boldsymbol{\gamma}}^{(L)} = \mathbf{X}\mathbf{R}_\ell\hat{\boldsymbol{\gamma}}^{(L)}$. Therefore, the PLS approximation of the vector of regression coefficients is

$$\hat{\boldsymbol{\beta}}_L^{(\text{PLS})} = \mathbf{R}_\ell\hat{\boldsymbol{\gamma}}^{(L)} = \mathbf{R}_\ell\mathbf{D}_L^{-2}\mathbf{T}_L^\top\mathbf{y}. \tag{3.7}$$

Alternatively, as introduced in Section 2.2.2, $\hat{\boldsymbol{\beta}}_L^{(\text{PLS})}$ can be viewed as the least squares approximation to $\boldsymbol{\beta}$ when the optimization is constrained to a Krylov subspace. In a multivariate space, a Krylov subspace is defined as follows:

Definition 3. *The Krylov subspace of order $L \leq D$ generated by the matrix $\mathbf{A} \in \mathbb{R}^{D\times D}$ and the vector $\mathbf{b} \in \mathbb{R}^D$, $\mathbf{b} \neq 0$ is*

$$\text{Kry}_\ell(\mathbf{A}, \mathbf{b}) = \text{span}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{L-1}\mathbf{b}\}.$$

Note that, so far in this chapter, we have introduced PLS regression as the result of least squares regression on the components extracted by NIPALS. The following proposition shows that the approximation obtained from the components coincides with the restriction of the least squares problem to the Krylov subspace.

Theorem 3.2.1. *The PLS approximation with L components defined in* (3.7) *is the solution to the least squares problem*

$$\hat{\boldsymbol{\beta}}_L^{(\text{PLS})} = \underset{\boldsymbol{\beta}\in\text{Kry}_\ell(\mathbf{\Sigma}_{XX},\mathbf{\Sigma}_{XY})}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2, \tag{3.8}$$

where $\text{Kry}_\ell(\mathbf{\Sigma}_{XX}, \mathbf{\Sigma}_{XY})$ is the Krylov subspace of order L generated by the matrix $\mathbf{\Sigma}_{XX}$ and the vector $\mathbf{\Sigma}_{XY}$.

Proof. Assume that the columns of $\mathbf{B}_\ell \in \mathbb{R}^{D\times L}$ constitute a basis of the Krylov subspace $\text{Kry}_\ell(\mathbf{\Sigma}_{\text{XX}}, \mathbf{\Sigma}_{\text{XY}})$. Then any $\boldsymbol{\beta} \in \text{Kry}_\ell(\mathbf{\Sigma}_{\text{XX}}, \mathbf{\Sigma}_{\text{XY}})$ can be expressed as $\boldsymbol{\beta} = \mathbf{B}_\ell\boldsymbol{\alpha}$ for some $\boldsymbol{\alpha} \in \mathbb{R}^L$. Thus, the constrained optimization problem given by (3.8) can be transformed into an unconstrained optimization problem in $\mathbb{R}^L$:

$$\underset{\boldsymbol{\beta}\in\text{Kry}_\ell(\mathbf{\Sigma}_{\text{XX}},\mathbf{\Sigma}_{\text{XY}})}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \underset{\boldsymbol{\alpha}\in\mathbb{R}^L}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\mathbf{B}_\ell\boldsymbol{\alpha}\|^2 = \mathbf{B}_\ell(\mathbf{B}_L^\top\mathbf{X}^\top\mathbf{X}\mathbf{B}_L)^{-1}\mathbf{B}_\ell^\top\mathbf{X}^\top\mathbf{y}. \tag{3.9}$$

As shown in Eldén (2004), the columns of the matrix $\mathbf{W}_\ell$ obtained after L iterations of NIPALS constitute a basis of $\mathrm{Kry}_\ell(\boldsymbol{\Sigma}_{\mathrm{XX}}, \boldsymbol{\Sigma}_{\mathrm{XY}})$. Therefore, (3.9) holds for $\mathbf{B}_\ell = \mathbf{W}_L$. It is then possible to show that $\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})}$ can be expressed in the form given by the rhs of (3.9) with $\mathbf{B}_\ell = \mathbf{W}_L$. To this end, Propositions 3.1.1 and 3.1.2 are applied repeatedly to (3.7):

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})} &= \mathbf{R}_\ell \mathbf{D}_\ell^{-2} \mathbf{T}_\ell^\top \mathbf{y} = \mathbf{W}_\ell (\mathbf{P}_\ell^\top \mathbf{W}_\ell^{-1}) \mathbf{D}_\ell^{-2} \mathbf{R}_\ell^\top \mathbf{X}^\top \mathbf{y} = \\
&= \mathbf{W}_\ell (\mathbf{P}_\ell^\top \mathbf{W}_\ell)^{-1} \mathbf{D}_\ell^{-2} (\mathbf{W}_\ell (\mathbf{P}_\ell^\top \mathbf{W}_\ell)^{-1})^\top \mathbf{X}^\top \mathbf{y} = \\
&= \mathbf{W}_\ell (\mathbf{W}_\ell^\top \mathbf{P}_\ell \mathbf{D}_\ell^2 \mathbf{P}_\ell^\top \mathbf{W}_\ell)^{-1} \mathbf{W}_\ell^\top \mathbf{X}^\top \mathbf{y} = \\
&= \mathbf{W}_\ell (\mathbf{W}_\ell^\top \mathbf{X}^\top \mathbf{T}_\ell \mathbf{P}_\ell^\top \mathbf{W}_\ell)^{-1} \mathbf{W}_\ell^\top \mathbf{X}^\top \mathbf{y} = \\
&= \mathbf{W}_\ell (\mathbf{W}_\ell^\top \mathbf{X}^\top (\mathbf{X} - \mathbf{X}_\ell) \mathbf{W}_\ell)^{-1} \mathbf{W}_\ell^\top \mathbf{X}^\top \mathbf{y} = \\
&= \mathbf{W}_\ell (\mathbf{W}_\ell^\top \mathbf{X}^\top \mathbf{X} \mathbf{W}_\ell)^{-1} \mathbf{W}_\ell^\top \mathbf{X}^\top \mathbf{y},
\end{aligned}$$

where the last step holds because of the orthogonality between $\mathbf{X}_\ell$ and $\mathbf{W}_\ell$ (Proposition 3.1.1). □

This proof makes explicit use of the properties of the matrices defined in the NIPALS algorithm. Other approaches can be adopted to derive this result. In Eldén (2004), a proof is given that is based on the relation of PLS with the Lanczos bidiagonalization algorithm. An alternative derivation is given in Takane and Loisel (2016), leveraging the properties of some bidiagonal and tridiagonal matrices in the NIPALS algorithm (Noonan & Wold, 1977). The approach followed in Chapter 2 is yet another alternative. In Chapter 2, the optimization problem was the starting point, and NIPALS was proved to yield equivalent quantities.

The expression of the vector of PLS regression coefficients given by (3.8) opens up the possibility of using numerical optimization algorithms that accept linear constraints to compute $\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})}$. It suffices to minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ subject to $\boldsymbol{\beta}$ belonging to $\mathrm{Kry}_\ell(\boldsymbol{\Sigma}_{\mathrm{XX}}, \boldsymbol{\Sigma}_{\mathrm{XY}})$. In particular, the conjugate gradient algorithm introduced in the previous chapter is an iterative algorithm that, in the L-th iteration, minimizes the quadratic form $\psi(\mathbf{z}) = \mathbf{z}^\top \mathbf{A}\mathbf{z} - \mathbf{b}^\top \mathbf{z}$ where the exploration is restricted to $\mathrm{Kry}_\ell(\mathbf{A}, \mathbf{b})$ (Nocedal & Wright, 1999). Thus, the optimization problem in Theorem 3.2.1 can be solved using the conjugate gradient algorithm with $\mathbf{A} = \boldsymbol{\Sigma}_{\mathrm{XX}}$ and $\mathbf{b} = \boldsymbol{\Sigma}_{\mathrm{XY}}$. In the next section, we take advantage of this property to study the evolution of difference between the PLS and OLS approximation to the regression coefficients, as a function of the number of components considered by PLS.

The following theorem establishes a link between the estimations obtained by both methods.

Theorem 3.2.2. *The OLS estimation of the regression coefficients* $\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}$ *is contained in* $\mathrm{Kry}_M(\boldsymbol{\Sigma}_{XX}, \boldsymbol{\Sigma}_{XY})$*, where* M *is the number of distinct eigenvalues of* $\boldsymbol{\Sigma}_{XX}$*.*

Proof. As a consequence of the Cayley-Hamilton theorem (Bronson & Costa, 2009, p.220), since $\mathbf{\Sigma}_{\mathrm{XX}}$ is a non-singular symmetric matrix, there exists a polynomial $P_{\mathbf{\Sigma}_{\mathrm{XX}}}$ of degree $M-1$ such that $P_{\mathbf{\Sigma}_{\mathrm{XX}}}(\mathbf{\Sigma}_{\mathrm{XX}})\mathbf{\Sigma}_{\mathrm{XX}} = \mathbf{I}$, where M is the number of different eigenvalues of $\mathbf{\Sigma}_{\mathrm{XX}}$. Applying this result to the usual formula of OLS, we obtain $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = P_{\mathbf{\Sigma}_{\mathrm{XX}}}(\mathbf{\Sigma}_{\mathrm{XX}})\mathbf{\Sigma}_{\mathrm{XY}} \in \mathcal{K}_M(\mathbf{\Sigma}_{\mathrm{XX}}, \mathbf{\Sigma}_{\mathrm{XY}})$. □

Corollary 3.2.1. *The PLS and OLS estimations of the regression coefficients coincide after M iterations, where M is the number of different eigenvalues of $\mathbf{\Sigma}_{XX}$:*

$$\hat{\boldsymbol{\beta}}^{(M)}_{\mathrm{PLS}} = \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}.$$

Proof. It is a direct consequence of Theorem 3.2.2 and the definition of $\hat{\boldsymbol{\beta}}^{(\mathrm{PLS})}_L$ as the solution of a restricted least squared problem in Theorem 3.2.1. □

3.3 Relation between partial least squares and ordinary least squares

As described in the previous section, the PLS estimation of the vector of coefficients of a linear regression model with L components converges to the ordinarly least squares estimator as L increases. Furthermore, they coincide when $L \geq M$, the number of distinct eigenvalues of $\mathbf{\Sigma}_{\mathrm{XX}}$. The goal of this section is to provide an upper bound for the distance between $\hat{\boldsymbol{\beta}}^{(\mathrm{PLS})}_L$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}$. In order to do so, we take advantage of the formulation of PLS in Theorem 3.2.1, as a constrained optimization problem that can be solved using conjugate gradients. The first part of this section follows the convergence analysis for the conjugate gradient method presented in Nocedal and Wright (1999). First, the PLS estimation is defined as the solution of yet another optimization problem, in which a distance to the OLS estimator is minimized subject to some constrains.

Proposition 3.3.1. *The PLS approximation to the vector of coefficients of a linear regression model with L components is the solution to the optimization problem*

$$\hat{\boldsymbol{\beta}}^{(\mathrm{PLS})}_L = \underset{\boldsymbol{\beta}\in \mathrm{Kry}_L(\mathbf{\Sigma}_{XX}, \mathbf{\Sigma}_{XY})}{\operatorname{argmin}} \left\| \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \right\|^2_{\mathbf{\Sigma}_{XX}}, \tag{3.10}$$

where $\|\mathbf{z}\|^2_{\mathbf{\Sigma}_{XX}} = \mathbf{z}^\top\mathbf{\Sigma}_{XX}\mathbf{z}$, the square of the quadratic-form norm with the positive definite matrix $\mathbf{\Sigma}_{XX}$.

Proof. This result is a consequence of the definition of the PLS approximation with L

components provided in Theorem 3.2.1:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})} &= \underset{\boldsymbol{\beta}\in \mathrm{Kry}_L(\boldsymbol{\Sigma}_{\mathrm{XX}},\boldsymbol{\Sigma}_{\mathrm{XY}})}{\operatorname{argmin}} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 = \\
&= \underset{\boldsymbol{\beta}\in \mathrm{Kry}_L(\boldsymbol{\Sigma}_{\mathrm{XX}},\boldsymbol{\Sigma}_{\mathrm{XY}})}{\operatorname{argmin}} \left(\mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\right) = \\
&= \underset{\boldsymbol{\beta}\in \mathrm{Kry}_L(\boldsymbol{\Sigma}_{\mathrm{XX}},\boldsymbol{\Sigma}_{\mathrm{XY}})}{\operatorname{argmin}} \left(\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y}\right) = \\
&= \underset{\boldsymbol{\beta}\in \mathrm{Kry}_L(\boldsymbol{\Sigma}_{\mathrm{XX}},\boldsymbol{\Sigma}_{\mathrm{XY}})}{\operatorname{argmin}} \left(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\mathrm{XX}}\,\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\mathrm{XX}}\,\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})} + (\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})})^\top\boldsymbol{\Sigma}_{\mathrm{XX}}\,\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}\right) = \\
&= \underset{\boldsymbol{\beta}\in \mathrm{Kry}_L(\boldsymbol{\Sigma}_{\mathrm{XX}},\boldsymbol{\Sigma}_{\mathrm{XY}})}{\operatorname{argmin}} \left\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}\right\|^2_{\boldsymbol{\Sigma}_{\mathrm{XX}}},
\end{aligned}$$

where we have used that $\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})} = \boldsymbol{\Sigma}_{\mathrm{XX}}\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}$. □

The quadratic-form norm $\|\cdot\|_{\boldsymbol{\Sigma}_{\mathrm{XX}}}$ is related to the Mahalanobis distance with the covariance matrix of the OLS estimator of $\boldsymbol{\beta}$. The following observation motivates the use of this norm as a natural way to quantify the differences between $\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}$.

Corollary 3.3.1. *The PLS estimation of the vector of coefficients of a linear regression model with L components is the solution of the optimization problem*

$$\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})} = \underset{\boldsymbol{\beta}\in \mathrm{Kry}_L(\boldsymbol{\Sigma}_{XX},\boldsymbol{\Sigma}_{XY})}{\operatorname{argmin}} d_M(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}),$$

where d_M is the Mahalanobis distance with respect to the matrix $\frac{1}{\sigma^2}{\boldsymbol{\Sigma}_{XX}}^{-1}$, which is the covariance matrix of the OLS estimator of the regression coefficients conditioned to the observations of X.

Proof. From (3.6), the variance of the OLS estimator conditioned to $\mathbf{x}_1, \dots \mathbf{x}_N$ is $\mathbf{C}_{\mathrm{OLS}} = \mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}|\mathbf{x}_1, \breve{\mathrm{a}} \dots \mathbf{x}_N) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$, where we have used that the $\mathrm{var}(\mathbf{y}|\mathbf{x}_1, \dots \mathbf{x}_N) = \mathrm{var}(\boldsymbol{\epsilon})$, and that the observations of ϵ are iid random variables with variance σ^2. As a result, the squared Mahalanobis distance between the $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ estimator and some other estimator $\hat{\boldsymbol{\beta}}$ can be expressed as

$$\begin{aligned}
d_M\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)^2 &= \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)^\top \mathbf{C}_{\mathrm{OLS}}^{-1}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right) = \\
&= \frac{1}{\sigma^2}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)^\top \left(\mathbf{X}^\top\mathbf{X}\right)\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right) = \\
&= \frac{1}{\sigma^2}\left\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right\|^2_{\boldsymbol{\Sigma}_{\mathrm{XX}}}.
\end{aligned}$$

Thus, the distance induced by the quadratic form norm $\|\cdot\|_{\boldsymbol{\Sigma}_{\mathrm{XX}}}$ is proportional to the Mahalanobis distance with $\sigma^2{\boldsymbol{\Sigma}_{\mathrm{XX}}}^{-1}$, the covariance matrix of the OLS estimator. □

Therefore, with L components, PLS finds the approximation to the regression coefficients that is closest to $\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}$ with respect to the Mahalanobis distance with the covariance matrix of the OLS estimator, in the Krylov subspace of order L generated by $\boldsymbol{\Sigma}_{\mathrm{XX}}$ and $\boldsymbol{\Sigma}_{\mathrm{XY}}$. The Mahalanobis distance provides a natural measure of differences in the space of estimations, one that captures its geometry better than the Euclidean distance. Furthermore, the Mahalanobis distance between the estimations is deeply related to the Euclidean distance between the predictions:

$$d_M\left(\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})}, \hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}\right)^2 = \frac{1}{\sigma^2}\left(\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})} - \hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}\right)^\top \left(\mathbf{X}^\top\mathbf{X}\right)\left(\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})} - \hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}\right) = \\ = \frac{1}{\sigma^2}\left\|\hat{\mathbf{y}}_{\mathrm{OLS}} - \hat{\mathbf{y}}_{\mathrm{PLS}}^{(L)}\right\|^2.$$

Additional relations can be unveiled by the observation that each element in a Krylov subspace of order L is associated with a polynomial of degree $L-1$. As a result, the optimization problem in Proposition 3.3.1, is equivalent to the polynomial fitting problem given in the following corollary:

Corollary 3.3.2. *The PLS estimation with L component is $\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})} = P_{L-1}^*(\boldsymbol{\Sigma}_{XX})\boldsymbol{\Sigma}_{XY}$, where*

$$P_{L-1}^* = \underset{P\in\mathcal{P}_{L-1}}{\operatorname{argmin}} \left\|P(\boldsymbol{\Sigma}_{XX})\boldsymbol{\Sigma}_{XY} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right\|_{\boldsymbol{\Sigma}_{XX}}^2, \tag{3.11}$$

and $\mathcal{P}_{L-1}$ is the space of polynomials of degree lower or equal to $L-1$.

Proof. Since $\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})} \in \mathrm{Kry}_\ell(\boldsymbol{\Sigma}_{\mathrm{XX}}, \boldsymbol{\Sigma}_{\mathrm{XY}})$, it can be expressed as $\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})} = P_{L-1}^*(\boldsymbol{\Sigma}_{\mathrm{XX}})\boldsymbol{\Sigma}_{\mathrm{XY}}$. By substituting this expression into (3.10), we obtain (3.11). □

As stated in the following theorem, the difference between the PLS and OLS estimations can be expressed as an optimization problem in a space of polynomials:

Theorem 3.3.1. *The distance between the PLS and OLS approximations fulfills*

$$\left\|\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right\|_{\boldsymbol{\Sigma}_{XX}}^2 = \min_{Q_\ell\in\Omega_L} \sum_{d=1}^{D} Q_\ell(\lambda_d)^2\lambda_d\xi_d^2, \tag{3.12}$$

where $\{\lambda_d\}_{d=1}^D$ are the eigenvalues of $\boldsymbol{\Sigma}_{XX}$, $\{\xi_d\}_{d=1}^D$ are the coefficients of the expansion of $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ in $\{\mathbf{u}_d\}_{d=1}^D$, the basis of eigenvectors of $\boldsymbol{\Sigma}_{XX}$, and $\Omega_\ell = \{Q_L \in \mathcal{P}_L : Q_L(0) = -1\}$. Additionally, for each L, the minimum is reached for $Q_\ell^*(t) = tP_{L-1}^*(t) - 1$.

Proof. Since $\boldsymbol{\Sigma}_{\mathrm{XX}} = \mathbf{X}^\top\mathbf{X}$ is a real, symmetric matrix, it is possible to find a sequence of non-negative eigenvalues $\{\lambda_1, \ldots, \lambda_D\}$ and orthonormal eigenvectors: $\{\mathbf{u}_1, \ldots, \mathbf{u}_D\}$ such that $\boldsymbol{\Sigma}_{\mathrm{XX}} = \sum_{d=1}^D \lambda_d\mathbf{u}_d\mathbf{u}_d^\top$. This eigenvalue decomposition has three properties:

- First, the eigenvectors span the entire $\mathbb{R}^D$ space. Therefore, D scalars $\{\xi_d\}_{d=1}^D$ can be found such that $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \sum_{d=1}^D \xi_d\mathbf{u}_d$.

- Second, the norm of a vector can be calculated as $\|\mathbf{z}\|^2_{\mathbf{\Sigma}_{\mathrm{XX}}} = \mathbf{z}^\top \mathbf{\Sigma}_{\mathrm{XX}} \mathbf{z} = \sum_{d=1}^{D} \lambda_d (\mathbf{u}_d^\top \mathbf{z})^2$.
- Third, for any polynomial P, it holds that $P(\mathbf{\Sigma}_{\mathrm{XX}})\mathbf{u}_d = P(\lambda_d)\mathbf{u}_d$, for $d = 1, \ldots, D$.

Using these properties we can now find an expression to calculate the lhs of (3.12) in terms of the polynomials P_L^*.

$$\begin{aligned}
\left\| \hat{\boldsymbol{\beta}}_L - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \right\|^2_{\mathbf{\Sigma}_{\mathrm{XX}}} &= \left\| (P^*_{l-1}(\mathbf{\Sigma}_{\mathrm{XX}})\mathbf{\Sigma}_{\mathrm{XX}} - \mathbf{I}) \hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \right\|^2_{\mathbf{\Sigma}_{\mathrm{XX}}} = \\
&= \left\| \sum_{d=1}^{D} (P^*_{l-1}(\mathbf{\Sigma}_{\mathrm{XX}})\mathbf{\Sigma}_{\mathrm{XX}} - \mathbf{I}) \xi_d \mathbf{u}_d \right\|^2_{\mathbf{\Sigma}_{\mathrm{XX}}} = \\
&= \sum_{d=1}^{D} Q^*_L(\lambda_d)^2 \lambda_d \xi_d^2,
\end{aligned} \tag{3.13}$$

where $Q^*_l(t) = tP^*_{l-1}(t) - 1$, a polynomial of degree l that fulfills $Q^*_L(0) = -1$.

Additionally, Corollary 3.3.2 shows that P^*_{l-1} is the polynomial that minimizes the RHS of (3.13) over all polynomials of degree $l-1$. Therefore, Q^*_l minimizes that same quantity over all the polynomials Q_L of degree l such that $Q_L(0) = -1$. That is to say, over Ω_L. □

This theorem implies that any polynomial $Q_\ell \in \Omega_L$ can be used to provide an upper bound for the distance between the OLS and PLS estimations:

$$\left\| \hat{\boldsymbol{\beta}}^{(\mathrm{PLS})}_L - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \right\|^2_{\mathbf{\Sigma}_{\mathrm{XX}}} \leq \sum_{d=1}^{D} Q_\ell(\lambda_d)^2 \lambda_d \xi_d^2, \quad \text{for all } Q_L \in \Omega_L.$$

Furthermore, it is possible to obtain an upper bound also in terms of the norm of the OLS estimator and of Q_ℓ evaluated at the eigenvalues of $\mathbf{\Sigma}_{\mathrm{XX}}$.

Corollary 3.3.3. *Given a function $H : \Omega_\ell \to \mathbb{R}$ that, for any polynomial $R \in \Omega_\ell$, fulfills $R(\lambda_d)^2 \leq H(R)$ over all $d \in \{1, \ldots, D\}$, and given a particular polynomial $Q_\ell \in \Omega_L$,*

$$\left\| \hat{\boldsymbol{\beta}}^{(\mathrm{PLS})}_L - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \right\|^2_{\mathbf{\Sigma}_{XX}} \leq H(Q_\ell) \left\| \hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \right\|^2_{\mathbf{\Sigma}_{XX}}, \quad \textit{for all } Q_\ell \in \Omega_L.$$

Proof. From Theorem 3.3.1, and the condition $R(\lambda_d)^2 \leq H(R)$ for $d = 1, \ldots, D$,

$$\begin{aligned}
\left\| \hat{\boldsymbol{\beta}}^{(\mathrm{PLS})}_L - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \right\|^2_{\mathbf{\Sigma}_{\mathrm{XX}}} &= \min_{R \in \Omega_\ell} \sum_{d=1}^{D} R(\lambda_d)^2 \lambda_d \xi_d^2 \leq \\
&\leq \min_{R \in \Omega_\ell} H(R) \sum_{d=1}^{D} \lambda_d \xi_d^2 = \\
&= \min_{R \in \Omega_\ell} H(R) \left\| \hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \right\|^2_{\mathbf{\Sigma}_{\mathrm{XX}}} \leq \\
&\leq H(Q_\ell) \left\| \hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \right\|^2_{\mathbf{\Sigma}_{\mathrm{XX}}}.
\end{aligned}$$

□

Therefore, by choosing an H function and a specific polynomial Q_ℓ, an upper bound on the PLS error can be obtained. There are different choices for H. In Nocedal and Wright (1999), a number of results are given using the upper bound $H_1(Q_\ell) = \max_d Q_L(\lambda_d)^2$. However, this bound has a major disadvantage: it is not straightforward to calculate the polynomial Q_ℓ that minimizes H_1. In the remainder of this section, the simpler upper bound $H_2(Q_\ell) = \sum_{d=1}^{D} Q_L(\lambda_d)^2$ is considered. The following theorem provides an uper bound on the PLS error by calculating the polynomial in Ω_ℓ that minimizes H_2.

Theorem 3.3.2. *The following bound for the squared norm of the difference between the OLS and L-th PLS estimation holds:*

$$\left\|\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right\|_{\boldsymbol{\Sigma}_{XX}}^2 \leq C_L \left\|\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right\|_{\boldsymbol{\Sigma}_{XX}}^2, \tag{3.14}$$

where

$$C_L = D(1 - \mathbf{c}_L^\top \mathbf{H}_L^{-1} \mathbf{c}_L), \qquad \mathbf{H}_\ell = \begin{pmatrix} \mu'_2 & \cdots & \mu'_{L+1} \\ \vdots & \ddots & \vdots \\ \mu'_{L+1} & \cdots & \mu'_{2L} \end{pmatrix}, \qquad \mathbf{c}_\ell = \begin{pmatrix} \mu'_1 \\ \vdots \\ \mu'_\ell \end{pmatrix},$$

and μ'_ℓ is the ℓ-th raw moment of the distribution of the eigenvalues of $\boldsymbol{\Sigma}_{XX}$.

Proof. All polynomials in Ω_L can be expressed as $R_L(t) = -1 + a_1 t + \cdots + a_L t^L$ for some coefficients $a_1, \ldots, a_L$. Therefore, as a function of the coefficients of the polynomials, the bound can be expressed as $h_L(a_1, \ldots, a_L) = \sum_{d=1}^{D} \left(-1 + a_1\lambda_d + \cdots + a_L\lambda_d^L\right)^2$. To minimize this function, we calculate its gradient and determine the coefficients for which it is zero:

$$\begin{aligned} \frac{\partial h_L}{\partial a_l} &= 2\sum_{d=1}^{D}(-1 + a_1\lambda_d + \cdots + a_L\lambda_d^L)\lambda_d^l = \\ &= -2\sum_{d=1}^{D}\lambda_d^\ell + 2a_1\sum_{d=1}^{D}\lambda_d^{\ell+1} + \cdots + 2a_L\sum_{d=1}^{D}\lambda_d^{\ell+L} = \\ &= 0, \end{aligned}$$

for all $\ell = 1, \ldots, L$.

This set of equations can be rewritten in terms of the sample raw moments of the eigenvalues:

$$-2D\mu'_\ell + 2a_1 D\mu'_{\ell+1} + \cdots + 2a_L D\mu'_{\ell+L} = 0 \iff a_1\mu'_{\ell+1} + \cdots + a_L\mu'_{\ell+l} = \mu'_\ell,$$

for all $\ell = 1, \ldots, L$.

These equations can be expressed as the system $\mathbf{H}_L\mathbf{a}_L = \mathbf{c}_L$. Therefore, the coeffi-

cients that minimize h_L are $\mathbf{a}_L^* = \mathbf{H}_L^{-1}\mathbf{c}_L$. Additionally, we express h_L as

$$h_L(a_1,\dots,a_L) = (-1,\mathbf{a}_L)\begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ \lambda_1^L & \dots & \lambda_D^L \end{pmatrix}\begin{pmatrix} 1 & \dots & \lambda_1^L \\ \vdots & \ddots & \vdots \\ 1 & \dots & \lambda_D^L \end{pmatrix}\begin{pmatrix} -1 \\ \mathbf{a}_L \end{pmatrix} = \\ = (-1,\mathbf{a}_L)\begin{pmatrix} D & D\mathbf{c}_L^\top \\ D\mathbf{c}_L & D\mathbf{H}_L \end{pmatrix}\begin{pmatrix} -1 \\ \mathbf{a}_L \end{pmatrix}.$$

Substituting the expression for $\mathbf{a}_L^*$ in the previous formula and multiplying blockwise:

$$h_L(\mathbf{a}_L^*) = (-1,\mathbf{H}_L^{-1}\mathbf{c}_L)\begin{pmatrix} D & D\mathbf{c}_L^\top \\ D\mathbf{c}_L & D\mathbf{H}_L \end{pmatrix}\begin{pmatrix} -1 \\ \mathbf{H}_L^{-1}\mathbf{c}_L \end{pmatrix} = \\ = (-1,\mathbf{H}_L^{-1}\mathbf{c}_L)\begin{pmatrix} -D + D\mathbf{c}_L^\top\mathbf{H}_L^{-1}\mathbf{c}_L \\ 0 \end{pmatrix} = \\ = D(1 - \mathbf{c}_L^\top\mathbf{H}_L^{-1}\mathbf{c}_L).$$

Additionally, the obtained coefficients $\mathbf{a}_L^*$ define the polynomial

$$R_L^*(t) = -1 + a_1^* t + \dots + a_L^* t^L,$$

which has the minimal sum of squared values when evaluated at the eigenvalues $\lambda_1,\dots,\lambda_D$. □

This result provides an upper bound for the distance between $\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ that depends only on the distribution of the eigenvalues of the regressor covariance matrix. Explicit expressions of this bound can be derived for PLS regression with one and two components.

Corollary 3.3.4. *The bounds given in* (3.14) *for* $L = 1$ *and* $L = 2$ *can be expressed as a function of the coefficient of variation* ($c_v = \sigma/\mu$)*, the coefficient of asymmetry* (γ) *and the kurtosis* (κ) *of the eigenvalues of* $\boldsymbol{\Sigma}_{xx}$.

$$C_1 = D\frac{c_v^2}{1+c_v^2}, \qquad C_2 = D\frac{c_v^4(\kappa-\gamma^2-1)}{(\kappa-\gamma^2)c_v^4 + (\kappa-3-2\gamma)c_v^3 - 2\gamma\, c_v + 1}.$$

Proof. These identities are obtained by expressing the raw moments that appear in C_ℓ in terms of μ, σ, γ and κ, and then simplifying the resulting formulas. □

From these expressions it is apparent that the more concentrated the eigenvalues of $\boldsymbol{\Sigma}_{\mathrm{XX}}$ ($c_v \to 0$), the fewer PLS components are needed to approximate $\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}$ with a given accuracy. The bound for $L = 1$ depends only on the coefficient of variation of the distribution of eigenvalues $c_v = \sigma/\mu$. It is proportional to c_v^2 in the limit $c_v \to 0^+$. Therefore, if the eigenvalues of $\boldsymbol{\Sigma}_{\mathrm{XX}}$ are grouped in one tight cluster, keeping a single PLS component yields an accurate approximation of $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$. The value of the bound for $L = 2$ depends not only

on c_v but also on the coefficient of asymmetry and the kurtosis, which makes it harder to interpret. However, it is proportional to c_v^4 in the limit $c_v \to 0^+$.

Additionally, from Pearson's inequality ($\kappa \geq 1 + \gamma^2$), the quantity $\kappa - \gamma^2 - 1$ is non-negative (Sharma & Bhandari, 2015). This quantity is zero for dichotomous distributions. Thus, C_2 should be small when the eigenvalues are distributed in two tightly packed clusters. In the next section, we provide numerical illustrations of the dependence of distance between $\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}$ as a function of L, for different distributions of the eigenvalues of $\boldsymbol{\Sigma}_{\mathrm{xx}}$.

3.4 Empirical study

In this section, an empirical study is carried out to investigate the effect of the eigenvalue distribution of the regressor covariance matrix on PLS. Specifically, we analyze the dependence of the quadratic-form distance between $\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}$, the upper bound established in Theorem 3.3.2 for this distance, and the accuracy of the linear predictor as a function of the number of PLS components considered. The analysis is first performed in regression problems with synthetic data for different forms of the distribution of eigenvalues. The corresponding analysis is then performed for the California Housing dataset (Kelley Pace & Barry, 1997).

3.4.1 Synthetic data

In this section, synthetic data are used to illustrate the behavior of the PLS method depending on the eigenvalue distribution of the regressor covariance matrix. Five regression problems are considered. In these problems, X is modelled as a multivariate normal vector $X \sim N(0, \boldsymbol{\Sigma})$. The eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$, $\{\lambda_d\}_{d=1}^D$, are sampled from different distributions with specific characteristics. Specifically, $D = 30$ eigenvalues are selected with the following characteristics:

1. 30 equally spaced eigenvalues from 2.5 to 7.5.
2. One cluster of 30 eigenvalues sampled from $N(5, 0.1)$.
3. Two clusters of 15 eigenvalues, each sampled from $N(2.5, 0.1)$ and $N(7.5, 0.1)$.
4. Three clusters of 10 eigenvalues sampled from $N(2.5, 0.1)$, $N(5, 0.1)$, and $N(7.5, 0.1)$.
5. Three clusters of 10 eigenvalues sampled from $N(0.2, 0.1)$, $N(5, 0.1)$, and $N(7.5, 0.1)$; so that one of the clusters is very close to zero.

These eigenvalue distributions are displayed in Figure 3.1. The actual covariace matrix is generated by a random rotation of the diagonal eigenvalue matrix: $\boldsymbol{\Sigma} = \mathbf{Q}^\top \mathrm{diag}(\lambda_1, \ldots, \lambda_D)\mathbf{Q}$, where $\mathbf{Q}$ is a uniformly-distributed orthogonal random matrix. The rotation matrix $\mathbf{Q}$ is obtained from the QR decomposition of a random matrix whose entries are sampled from a standard normal distribution (Mezzadri, 2007). Finally, the data matrix, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top$ is obtained by stacking $N = 1000$ samples from this random vector.

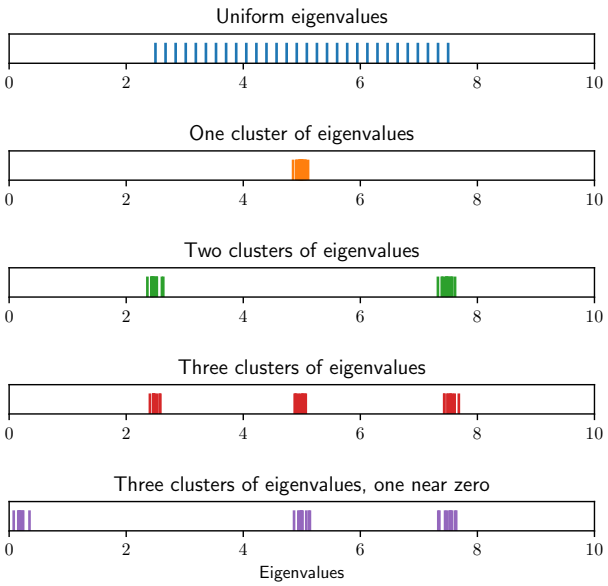

Figure 3.1: Eigenvalue distributions in the synthetic regression problems

To generate the response data, the linear model with additive noise presented in (3.5) is used. The $\boldsymbol{\beta}$ parameter is a random vector whose components are sampled from a uniform distribution in $[0, 1]$. The noise $\boldsymbol{\epsilon}$ is sampled from a $N(0, \sigma^2)$ distribution, where $\sigma = 0.1\,\mathrm{std}(\mathbf{X}\boldsymbol{\beta})$, so that the model is not dominated by the noise. Finally, the response vector is computed as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

In the experiments carried out, the closeness between $\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}$ is quantified in terms of the normalized estimation difference:

$$\mathrm{NED}_\ell = \frac{\|\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\|^2_{\boldsymbol{\Sigma}_{\mathrm{XX}}}}{\|\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\|^2_{\boldsymbol{\Sigma}_{\mathrm{XX}}}}.$$

From (3.14), it is apparent that C_ℓ is an upper bound on NED_L. The results reported are averages over 20 realizations of the data.

The plots in the left column of Figure 3.2 display the dependence of the normalized differences between the estimation, NED_ℓ, and of the corresponding upper-bound, C_L, on L, the number of PLS components considered. These plots show how, as L increases, the decrease of the bound introduced in Theorem 3.3.2 parallels that of the difference between the estimations. As discussed in the previous section, PLS can be formulated as a polynomial fitting problem. In particular, Theorem 3.3.1 provides a way of expressing the error of the estimation with L iterations as a function of the values of some polynomial Q_ℓ, of degree lower or equal to L that fulfills $Q_L(0) = -1$. The optimal polynomials Q_ℓ^* defined in Theorem 3.3.1 are plotted in the right column of Figure 3.2 .

It is possible to interpret the features of the curves displayed in the left column of Figure 3.2 from the characteristics of the polynomials plotted in the right column of this figure. In the first scenario, in which the eigenvalues are uniformly distributed in an interval separate

from zero, considering more components allows to find polynomials that fulfill $Q_\ell(0) = -1$ and take small values for all the eigenvalues. In the second one, the decrease of NED_ℓ is much steeper because having the eigenvalues closely packed in a single cluster makes the polynomial fitting problem much simpler. For a given numbers of components, the corresponding polynomials take smaller values on the eigenvalues in the second scenario than in the first one. This result is consistent with the dependency of C_1 and C_2 on c_v given in Corollary 3.3.4.

Figure 3.2 also shows that the decrease of NED_ℓ with L follows different patterns depending on the number of clusters in which the eigenvalues are grouped. In particular, the decrease is sharper for specific numbers of components. When the eigenvalues are grouped in two clusters, the first abrupt decrease of NED_ℓ occurs between $L = 1$ and $L = 2$. This observation can be explained by noting it is not possible to find a polynomial of degree one (i.e., a straight line) that takes small values on both clusters and passes through the point $(0, -1)$. However, a polynomial of degree two (i.e., a parabola) provides a reasonable fit. Significant improvements are observed also for $L = 4$ and $L = 6$. This is due to the fact that, in those cases, it is possible to find polynomials that pass through (0,-1) with equal numbers of roots located in the vicinity of each of the clusters. A similar analysis can be carried out for the fourth scenario, in which the eigenvalues are clustered in 3 groups. In this case, sharper improvements are found for $L = 3$ and $L = 6$.

To complete the analysis, we consider a case in which one of the clusters of eigenvalues is close to 0. From the plot in the bottom left of Figure 3.2 it is apparent that the decrease of NED_ℓ with L is rather slow. The reason for this is that, since the fitted polynomial has to go through $(0, -1)$, large values of L are needed so that the polynomial can take simultaneously small values for the eigenvalues in the vicinity of 0 and in the other clusters.

We now compare the performance of PCA and PLS regression as a function of L, the number of components considered. The quality of the predictions is measured in terms of the coefficient of determination (R^2 score), which represents the proportion of explained variance. In most regression problems PLS is expected to outperform PCA because, in the definition of the components, the correlations between the regressor and response variables are taken into account in the former, but not in the latter (Frank & Friedman, 1993). Since the properties of PLS depend on the distribution of the eigenvalues of $\mathbf{\Sigma}_{\text{XX}}$, the regressor covariance matrix, we carry out the analysis for the five scenarios described earlier. In Figure 3.3 we compare the curves that trace the dependence R^2 on L, for PLS (left plots) and PCA (right plots) in the first two synthetic datasets. This comparison illustrates the differences between problems in which the eigenvalues of the regressor covariance matrix are uniformly distributed and problems in which they are clustered around a particular value, different from zero. As expected, PLS obtains better results when the eigenvalues concentrate around a few different values. In fact, when they are clustered in a single tight group, the PLS regression model with only one component provides a very accurate prediction of the response. By contrast, when the eigenvalues are uniformly distributed, more components are needed. The behavior of PCA is markedly different. In the case of clustered eigenvalues, the R^2 score of PCA increases linearly with the number of eigenvalues considered. This is to be expected since the increment in explained variance is proportional to the eigenvalue

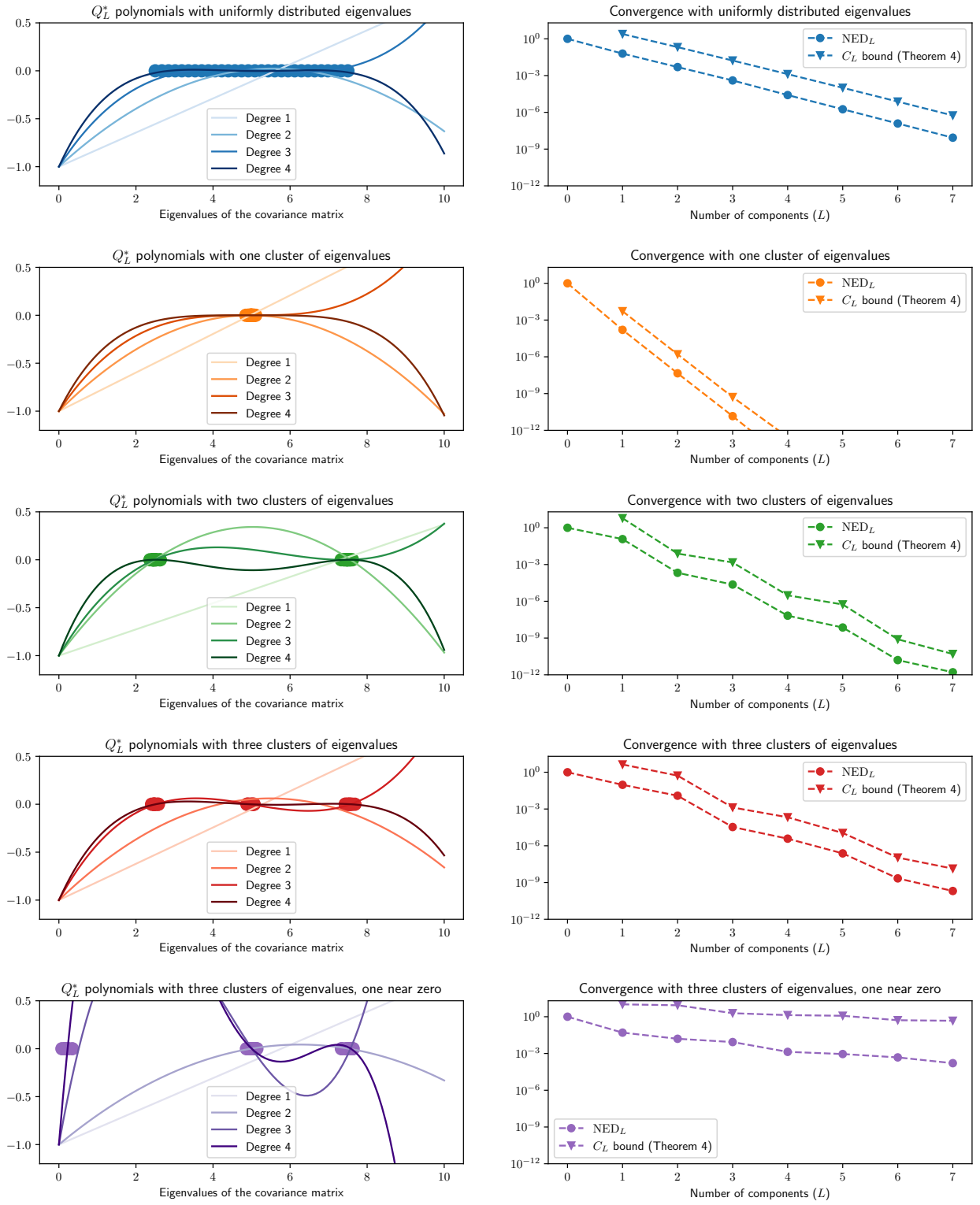

Figure 3.2: PLS estimation distance analysis with different distributions of eigenvalues of the regressor covariance matrix

that corresponds to the eigenvector considered by PCA at each step. If the eigenvalues are spread out uniformly, PCA considers first the components that correspond to the largest eigenvalues. Therefore, the magnitude of the eigenvalues decreases as more components are considered, which leads to a reduction of the rate at which R^2 increases as a function of L. Additionally, Figure 3.3 also shows that PCA needs many more components to achieve the same R^2 scores as PLS.

The plots displayed in Figure 3.4 illustrate the properties of the evolution of the R^2 score as a function of L, depending on the number of clusters in which the eigenvalues are grouped. From these results we conclude that, in this case, the number of PLS components necessary to obtain a value of R^2 close to 1 (perfect prediction) coincides with the number of eigenvalue clusters. This is consistent with the analysis of the differences between $\hat{\boldsymbol{\beta}}_L^{(\mathrm{PLS})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{OLS})}$ for these datasets. Regarding PCA, we can see how the number of clusters of eigenvalues has only minor effects in dependence of the R^2 scores with L. For example, with two clusters, the R^2 increases faster during the first 15 iterations, which corresponds to the cluster with the largest 15 eigenvalues. For the scenario with three clusters of eigenvalues, the rate of increase drops after 10 and 20 components have been considered. These correspond to having included in the model all the components in the first, and in the first and second largest clusters, respectively.

Finally, we use the last two scenarios to investigate the impact of having a cluster of eigenvalues close to zero. Figure 3.5 shows how that for $L > 1$, the performance of PLS deteriorates when there is a cluster of small eigenvalues. This is again to be expected from the theoretical analysis carried out in the previous section because of the difficulties of fitting a polynomial that goes throgh $(0, -1)$ and takes small values at the locations of the eigenvalues in the clusters. By contrast, PCA achieves better results when a sizeable fraction of the eigenvalues are close to zero. In fact, the maximum value of R^2 is attained for $L = 20$, once all the components that correspond to eigenvalues significantly larger than zero have been selected. Nonetheless, for a given number of components, PLS outperforms PCA regression also in these scenarios.

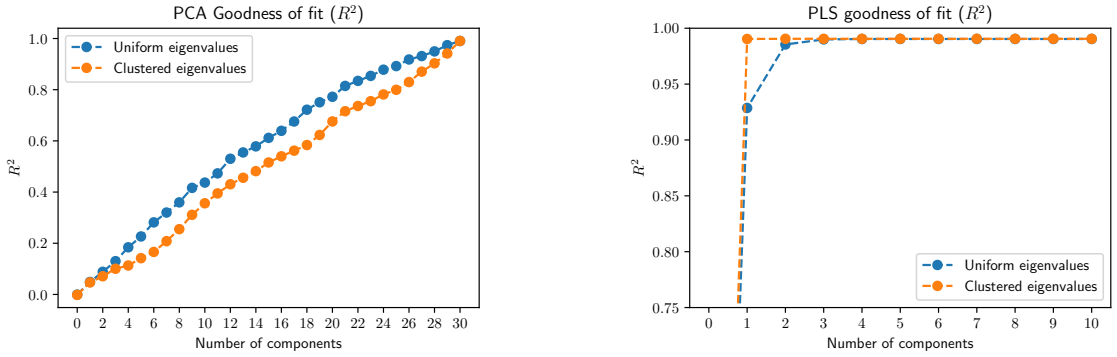

Figure 3.3: Accuracy of the predictions of PCA and PLS regression measured in terms of the R^2 score depending on whether the eigenvalues are concentrated or spread out uniformly.

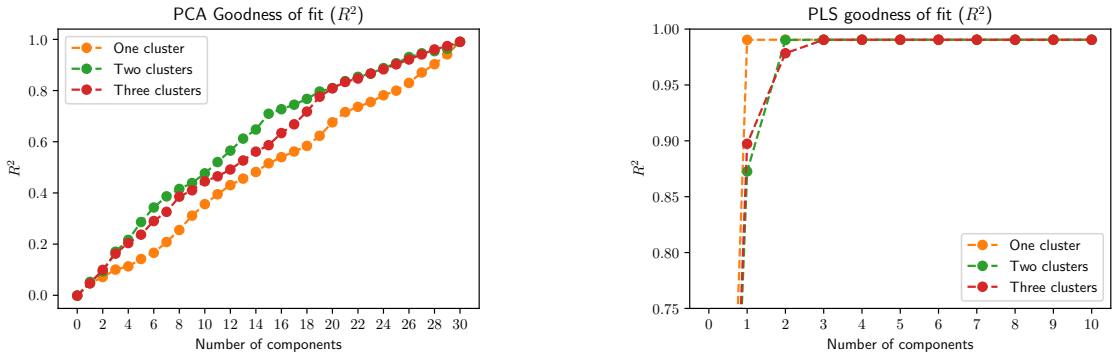

Figure 3.4: Accuracy of the predictions of PCA and PLS regression measured in terms of the R^2 score depending on the number of clusters in which the eigenvalues are grouped

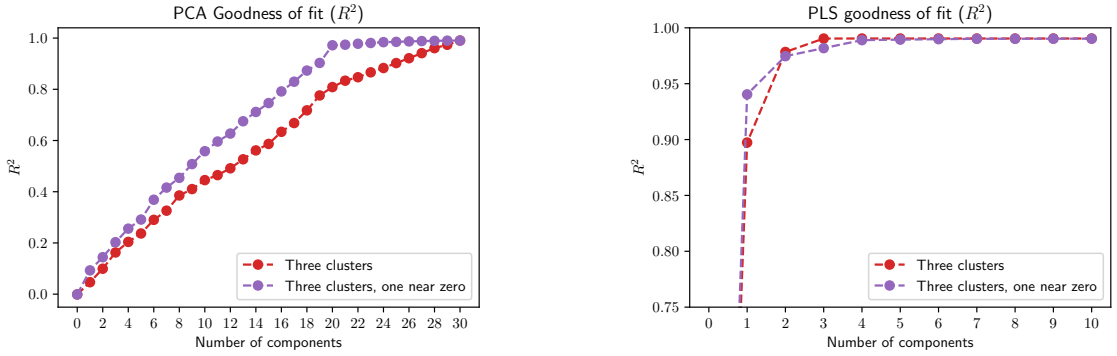

Figure 3.5: Accuracy of the predictions of PCA and PLS regression measured in terms of the R^2 score, depending on whether there is a cluster of eigenvalues near zero

3.4.2 The Californian Housing dataset

In this section we analyze the properties of PLS regression for the California Housing dataset (Kelley Pace & Barry, 1997) In this problem, the goal is to predict the median house value in a particular block group in a California district using 8 attributes ($D = 8$): the median house age, the average number of rooms, the average number of bedrooms, the number of people residing within the block, the average number of household members, and the latitude and longitude of the block group. As a preprocessing step, both the regressor vector and the response variable are centered so that they have zero mean. Each column of $\mathbf{X}$ is scaled so that it has unit variance. In the original dataset, a median house value of $500,000$\$ is assigned to instances whose actual value is above that threshold. To avoid distortions associated to this thresholding, these examples have been discarded.

Figure 3.6 shows the eigenvalue distribution of the regressor covariance matrix for the California Housing dataset. The eigenvalues are roughly grouped in three clusters, one of them close to zero. This pattern is similar to the last synthetic dataset analyzed in the previous section. However, the eigenvalues in the central cluster are more spread out. This dispersion hinders somewhat the performance of PLS, which is nonetheless fairly good. The differences between the PLS and the OLS approximations as a function of L are analyzed in

Figure3.7. The left plot displays the dependence of these differences, quantified by NED_ℓ, and of C_L, the upper bound of these differences derived in this work, as a funtion of L, the number of PLS components considered. Note that, for $L = 8$ the PLS coincides with the OLS estimator. As expected, the distance between the estimations decreases slowly, because of the presence of the small eigenvalues and, to a lesser extent, the dispersion of the medium-sized eigenvalues.

Figure 3.8 presents the results of a comparison between PCA and PLS regression. The left plot displays the curves that trace the dependence of the R^2 score, a measure of the quality of the predictions, with L, the number of components considered. From these results one concludes that PLS obtains better results than PCA, and needs fewer components to achieve an accuracy comparable to OLS. The evolution of C_ℓ as a function of L is displayed in the right plot of this figure. Note that the descent of the bound mirrors the increase of the R^2 score as L increases. This illustrates that the upper bound defined in Theorem 3.3.2 provides an effective way to monitor the performance of PLS.

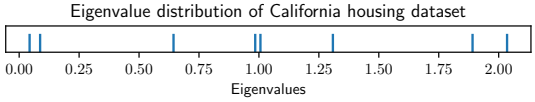

Figure 3.6: Eigenvalue distribution in the California Housing dataset

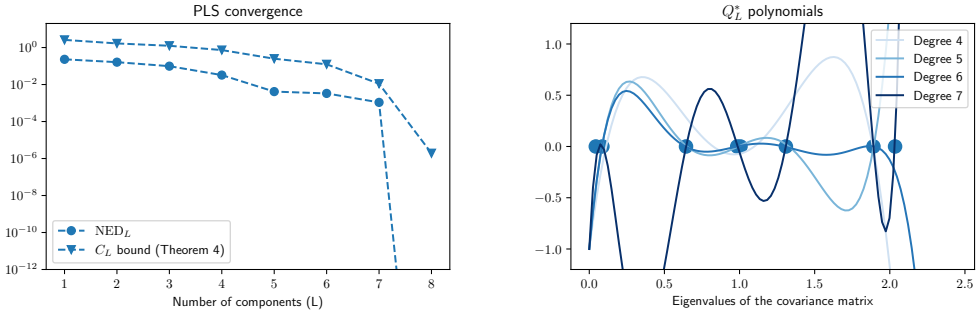

Figure 3.7: PLS estimation distance analysis in the California Housing dataset

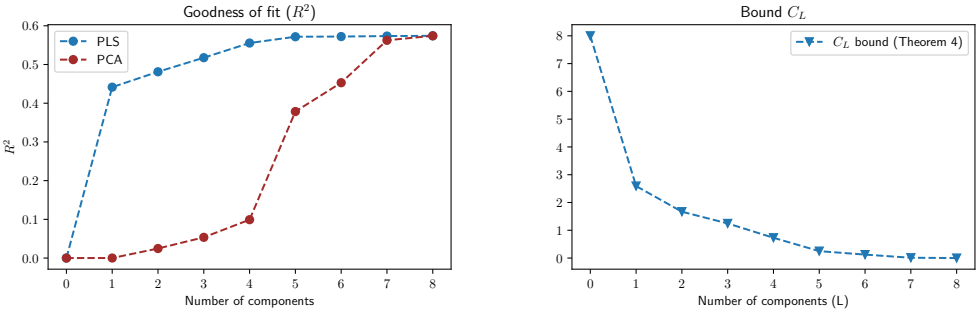

Figure 3.8: Accuracy of the predictions of PCA and PLS regression measured in terms of the R^2 score in the California Housing dataset

Chapter 4

Empirical comparison of PCA and PLS Regression

In the previous chapter we analyzed the impact of the eigenvalue distribution of the regressor covariance matrix on the effectiveness of partial least squares and principal component regression. As shown in Section 3.4, PLS regression can provide better results than PCA regression when these eigenvalues are grouped in clusters. The goal of this chapter is to investigate how this translates to real-world performance, by analyzing the performance of PLS, with respect to PCA in a variety of real-world datasets. The datasets selected for this study span different domains, and posses varying characteristics, allowing us to produce a representative comparison of the two dimensionality reduction techniques.

We also seek to study how these dimensionality reduction methods perform when applied as part of the preprocessing step on a regression problem. First, we analyze the impact of scaling the data before performing dimensionality reduction. Second, we consider their use in combination with non-linear predictors and regularized methods. Doing so enables us to study if PLS is an effective technique when non-linear relationships have to be considered. This analysis is performed for both finite-dimensional regressors (Section 4.1) and functional regressors (Section 4.2).

4.1 Multiple regression

In this empirical study, ten regression problems from different domains have been considered. In what follows each dataset is described, alongside with the preprocessing steps applied. As a summary, the main properties of each dataset are enumerated in Table 4.1.

Diabetes

The original source of the dataset is Efron et al. (2004), and it was downloaded through *scikit-learn*'s dataset module. The target variable is a quantitative measure of disease progression during one year. Ten regressors are considered: age, sex, body mass index, average blood

	Samples	Regressors	Source
Diabetes	442	10	(Efron et al., 2004)
California Housing	20640	8	(Kelley Pace & Barry, 1997)
Cancer Registry	3047	29	clinicaltrials.gov, cancer.gov, census.gov
Wine Quality	1599	11	(Cortez et al., 2009)
US Census	72727	31	US Census Bureau, tables DP03 and DP05.
Life Expectancy	2938	19	Global Health Observatory, download link
Bike Sharing	8760	12	(Sathishkumar et al., 2020)
AIDS Clinical Trials	2139	23	(Hammer et al., 1996)
Obesity	2111	22	(Palechor & Manotas, 2019)

Table 4.1: Multivariate datasets considered

pressure, and six blood serum measurements. The data was collected for 442 patients.

California Housing

The original source of the dataset is Kelley Pace and Barry (1997), and it was retrieved through *scikit-learn*'s dataset module. The target variable is the median house price in geographically compact blocks of California real state. Eight regressor variables are considered: the median income in the block, the average number of rooms, the average number of bedrooms, the population, the average of household members, and the latitude and longitude. The dataset contains 20640 samples.

The target variable has been thresholded at 5 hundred of thousands of dollars. Since this affects less than 5 % of the observations, the affected rows have been discarded. Moreover, highly atypical values were detected in the columns corresponding to the average number of bedrooms, rooms, and occupancy (count of household members), as well as the population. To solve this issue, a threshold of 5 times the interquartile range was empirically determined, and the observations taking values outside this range were dropped. Figure 4.1 shows the data after discarding these observations.

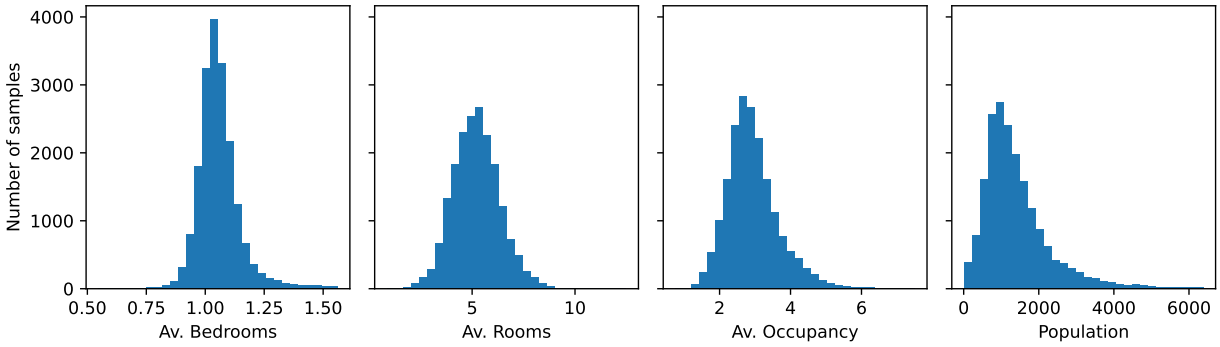

Figure 4.1: Result of outlier removal in the California Housing dataset

Cancer Registry

This dataset was created by Noah Rippner, as the result of joining data cancer trial data from clinicaltrials.gov and cancer.gov, along with demographic data from census.gov. The version utilized for the experiments that follow was downloaded from https://data.world/nrippner/cancer-trials. The target variable is the age-adjusted [1] mortality rate per county in the United States. Twenty-nine regressor variables are considered, including the incidence rate, education statistics, income data, and racial information. The data was collected for 3047 counties in the US.

The same outlier removal procedure as in the previous dataset was applied considering the values from the columns containing the median age, the number of cancer trials in the county per capita and the population estimate. Additionally, the column containing the name of the county was dropped, along with the total number of deaths by year and the average number of deaths per year. These last two columns were discarded as they contained very similar information to the target variable (age-adjusted death rate).

Wine Quality

The original source of the dataset is Cortez et al. (2009), and the dataset was downloaded from https://www. kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009. The target variable is the mean score given to a red wine sample by three sensory assessors. Eleven regressor variables are considered, containing the results of common psychochemical tests: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, and alcohol contents. The data was collected for 1599 wine samples.

US Census

This dataset is the result of joining the DP03 and DP05 tables from the 2015 American Community Survey 5-year estimates study conducted by the Census Bureau of the United States. The data has been downloaded from https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data. The target variable is the unemployment rate in a census tract in the US. Census tracts are geographical subdivision of US counties defined by the census bureau. Thirty-one regressor variables are considered, including racial, income and occupation statistics. The data was collected for 72727 tracts.

As part of the preprocessing, the identifier of the census tract was removed, along with the county and state name. Additionally, the count of employed inhabitants was also dropped, due to its similarity with the target variable (unemployment rate).

[1] https://www.health.ny.gov/diseases/chronic/ageadj.htm

Life Expectancy

This dataset was created by Kumar Rajarshi, as the result of joining data sources provided in the Global Health Observatory by the World Health Organization, and performing some preprocessing on them. The dataset, alongside with a description of the preprocessing steps is available at https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?resource=down-. The target variable is the life expectancy of a country. Nineteen regressor variables are considered, including demographic aspects such as population or infant mortality, economic indicators such as GDP and Human Development Index, and the prevalence of some diseases such as Hepatitis or HIV. The data was collected for 193 countries during the years 2000-2015. In some countries, data was not available for all years, leading to some rows being absent. The total number of observations is 2938.

The Life Expectancy dataset required little processing. The country and year columns were dropped. Furthermore, the status variable, which contained either "Developing" or "Developed" was transformed to a binary variable.

Bike Sharing

The original source of the dataset is Sathishkumar et al. (2020), and the dataset was downloaded from UC Irvine's Machine Learning Repository (Markelle, Longjohn, & Nottingham, 1998). The target variable is the number of bikes rented during an hour. Twelve regressor variables are considered, including weather information, the season, and the presence of holidays. The data was collected for 8760 windows of one hour.

As part of the preprocessing, the date column was dropped. Two more columns were modified: the textual season column was encoded employing one-hot encoding, and the holiday column was encoded as 0 (no holiday) and 1 (holiday).

AIDS Clinical Trials

The original source of the dataset is Hammer et al. (1996), and the dataset was downloaded from UC Irvine's Machine Learning Repository (Markelle et al., 1998). The target variable is the time of failure of each subject. Twenty-three regressor variables are considered including physical statistics, drug history and the outcome of medical tests. The data was collected for 2139 patients.

Obesity

The original source of the dataset is Palechor and Manotas (2019), and the dataset was downloaded from UC Irvine's Machine Learning Repository (Markelle et al., 1998). The target variable is the obesity level. The observations have been labeled with one of seven obesity levels, from "Insufficient weight" to "Obesity Type III". Twenty-two regressor variables are considered, including physical traits, family history, and eating habits. The dataset contains 2111 observations.

The main preprocessing step in the Obesity dataset was to transform the textual columns into numeric values. The columns CAEC (Frequency of food consumption between meals) and CALC (Frequency of alcohol consumption) contained one of: "no", "Sometimes", "Frequently", or "Always" and were encoded numerically in increasing frequency order. The binary features were encoded as 0 or 1. Finally, the obesity level contained one of "Insuficient_Weight", "Normal_Weight", "Overweight_Level_I", "Overweight_Level_II", "Obesity_Level_I", "Obesity_Level_II", or "Obesity_Level_III"; and was encoded numerically in increasing order.

Communities and Crime

The original source of the dataset is Redmond and Baveja (2002), and the dataset was downloaded from UC Irvine's Machine Learning Repository (Markelle et al., 1998). The target variable is the per-capita violent crime rate. A total of 101 regressor variables are considered, including income data, racial statistics, average housing characteristics and demographic data. The dataset contains data from 1993 counties.

In this dataset, there were many rows with missing values (94% of the observations). However, the missing values were concentrated on a few columns. In particular, these missing values were mostly in 22 of the columns. Since the dataset contains 127 variables, dropping the columns is a better option. After dropping these columns, only one row has to be dropped, to address the missing value in the column "OtherPerCap"(other racial group per capita). Finally, the textual columns "state", "county" and "communityname" were also discarded.

4.1.1 Results

In this section, we present the results obtained by applying principal component regression and partial least squares regression to the problems described in the previous section. The experiments have been carried out using both unscaled and scaled data. Scaling has been performed by subtracting the sample mean and dividing each variable by its sample standard deviation. In all cases, the models were fitted on a fixed training partition, and evaluated in a test partition (20% of the observations). Figure 4.2 displays the results obtained by the four configurations as the number of components vary. To ensure that the differences in the first components could be clearly seen, the number of components were limited to 25 even if some datasets contained more regressor variables.

Analyzing the results, two clear patterns are observed: PLS outperforms PCA in the first components, and the standardization of the variables generally yields better results. Since both the variance (PCA's criterion) and the covariance (PLS's criterion) are affected by scale changes, the impact of the standardization is expected. However, it is interesting to note that, while in PLS the standardization led to better results in all the problems investigated, this is not always the case for PCA. For example, in the case of the California Housing problem, from four to seven components, PCA obtains better results when the regressors are not scaled. This dataset combines variables with drastically different scales.

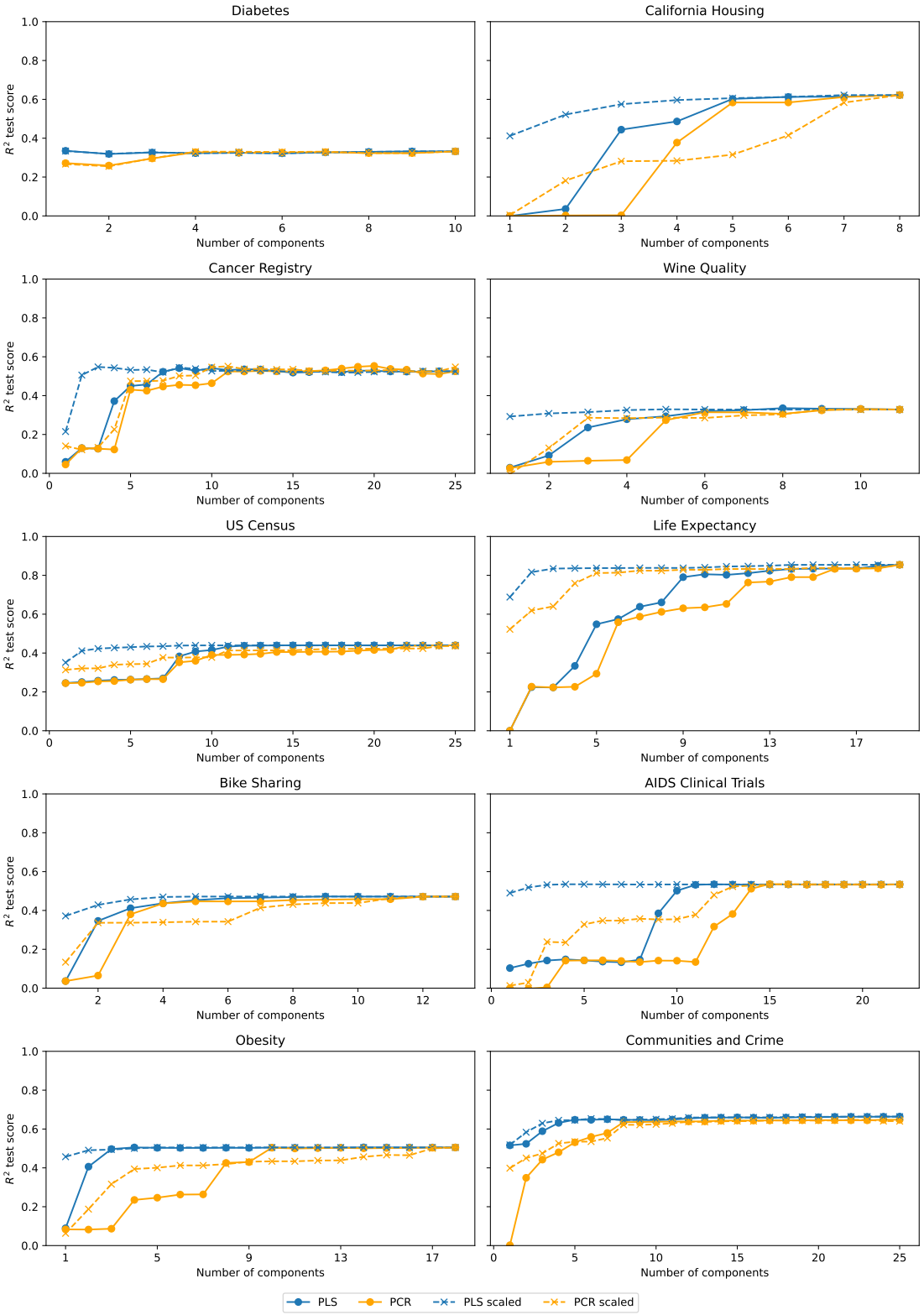

Figure 4.2: PCA and PLS Regression R^2 test scores (limited to 25 components)

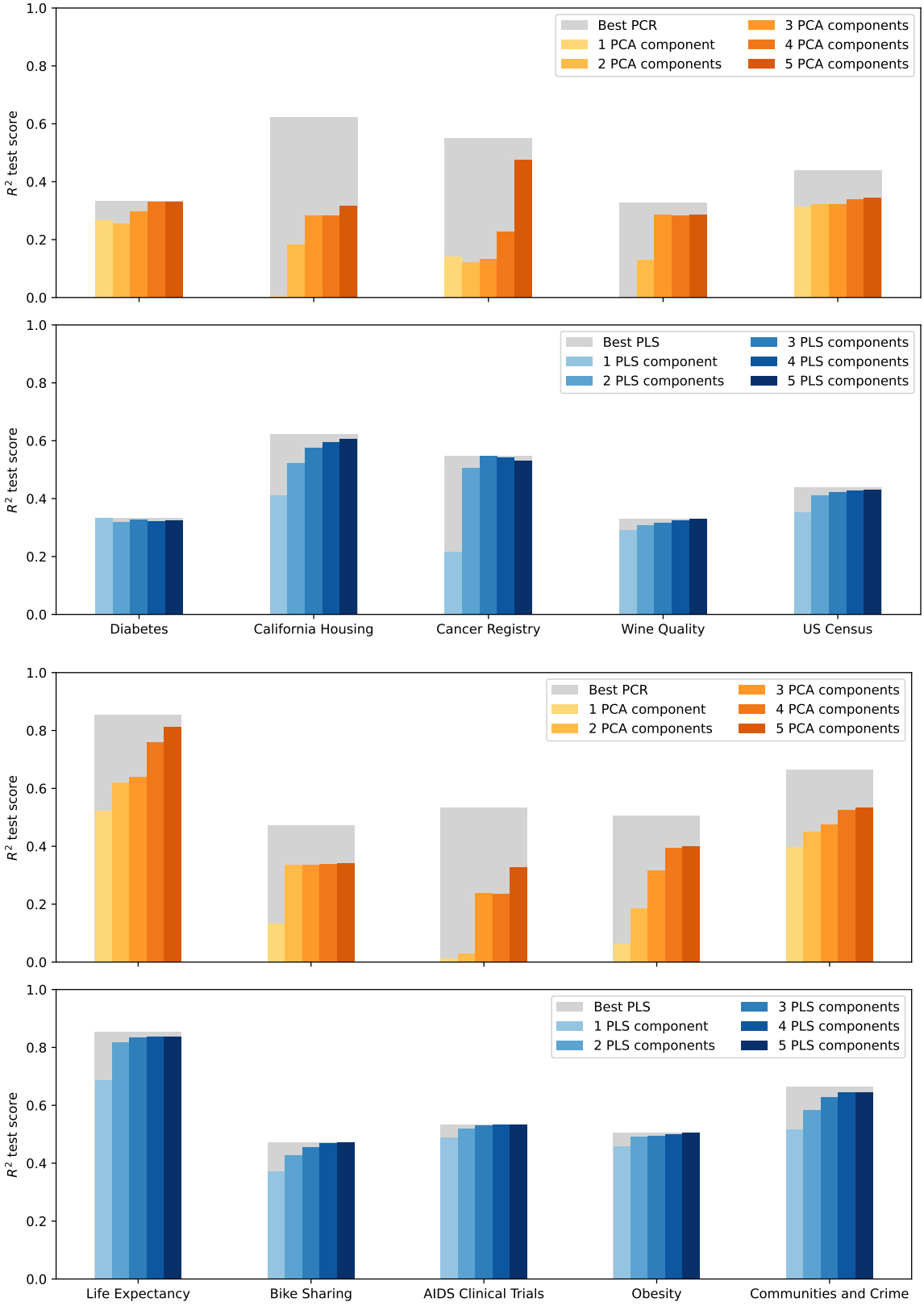

Figure 4.3: Comparison of the first components of PLS and PCA

Therefore, some features have a higher variance simply due to the scale. For example, the population variable takes values in the thousands, and has a variance of the order of 10^6 while the average number of bedrooms takes values between zero and two, with a variance of only 10^{-3}. As a result, the selection of the PCA components is mostly determined by the scale of the variables. In fact, if we sort the variables in order of decreasing variance, the variable with the highest correlation with the target is the fifth variable, which matches the number of components for which the test R^2 score converges to its maximum value.

As we have seen, the performance of scaled PLS regression is, in general, better than the other models considered. Typically, the vast majority of the variance can be explained using the first two components. This contrasts with the evolution of the scores of PCA, in which many more components must be considered to reach similar results. To highlight this difference, Figure 4.3 compares the scores achieved by both methods considering the first five components. On average, the first two components of PLS can explain 92.7% of the variance that can be explained by the best linear model, while this metric drops to 49.5% in the case of PCA.

To improve the accuracy of the predictors, we have considered the use of penalized regression and of non-linear regression techniques. As discussed in Chapter 2, PLS regression can be understood as a two-step process. The first step is the projection of the original data onto a lower-dimensionality space. Then, a linear regression model is fitted by least squares. This same separation holds true with PCA, since the difference between both methods lies on how the original data is projected onto the lower-dimensionality space. Therefore, one can combine these methods with any multivariate predictor by replacing the linear regression of the second step with a different regression technique.

As regularization techniques, ridge, lasso and elastic-net regression were considered. The performance of each of these methods, as well as linear regression (as the baseline) is plotted in the left column of Figures 4.4 and 4.5 for all datasets. A key difference with linear regression is that these methods have hyperparameters that have to be selected. To find the optimum combination of hyperparameters for each number of components, a grid search was performed. The combination of hyperparameters selected is the one that yields the best cross validation score (with five folds) in the training dataset. For the sake of completeness, the hyperparameter values considered are included in Table 4.2. The hyperparameters whose values are not given in the tables are left to the default values in *scikit-learn*, which is the library we used for these experiments.

Ridge and Lasso	
Alpha	$0.001, 0.01, 0.1, 1, 10, 100$

Elastic-net	
Alpha	$0.001, 0.01, 0.1, 1, 10, 100$
L_1 ratio	$0.001, 0.01, 0.1, 0.5, 0.7, 0.9$

Table 4.2: Hyperparameters considered in regularized regressors

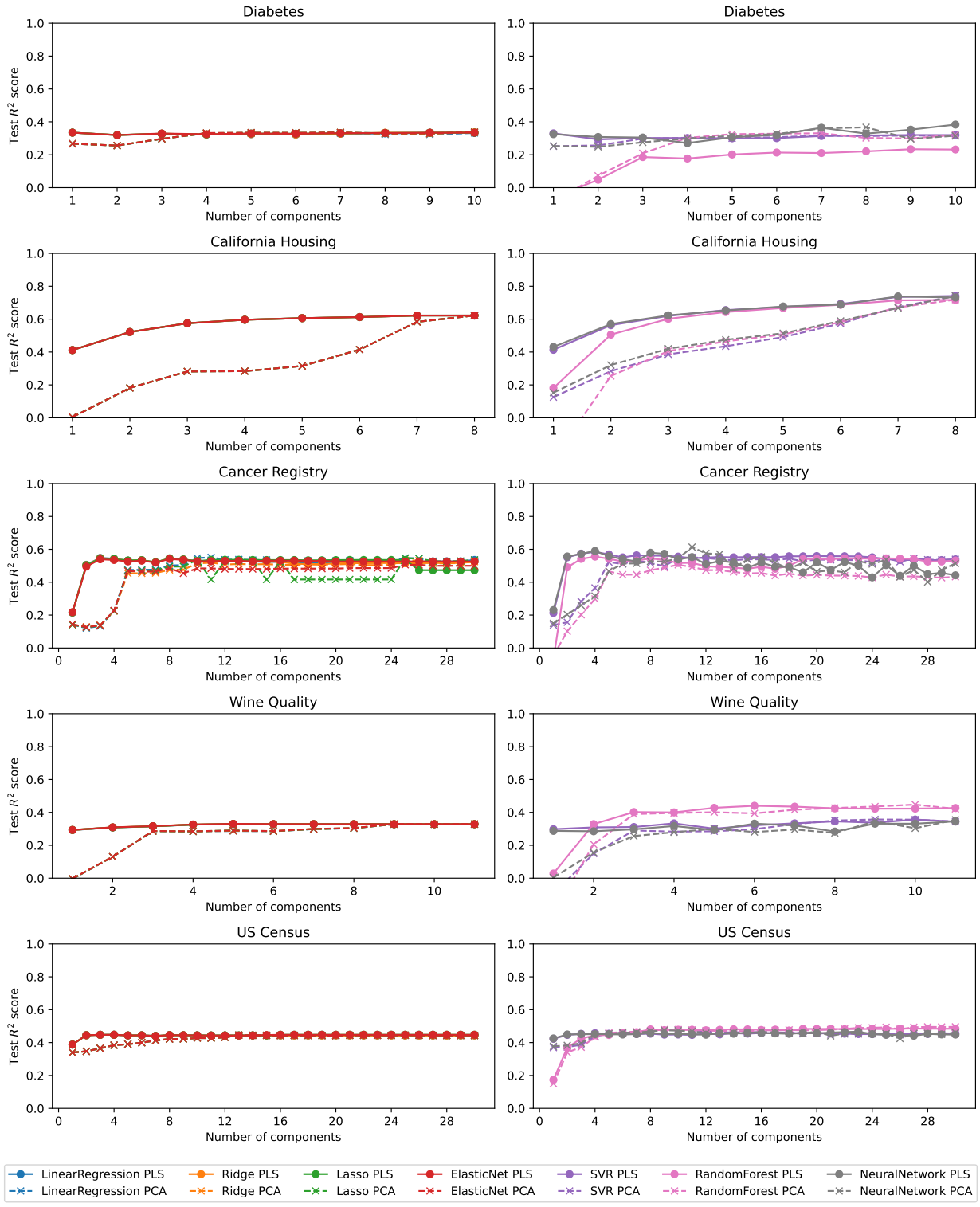

Figure 4.4: PCA and PLS Regression R^2 test scores

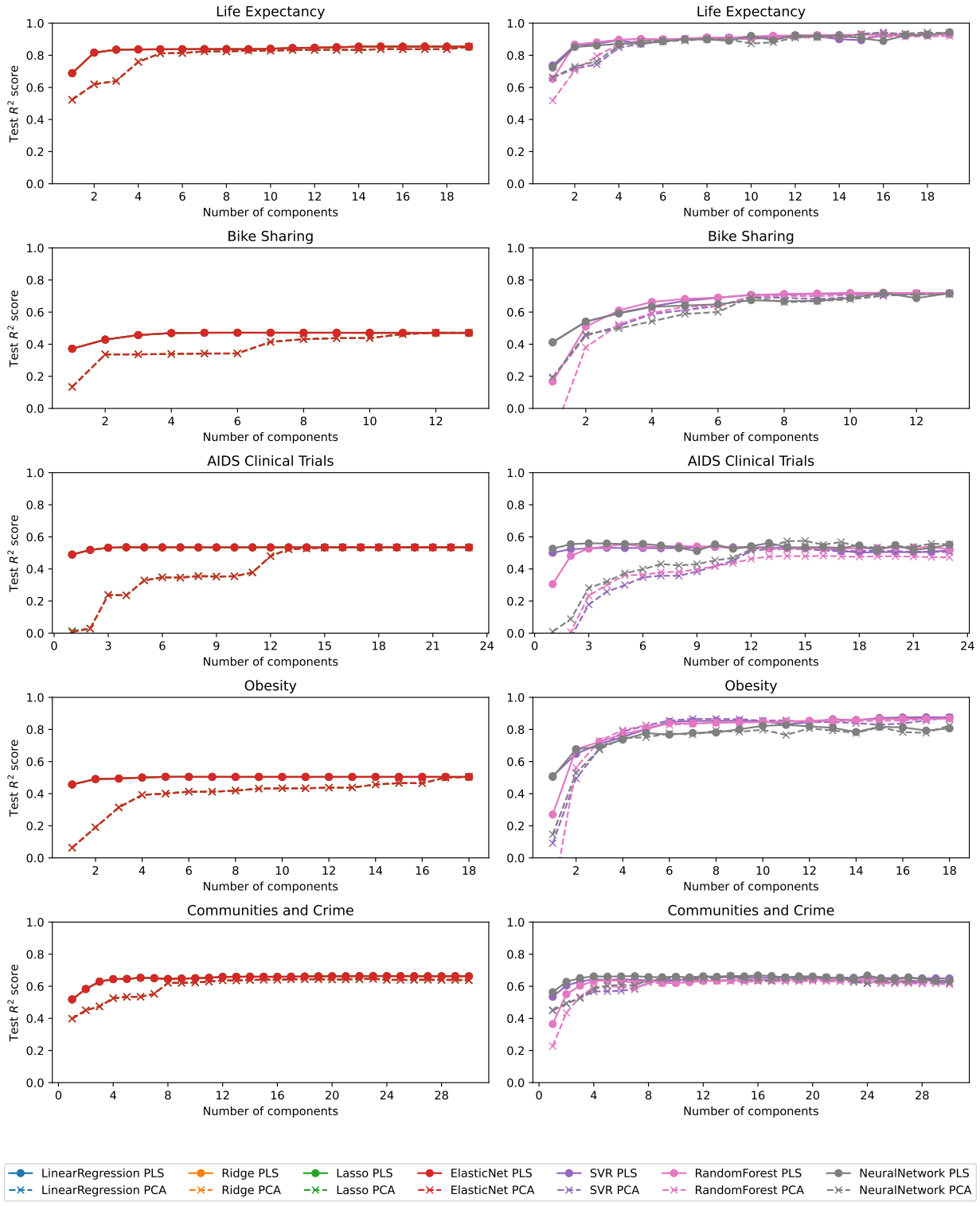

Figure 4.5: PCA and PLS Regression R^2 test scores

		LR	Ridge	Lasso	ElasticNet	SVR	RF	NN
Diabetes	PCA	0.323 (8)	**0.336** (10)	0.330 (8)	0.328 (8)	0.300 (4)	0.273 (9)	0.304 (5)
	PLS	0.328 (3)	0.326 (6)	**0.332** (3)	0.331 (3)	0.280 (2)	0.232 (10)	0.291 (4)
California Housing	PCA	0.622 (8)	0.622 (8)	0.622 (8)	0.622 (8)	0.742 (7)	0.738 (7)	**0.746** (7)
	PLS	0.622 (8)	0.622 (8)	0.622 (8)	0.622 (8)	**0.741** (8)	0.717 (8)	0.731 (8)
Cancer Registry	PCA	0.524 (27)	0.515 (27)	**0.531** (27)	0.524 (27)	0.525 (26)	0.427 (16)	0.479 (14)
	PLS	0.548 (3)	0.514 (16)	0.544 (3)	0.524 (16)	**0.552** (15)	0.482 (17)	0.536 (17)
Wine Quality	PCA	0.332 (10)	0.332 (10)	0.332 (10)	0.333 (10)	0.358 (10)	**0.414** (9)	0.313 (9)
	PLS	0.315 (3)	0.316 (3)	0.329 (5)	0.329 (5)	0.357 (5)	**0.426** (11)	0.332 (6)
US Census	PCA	0.442 (22)	0.442 (22)	0.442 (22)	0.442 (22)	0.458 (22)	**0.473** (26)	0.464 (22)
	PLS	0.443 (23)	0.443 (23)	0.444 (23)	0.444 (23)	0.463 (14)	**0.484** (29)	0.461 (15)
Life Expectancy	PCA	0.854 (19)	0.855 (19)	0.854 (19)	0.855 (19)	**0.937** (19)	0.919 (19)	0.928 (19)
	PLS	0.854 (16)	0.854 (16)	0.854 (16)	0.854 (16)	0.937 (15)	0.929 (19)	**0.944** (19)
Bike Sharing	PCA	0.471 (12)	0.471 (12)	0.471 (12)	0.471 (12)	0.712 (11)	**0.730** (13)	0.697 (7)
	PLS	0.472 (8)	0.472 (8)	0.471 (8)	0.471 (8)	0.714 (8)	0.719 (12)	**0.720** (11)
AIDS Clinical Trials	PCA	0.533 (18)	0.533 (18)	0.533 (18)	0.533 (18)	0.510 (17)	0.532 (16)	**0.563** (16)
	PLS	0.535 (5)	0.535 (5)	0.535 (5)	0.535 (5)	0.537 (4)	0.540 (9)	**0.552** (2)
Obesity	PCA	0.505 (18)	0.505 (18)	0.505 (18)	0.505 (18)	0.876 (17)	**0.877** (18)	0.785 (16)
	PLS	0.505 (7)	0.505 (7)	0.505 (11)	0.505 (8)	**0.876** (4)	0.868 (18)	0.807 (18)
Communities and Crime	PCA	**0.661** (83)	0.660 (83)	0.653 (83)	0.654 (83)	0.644 (94)	0.631 (8)	0.654 (10)
	PLS	0.653 (11)	0.653 (11)	0.662 (20)	0.662 (20)	0.653 (16)	0.639 (5)	**0.662** (4)

Table 4.3: Test scores for the optimum number of components determined by cross validation on the training partition. The best score for each dataset and method is highlighted in bold. Additionally, the best score for each dataset is underlined.

Furthermore, the number of components that produced the best cross validation score on the training dataset (with any combination of hyperparameters) was also calculated. The results are displayed in the first columns of Table 4.3, along with the test score of that regressor refitted on the entire training partition.

When compared to the results obtained with linear regression, there were very minor improvements on the Diabetes, and Cancer Registry datasets. However, in general, the regularization did not improve the results. This is not surprising since one of the main uses of regularization is to reduce overfitting. However, the dimensionality reduction step (PLS or PCA) was probably effective at limiting the amount of overfitting. In fact, one can think of dimensionality reduction as a regularization technique, as it limits the search space of the coefficients. Therefore, considering additional regularization penalties in the regression step may not be necessary.

Finally, we consider the use of non-linear predictors after the dimensionality reduction step. In particular, we consider support vector machine (SVR), random forests (RF), and neural networks (NN). As before, the best hyperparameters are selected using 5-fold cross-validation. The hyperparameter values considered are including in Table 4.4. In the case of the neural network, to avoid overfitting, early stopping was utilized, with a validation partition of 10% and a patience of 20 epochs. That is to say, the training process is halted if there is no improvement in the validation score during 20 epochs.

SVR	
C	$0.1, 1, 10, 100$
gamma	$1, 0.1, 0.001, 0.0001$
kernel	rbf, linear
epsilon	$0.1, 0.01, 0.001$

Random Forest	
Max features	all features, sqrt, log2
Number of estimators	1000

Neural Network	
Hidden layer	(5), (10), (20), (5,5), (10,10), (20,20)
Activation function	logistic, linear, relu
Max epochs	1000
Batch size	100
Optimiser	adam

Table 4.4: Hyperparameters considered in non-linear regressors

As we can see in Table 4.3, except for the Diabetes dataset, the best score is achieved by a non-linear method. However, not all datasets present a significant improvement. We can highlight the accuracy improvements from 0.66 to 0.74 in California Housing, from 0.85 to 0.94 in Life Expectancy and from 0.50 to 0.87 in Obesity. This shows that, even though the dimensionality reduction technique is linear, the use of a non-linear predictor can improve the results notably. The results in the table also highlight that, in many cases, PLS can obtain equivalent results to PCA with fewer components. This is apparent when we compare the number of components that yield in the best cross validation score. For example, in the case of Diabetes, PCA reached its best result with 10 components, in contrast to only 3

for PLS. This same pattern repeats in the Cancer Registry (27 to 15), AIDS Clinical Trials (16 to 2), and Communities and Crime (83 to 4). As we saw when comparing the results with linear regression, both methods obtain very similar scores, but PCA usually needs many more components. This is particularly beneficial for non-linear methods since it can decrease notably the fitting time of complex regressors.

This same behavior can be observed in the plots in Figures 4.4 and 4.5. In particular, the difference is clear in California Housing, Cancer Registry, and AIDS Clinical Trials, where the results of PCA (dashed lines) stay notably below of PLS's results for the first few components. However, in most cases, the advantage of PLS over PCA is slightly lower than when comparing linear regression results.

In conclusion, the experiments in this section show that PLS has a notable advantage over PCA when few components are used. In no dataset had PCA a significant advantage over PLS for any regression technique on any number of components. Moreover, we saw how introducing non-linear regressors can lead to better predictive capabilities for the PLS component even when considering only the first few components.

4.2 Functional regression with scalar response

In this section we carry out an empirical evaluation of PLS in functional regression problems. As discussed in Chapter 2, the functional observations are discretized in a grid fine enough so that the functional nature of the data is apparent. As an example, Figure 4.6 shows the samples of one of the datasets considered, the Tecator dataset.

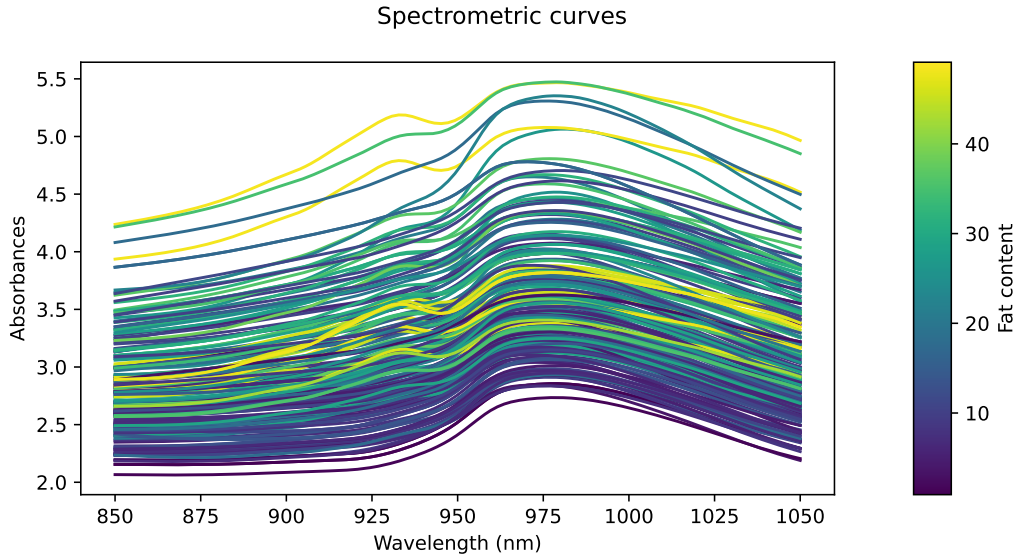

Figure 4.6: Tecator dataset. Each trajectory represents the absorbance of a sample depending on the wave length.

Four datasets have been considered. In the following, each dataset is described, together with any preprocessing steps that were performed. As a summary, Table 4.5 summarizes the most relevant characteristics of each dataset.

	Samples	**Grid size**	**Source**
Tecator	215	100	Tecator company, download: lib.stat.cmu.edu
Sugar content	268	3997	(Munck et al., 1998)
Octane	60	401	UIO Guided Wave Inc., (Segaert et al., 2024)
AEMET	73	365	AEMET, (Febrero-Bande & de la Fuente, 2012)

Table 4.5: Summary of functional datasets

Tecator

The tecator dataset contains the absorbance spectra of 215 meat samples, measured with a Tecator Infratec Food and Feed Analyzer in the wavelength range $850 - 1050$ nm. The goal is to predict the fat content of each sample utilizing the absorbance spectra. The dataset was downloaded from http://lib.stat.cmu.edu/datasets/tecator, while the original source is the Tecator company.

Aside from the original data, during the analysis of the Tecator dataset, the second derivatives of the regressors are usually considered, as they contain most of the information (Ferraty, 2006). Therefore, we have also considered the problem of predicting the fat content from the second derivatives of the regressors.

Sugar content

This dataset contains the emission spectra of 268 samples of sugar solutions. The spectra were measured in the range of $275 - 560$ nm, at different seven different excitation wavelengths. As a preprocessing steps, the seven measurements for each sample have been concatenated, obtaining a single (functional) regressor. The target variable is the ash content, which is an indicator of sugar quality. The data set was downloaded from https://jeffgoldsmith.com/IWAFDA/DataCode/Sugar.RDA, and its original source is Munck et al. (1998).

Octane

The Octane dataset is composed of the near infrared spectra of gasoline samples, with wavelengths in the range $1102 - 1552$ nm. The goal is to predict the octanes from the spectra of each sample. This dataset was downloaded from CRAN, in particular, from the package *mrfDepth* (Segaert et al., 2024). This dataset was originally provided by UOP Guided Wave Inc., USA (Esbensen, 2002, p.221).

AEMET

The AEMET dataset contains the average of the daily temperature, precipitation, and wind speed measured of 72 Spanish weather stations from 1980 to 2009. The goal in this case is to predict the average yearly precipitation from the average daily temperatures. The original source is the Meteorological State Agency of Spain (AEMET), while the data was downloaded from the R packet *fda.usc* (Febrero-Bande & de la Fuente, 2012). As a preprocessing step, the target variable has been transformed by applying a logarithm, with the aim of compensating for differences in scale.

4.2.1 Results

In this section we present first the results obtained by applying linear regression to the extracted PLS and PCA components. For functional data, in principle, there is no upper limit on the number of components that can be extracted. However, in practice, the number of components is limited by the number of discretization points in the grid. For this analysis, we have selected an upper limit of 80 components in all datasets. This limit is well below the number of discretization points in any of the problems, while being high enough to show the evolution of the scores as more components are considered.

The results obtained are displayed in Figure 4.7. As in the multivariate setting, the major advantage of PLS is the rate at which the performance increases within the first few components. This is particularly apparent in the results obtained with the second derivatives of Tecator and the Aemet dataset. Furthermore, in the Aemet dataset, the accuracy of PLS drops when considering more than eight components. To check if this could be due to overfitting, we have also included the R^2 scores on the training partition in the right column of Figure 4.7. We can see how PLS's R^2 score on the training partition already reaches 0.99 with the first eight components, and does not drop as more components are considered. In contrast, the train R^2 score for PCA with eight components is 0.9, and slowly increases as more components are taken into account, only reaching 0.99 after including 34 components or more. Therefore, the drop in the test scores of PLS seems to be caused by overfitting, which is less surprising once we recall that this dataset has only 73 samples.

Following the same logic as in the previous section, we also considered the use of regularization and non-linear predictors. The same hyperparameter grids included in Section 4.1 are used for all regressors except for the neural network. Since the number of samples of these datasets is very low, the runtime of all the algorithms is considerably faster, and we could utilize LBFGS, instead of ADAM as the optimizer. LBFGS is a quasi-Newton optimization method (Liu & Nocedal, 1989), that utilizes an approximation of the Hessian. Compared to ADAM, since LBFGS also takes into account the curvature of the space, it can converge faster in situations where ADAM struggles to converge. Moreover, in these experiments, no major overfitting was observed and, therefore, it was not necessary to utilize early stopping. The final hyperparameter considered are included in Table 4.6.

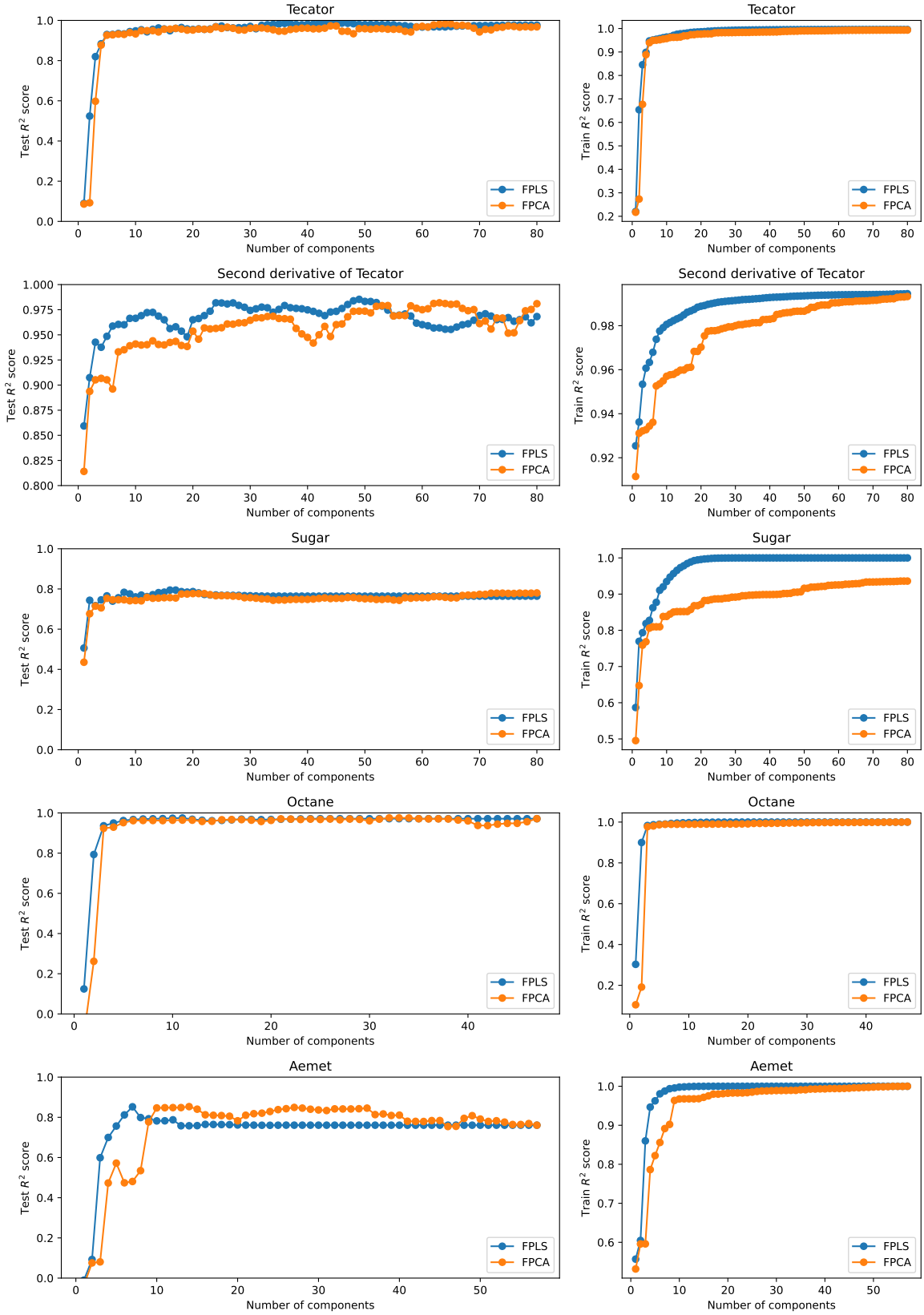

Figure 4.7: R^2 scores for functional PCA and PLS regression

Neural Network	
Hidden layer	(5), (10), (20), (5,5), (10,10), (20,20)
Alpha	0.0001, 0.001, 0.01, 0.1, 1, 10, 100
Activation function	logistic, linear, relu
Max epochs	1000
Optimiser	lbfgs

Table 4.6: Hyperparameters considered for the application of networks with functional regressors

The results obtained considering non-linear regressors are plotted in Figure 4.8. As in the previous section, the results are also presented in a tabular format in Table 4.7. Similarly to the multivariate case, the utilization of regularized methods does not yield notable improvements. The largest difference is obtained in the Sugar dataset, where elastic net achieved a test scores of 0.764 and 0.739 for PCA and PLS, which correspond to improvements of 2% and, 3% over linear regression respectively. This could be an indicator that this dataset is prone to overfitting since the additional regularization step (after dimensionality reduction) improves the results. However, the difference is very small, and we did not observe a decrease of the test score as the number of components increases.

With the exception of the Octane dataset, the best PLS result is obtained when using neural networks. Moreover, the combination of PCA and neural networks also obtains great results. In particular, in the Tecator dataset, the best results increase from 0.95 with linear regression to 0.99 with a neural network. This illustrates the effectiveness of combining linear dimensionality reduction techniques, such as PCA or PLS, with non-linear predictors.

As in previous experiments, the best scores obtained by PLS and PCA are almost equal, but PLS generally reaches the optimum value with fewer components. In the Tecator dataset, PLS reached its maximum with 8 components, compared to 12 for PCA and, in the Sugar dataset, 7 compared to 28. This same phenomenon can be observed in Figure 4.8, where the plots show how the scores obtained by PLS tend to be above those of PCA. However, if we compare the plots on the right and left column, we can see how the introduction of non-linear methods decreases the differences between PCA and PLS when only the first few components are considered. Clear examples of this are the Aemet dataset and the second derivatives of the Tecator dataset. A possible explanation for this is that the capabilities of non-linear methods to capture complex relationships might reduce the impact of the selection of the dimensionality reduction technique.

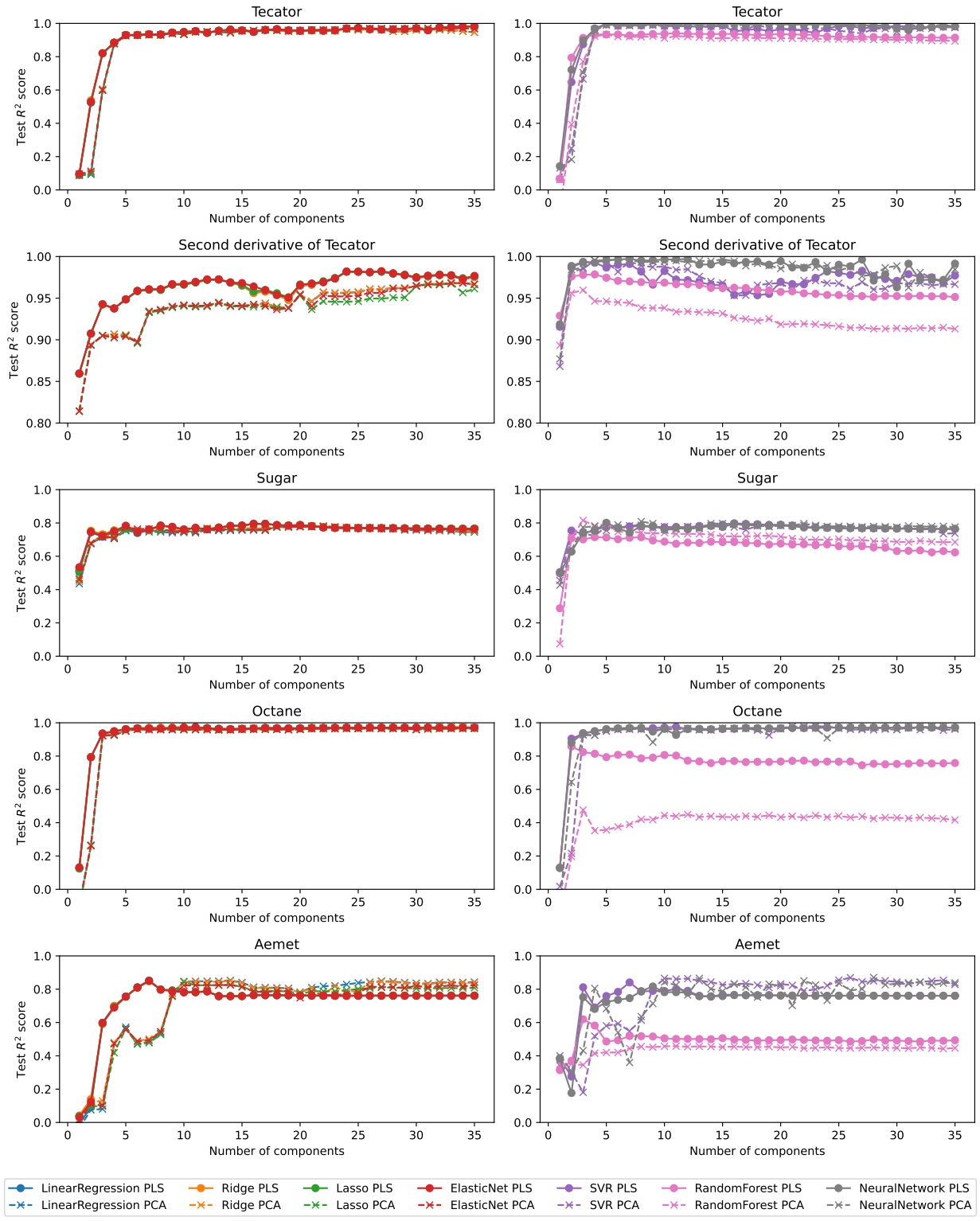

Figure 4.8: Non-linear methods

		L.R.	Ridge	Lasso	ElasticNet	SVR	R.F.	N.N.
Tecator	PCA	0.953(32)	0.953(32)	0.969(34)	0.968(31)	0.984(6)	0.941(5)	**0.999**(12)
	PLS	0.957(13)	0.963(14)	0.965(18)	0.965(18)	0.996(8)	0.930(5)	**0.997**(8)
Tecator derivatives	PCA	0.958(32)	0.958(32)	0.959(32)	0.959(32)	**0.997**(4)	0.973(2)	0.996(6)
	PLS	0.947(13)	0.948(13)	0.961(15)	0.957(14)	0.995(5)	0.977(2)	**0.998**(8)
Sugar	PCA	0.748(12)	0.758(12)	0.752(12)	**0.764**(12)	0.724(15)	0.754(13)	0.744(28)
	PLS	0.718(6)	0.730(6)	0.727(6)	0.739(6)	0.756(7)	0.726(3)	**0.777**(7)
Octane	PCA	0.934(5)	0.934(5)	0.934(5)	0.934(5)	**0.935**(5)	0.745(4)	0.934(5)
	PLS	**0.938**(4)	0.938(4)	0.937(4)	0.937(4)	0.933(3)	0.697(3)	0.938(4)
Aemet	PCA	**0.853**(35)	0.853(35)	0.851(35)	0.853(35)	0.848(28)	0.487(8)	0.780(19)
	PLS	0.801(9)	0.801(9)	0.801(9)	0.801(9)	0.811(8)	0.549(8)	**0.899**(6)

Table 4.7: Test scores for the optimum number of components determined by cross validation on the training partition. The best score for each dataset and method is highlighted in bold. Additionally, the best score for each dataset is underlined.

In summary, the experiments conducted on functional data showed that behavior of PLS does not deviate from the patterns observed on multivariate data. In most cases, the best scores obtained by PLS and PCA are similar. However, PLS converges to the optimum faster. This is to be expected since PLS takes into account the response variable during the dimensionality reduction step. Furthermore, the application of non-linear regression methods on the components extracted by PLS generally leads to better predictions, with neural networks being the most effective in the majority of datasets considered.

Chapter 5

Conclusions and future work

In this master thesis, we have explored the application of PLS to regression problems both with multivariate and with functional data. A generic formulation of PLS was given in which the only requirement is that the regressor variables can be characterized as elements of a Hilbert space. This formulation provides a unified framework to understand the relationships between different PLS variants found in the literature, and show their equivalence. In order to do so, we introduce PLS as an iterative process that builds a sequence of nested subspaces of increasing dimension. Each subspace is generated by a basis, composed of elements in the Hilbert space. At each iteration, PLS extends the basis, by incorporating the element of the Hilbert space that maximizes the covariance with the target variable, subject to some constraints. Depending on the constraints enforced, different PLS basis are obtained. The orthogonal basis is the result of imposing pair-wise orthogonality with respect to the inner product of the Hilbert space. The conjugate basis is obtained by enforcing orthogonality with respect to the conjugate inner product defined under the metric induced by the inverse of the covariance operator of the predictor variables. In both cases, the bases obtained span Krylov subspaces of increasing order. As a result, a third basis can be identified: the Krylov basis, containing the elements obtained by applying repeatedly the regressor covariance operator onto the cross-covariance. While the Krylov basis is useful in theoretical contexts, the two other basis correspond to two of the most widespread algorithms for PLS: NIPALS and conjugate gradients. Both algorithms build the corresponding basis iteratively. However, NIPALS exploits the properties of the orthogonal basis, while conjugate gradients relies on the properties of the conjugate basis.

Additionally, the differences between partial least squares and ordinary least squares regression was studied for multivariate data, and we concluded that this difference is largely determined by the structure of the eigenvalues of the covariance operator. In particular, the convergence of PLS to OLS can be related to a polynomial fitting problem, where the degree of the polynomial is given by the number of components considered, and the points to fit are given by the eigenvalues of the covariance operator. From this reformulation, an upper bound on the differences between regression coefficients obtained by PLS and OLS was derived, which depends only on the eigenvalue distribution of the covariance operator. The conclusions of this analysis are that PLS is expected to be most effective when the

eigenvalues are not close to zero and appear tightly grouped in a few clusters. Moreover, if there are only M distinct eigenvalues, it is sufficient to consider M PLS components.

Finally, the predictive capabilities of PLS regression are evaluated in a variety of problems, both functional and multivariate, and comparing its performance to principal component regression (PCR). The experiments show that the performance of PLS regression is comparable to that of PCR, but utilizing fewer components. Moreover, we also consider combining PCA and PLS with non-linear regression techniques, with the goal of exploiting more complex relationships between the regressors and the target variable. The results show that the combination of PLS and neural networks can be very effective. The dimensionality reduction step can significantly reduce the inputs of the regressor, simplifying and accelerating the fitting process, while the neural networks are capable of capturing complex relationships, greatly surpassing the predictive capabilities of linear methods in some cases.

A possible extension of this work would be to study in detail the impact of the eigenvalue distribution of the regressor covariance operator on the predictive performance of PLS for functional data. In the third chapter of this master thesis, we showed that the presence of eigenvalues that are close to zero can reduce the effectiveness of PLS for multivariate regressors. However, in the functional setting, the eigenvalues have an accumulation point at zero. One investigative line would be to study the performance of PLS when the data is projected onto the space generated by the eigenvectors of the eigenvalues above certain threshold. That is to say, when the information associated with small eigenvalues is discarded. This transformation should accelerate the convergence of PLS. However, it is unclear if this improvement would compensate the discarded information.

Alternatively, different optimization goals could be introduced in the definition of PLS to consider non-linear relationships between the regressors and the target. As an example, Kernel PLS (Rosipal & Trejo, 2001; Yifeng, Jian, & Long, 2006), has already been explored as an alternative to model non-linear relationships. By changing the maximization of the covariance to some other criterion that involves both target and regressor, and that considers non-linear relationships, it could be possible to introduce new variants of PLS.

Furthermore, the approach of the second chapter could be applied to other techniques in the PLS family such as SIMPLS (de Jong, 1993) or PLS-SVD (Wegelin, 2000). By reconsidering these methods utilizing the framework introduced in this work, it is likely that their relationship with the PLS bases described in Chapter 2 could be clarified. Moreover, in doing so, these methods could be generalized to consider regressors in Hilbert spaces.

To conclude, in this work we explored a unified approach that ties the most widely-spread functional PLS formulations with their multivariate counterparts by abstracting them to consider regressors contained in a Hilbert space. Furthermore, we clarified the relationship between the different PLS basis, and the algorithms that utilize them, along with the properties that they exploit. Additionally, we introduced a bound for the difference between the PLS and OLS approximations to the regression coefficients for multivariate regressors that only depends on the eigenvalue structure of the regressor covariance operator. To close this exploration of PLS, its performance was measured in functional and multivariate real-world problems, showing its effectiveness as a dimensionality reduction technique, and the potential of combining it with non-linear regressors.

References

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). WIREs Computational Statistics, 2(1), 97–106. doi: 10.1002/wics.51

Abdi, H., & Williams, L. J. (2010). Principal component analysis. WIREs Computational Statistics, 2(4), 433–459. doi: 10.1002/wics.101

Aguilera, A. M., Ocaña, F. A., & Valderrama, M. J. (1997). An approximated principal component prediction model for continuous-time stochastic processes. Applied Stochastic Models and Data Analysis, 13(2), 61–72. doi: 10.1002/(SICI)1099-0747(199706)13:2<61::AID-ASM296>3.0.CO;2-I

Babii, A., Carrasco, M., & Tsafack, I. (2024). Functional Partial Least-Squares: Optimal Rates and Adaptation (No. arXiv:2402.11134). arXiv. doi: 10.48550/arXiv.2402.11134

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. Journal of Chemometrics, 17(3), 166–173. doi: 10.1002/cem.785

Blazère, M., Gamboa, F., & Loubes, J.-M. (2014). PLS: A new statistical insight through the prism of orthogonal polynomials (No. arXiv:1405.5900). arXiv. doi: 10.48550/arXiv.1405.5900

Bronson, R., & Costa, G. B. (2009). 7 - Matrix Calculus. In R. Bronson & G. B. Costa (Eds.), Matrix Methods (Third Edition) (pp. 213–255). Boston: Academic Press. doi: 10.1016/B978-0-08-092225-6.50013-9

Burnett, A. C., Anderson, J., Davidson, K. J., Ely, K. S., Lamour, J., Li, Q., ... Serbin, S. P. (2021). A best-practice guide to predicting plant traits from leaf-level hyperspectral data using partial least squares regression. Journal of Experimental Botany, 72(18), 6175–6189. doi: 10.1093/jxb/erab295

Cardot, H., Ferraty, F., & Sarda, P. (1999). Functional linear model. Statistics & Probability Letters, 45(1), 11–22. doi: 10.1016/S0167-7152(99)00036-X

Cook, R. D., & Forzani, L. (2018). Big data and partial least-squares prediction. Canadian Journal of Statistics, 46(1), 62–78. doi: 10.1002/cjs.11316

Cook, R. D., & Forzani, L. (2019). Partial least squares prediction in high-dimensional regression. The Annals of Statistics, 47(2), pp. 884–908. doi: 10.1214/18-AOS1681

Cook, R. D., & Forzani, L. (2021). PLS regression algorithms in the presence of nonlinearity. Chemometrics and Intelligent Laboratory Systems, 213, 104307. doi: 10.1016/j.chemolab.2021.104307

Cook, R. D., Forzani, L., & Liu, L. (2023). Partial least squares for simultaneous reduction of response and predictor vectors in regression. Journal of Multivariate Analysis, 196, 105163. doi: 10.1016/j.jmva.2023.105163

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547–553. doi: 10.1016/j.dss.2009.05.016

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. Journal of Statistical Planning and Inference, 147, 1–23. doi: 10.1016/j.jspi.2013.04.002

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 18(3), 251–263. doi: 10.1016/0169-7439(93)85002-X

Delaigle, A., & Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. The Annals of Statistics, 40(1), 322–352. doi: 10.1214/11-AOS958

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. The Annals of Statistics, 32(2), 407–499. doi: 10.1214/009053604000000067

Eldén, L. (2004). Partial least-squares vs. Lanczos bidiagonalization—I: Analysis of a projection method for multiple regression. Computational Statistics and Data Analysis, 46(1), 11–31. doi: 10.1016/S0167-9473(03)00138-5

Ergon, R. (2009). Re-interpretation of NIPALS results solves PLSR inconsistency problem. Journal of Chemometrics, 23, 72–75. doi: 10.1002/cem.1180

Esbensen, K. (2002). Multivariate data analysis : In practice : An introduction to multivariate data analysis and experimental design. Oslo, Norway : CAMO. Retrieved 2024-08-06, from `http://archive.org/details/multivariatedata0000esbe`

Febrero-Bande, M., & de la Fuente, M. O. (2012). Statistical Computing in Functional Data Analysis: The R Package fda.usc. Journal of Statistical Software, 51, 1–28. doi: 10.18637/jss.v051.i04

Febrero-Bande, M., Galeano, P., & González-Manteiga, W. (2017). Functional Principal Component Regression and Functional Partial Least-squares Regression: An Overview and a Comparative Study. International Statistical Review, 85(1), 61–83. doi: 10.1111/insr.12116

Ferraty, F. (2006). Nonparametric functional data analysis. Springer. doi: 10.1007/0-387-36620-2

Frank, I. E., & Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. Technometrics, 35(2), 109–135. doi: 10.2307/1269656

Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2021). Reproducing Kernel Hilbert Space, Mercer's Theorem, Eigenfunctions, Nyström Method, and Use of Kernels in Machine Learning: Tutorial and Survey (No. arXiv:2106.08443). arXiv. doi: 10.48550/arXiv.2106.08443

Hall, P., & Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. The Annals of Statistics, 35(1), 70–91. doi: 10.1214/009053606000000957

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., ... Merigan, T. C. (1996). A Trial Comparing Nucleoside Monotherapy with Combination Therapy in HIV-Infected Adults with CD4 Cell Counts from 200 to 500 per Cubic Millimeter. New England Journal of Medicine, 335(15), 1081–1090. doi: 10.1056/NEJM199610103351501

Helland, I. S. (1990). Partial least squares regression and statistical models. Scandinavian Journal of Statistics, 17(2), 97–114. Retrieved 2024-04-15, from `http://www.jstor.org/stable/4616159`

Hestenes, M. R., & Stiefel, E. (1952). Methods of conjugate gradients for solving linear systems. Journal of research of the National Bureau of Standards, 49, 409–435. doi: 10.6028/jres.049.044

Höskuldsson, A. (1988). PLS regression methods. Journal of Chemometrics, 2(3), 211–228. doi: 10.1002/cem.1180020306

Kelley Pace, R., & Barry, R. (1997). Sparse spatial autoregressions. Statistics & Probability Letters, 33(3), 291–297. doi: 10.1016/S0167-7152(96)00140-X

Krishnan, A., Williams, L. J., McIntosh, A. R., & Abdi, H. (2011). Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. NeuroImage, 56(2), 455–475. doi: 10.1016/j.neuroimage.2010.07.034

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. Mathematical Programming, 45(1), 503–528. doi: 10.1007/BF01589116

Markelle, K., Longjohn, R., & Nottingham, K. (1998). UCI Machine Learning Repository. Retrieved 2024-06-30, from `https://archive.ics.uci.edu/`

Mehmood, T., & Ahmed, B. (2016). The diversity in the applications of partial least squares: An overview. Journal of Chemometrics, 30(1), 4–17. doi: 10.1002/cem.2762

Mercer, J., & Forsyth, A. R. (1997). XVI. Functions of positive and negative type, and their connection the theory of integral equations. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 209(441-458), 415–446. doi: 10.1098/rsta.1909.0016

Mezzadri, F. (2007). How to generate random matrices from the classical compact groups. Notices of the American Mathematical Society, 54(5), 592–604. doi: 10.48550/arXiv.math-ph/0609050

Moindjié, I.-A., Dabo-Niang, S., & Preda, C. (2023). Classification of multivariate functional data on different domains with Partial Least Squares approaches. Statistics and Computing, 10(1), 5. doi: 10.1007/s11222-023-10324-1

Munck, L., Nørgaard, L., Engelsen, S. B., Bro, R., & Andersson, C. A. (1998). Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. Chemometrics and Intelligent Laboratory Systems, 44(1), 31–60. doi: 10.1016/S0169-7439(98)00074-4

Nakua, H., Yu, J.-C., Abdi, H., Hawco, C., Voineskos, A., Hill, S., ... Ameis, S. H. (2024). Comparing the stability and reproducibility of brain-behaviour relationships found using canonical correlation analysis and partial least squares within the ABCD sample. Network Neuroscience, 1–52. doi: 10.1162/netn_a_00363

Nguyen, D., & Rocke, D. (2002). Tumor Classification by Partial Least Squares Using Microarray Gene Expression Data. Bioinformatics (Oxford, England), 18, 39–50. doi: 10.1093/bioinformatics/18.1.39

Nocedal, J., & Wright, S. J. (Eds.). (1999). Numerical Optimization. New York: Springer-Verlag. doi: 10.1007/b98874

Noonan, R., & Wold, H. (1977). NIPALS Path Modelling with Latent Variables. Scandinavian Journal of Educational Research - SCAND J EDUC RES, 21, 33–61. doi: 10.1080/0031383770210103

Okwuashi, O., Ndehedehe, C., & Attai, H. (2020). Tide modeling using partial least squares regression. Ocean Dynamics, 70(8), 1089–1101. doi: 10.1007/s10236-020-01385-1

Palechor, F. M., & Manotas, A. D. L. H. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data in Brief, 25, 104344. doi: 10.1016/j.dib.2019.104344

Preda, C., & Saporta, G. (2005). PLS regression on a stochastic process. Computational Statistics and Data Analysis, 48(1), 149–158. doi: 10.1016/j.csda.2003.10.003

Ramsay, J., & Silverman, B. (2013). Functional Data Analysis. Springer New York. doi: 10.1007/b98888

Redmond, M., & Baveja, A. (2002). A data-driven software tool for enabling cooperative information sharing among police departments. European Journal of Operational Research, 141(3), 660–678. doi: 10.1016/S0377-2217(01)00264-8

Rosipal, R., & Krämer, N. (2005). Overview and Recent Advances in Partial Least Squares. Lecture Notes in Computer Science, 3940, 34–51. doi: 10.1007/11752790_2

Rosipal, R., & Trejo, L. (2001). Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. Journal of Machine Learning Research, 2, 97–123. doi: 10.1162/15324430260185556

Sathishkumar, E., Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. Computer Communications, 153, 353–366. doi: 10.1016/j.comcom.2020.02.007

Segaert, P., Hubert, M., Rousseeuw, P., Raymaekers, J., & Vakili, K. (2024). mrfDepth: Depth Measures in Multivariate, Regression and Functional Settings. Retrieved 2024-08-25, from https://cran.r-project.org/web/packages/mrfDepth/index.html

Sharma, R., & Bhandari, R. (2015). Skewness, kurtosis and Newton's inequality. Rocky Mountain Journal of Mathematics, 45(5), 1639–1643. doi: 10.1216/RMJ-2015-45-5-1639

Ståhle, L., & Wold, S. (1987). Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. Journal of Chemometrics, 1(3), 185–196. doi: 10.1002/cem.1180010306

Takane, Y., & Loisel, S. (2016). On the PLS Algorithm for Multiple Regression (PLS1). In H. Abdi, V. Esposito Vinzi, G. Russolillo, G. Saporta, & L. Trinchera (Eds.), The Multiple Facets of Partial Least Squares and Related Methods (pp. 17–28). Cham: Springer International Publishing. doi: 10.1007/978-3-319-40643-5_2

Wang, H., Gu, J., Wang, S., & Saporta, G. (2019). Spatial partial least squares autoregression: Algorithm and applications. Chemometrics and Intelligent Laboratory Systems, 184, 123–131. doi: 10.1016/j.chemolab.2018.12.001

Wegelin, J. (2000). A Survey of Partial Least Squares (PLS) Methods, with Emphasis on the Two-Block Case. In Technical Report. Department of Statistics, University of Washington, Seattle. Retrieved 2023-10-23, from https://www.semanticscholar.org/paper/A-Survey-of-Partial-Least-Squares-(PLS)-Methods%2C-on-Wegelin/73a35426f4943e48c442fe060423289e0350b039

Williams, V. V. (2014). Multiplying matrices in $O(n^{2.373})$ time.. Retrieved 2024-06-03, from https://www.semanticscholar.org/paper/Multiplying-matrices-in-O(n-2%3A373-)-time-Williams/fa9e397b5fcac5010ba9acc33270a4e15ffbd4a1

Wold, S., Ruhe, A., Wold, H., & Dunn, W. J., III. (1984). The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. SIAM Journal on Scientific and Statistical Computing, 5(3), 735–743. doi: 10.1137/0905052

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2), 109–130. doi: 10.1016/S0169-7439(01)00155-1

Yifeng, B., Jian, X., & Long, Y. (2006). Kernel Partial Least-Squares Regression. In The 2006 IEEE International Joint Conference on Neural Network Proceedings (pp. 1231–1238). doi: 10.1109/IJCNN.2006.246832

Zhang, W., Han, J., & Deng, S. (2017). Heart sound classification based on scaled spectrogram and partial least squares regression. Biomedical Signal Processing and Control, 32, 20–28. doi: 10.1016/j.bspc.2016.10.004