Partial least squares for funtional data

Daniel Perdices Burrero

Máster en Matemáticas y Aplicaciones

MÁSTERES
DE LA UAM
2018 - 2019

Facultad de Ciencias

UAM Universidad Autónoma de Madrid

UNIVERSIDAD AUTÓNOMA DE MADRID

FACULTAD DE CIENCIAS

FACULTAD DE CIENCIAS
UNIVERSIDAD AUTÓNOMA DE MADRID

UAM
Universidad Autónoma de Madrid

Master's Degree in Mathematics and Applications

MASTER'S THESIS

PARTIAL LEAST SQUARES FOR FUNCTIONAL DATA

Author: Daniel Perdices Burrero

Advisor: José Ramón Berrendero Díaz

Department of Mathematics

September 2019

Partial Least Squares for Functional Data

Author: Daniel Perdices Burrero
Advisor: José Ramón Berrendero Díaz

Departament of Mathematics
Facultad de Ciencias
Universidad Autónoma de Madrid

September 2019

ABSTRACT

Abstract Linear regression is an useful tool to study the relationship between different variables or address classification and prediction problems. Furthermore, functional data analysis is a field of growing attention in statistics and provides a framework to deal with infinite-dimensional data. In this work, we present a study of different approaches to partial least squares (PLS) for functional data. For this purpose, we first examine the finite-dimensional multivariate case and present different points of view of PLS that lead to different properties. This provides different possible extensions of PLS for functional data, which are not totally straightforward since both data and parameter space have infinite dimension and, as a consequence, some issues arise. Moreover, we also provide a comparison with synthetic and real data that benchmarks PLS against principal component regression (PCR). The final objective is to show that PLS is a good alternative to PCR and it can perform better in some scenarios or, at least, it achieves the same performance.

AGRADECIMIENTOS

This chapter is intentionally in Spanish.

En primer lugar, el agradecimiento institucional a la Universidad Autónoma de Madrid por la ayuda para el inicio de los estudios de máster así como la de fomento de la investigación (que tuve finalmente que rechazar por mi situación laboral), que me ayudaron a afrontar las tasas.

Cuando uno finaliza el grado, tiene la perspectiva que la continuación natural es el máster y, por tanto, que el máster va a ser muy similar al grado. Para mi sorpresa, no tiene nada que ver. Durante el grado, uno se acostumbra a muchas pruebas de evaluación, a un profesor muy distante que ni te conoce y a clases abarrotadas. Sin embargo, el ambiente es sorprendentemente muy distinto: se ve a profesores muy implicados por el proceso de aprendizaje de los alumnos, clases de 15 personas, aprobar mediante ejercicios, etc.

También se nota en los compañeros y en el ambiente: los compañeros colaboramos entre nosotros para tareas como hacer los ejercicios o buscar la bibliografía, tienes compañeros que hablan contigo en inglés, que vienen de carreras muy diferentes... Por todo esto, hay que agradecer a los profesores y coordinadores del máster su trabajo y actitud y la construcción de un ambiente tan idílico. También a los compañeros, que en todo momento se ha creado un ambiente sano y agradable de cooperación, amistad y ayuda mutua. Especial mención a los compañeros de *la salita*, que tantos ejercicios hemos hecho juntos.

También, agradezco en especial a mi tutor, José Ramón Berrendero, la guía y ayuda a lo largo de este trabajo que ha pasado por etapas difíciles. En especial, agradecer todo el esfuerzo de corregir y revisar este documento en vacaciones y fines de semana.

También quiero agradecer a mi familia, en especial a mis padres, por todas esas veces que me han llevado a la universidad y al médico cuando me encontraba mal, que me han preparado la comida y apoyado en todo momento. También, a una pequeña *"coautora"* canina de este documento, Milka, que se ha pasado alguna vez por encima del teclado de mi ordenador y que se ha puesto a mi lado a dormir y soltar algún lametón.

Para acabar, me gustaría agradecerte todo, Ana, tu apoyo, amor, cariño y ánimos durante esta etapa. Ahora en el futuro inminente te toca a ti pasar por esto y espero poder corresponderte al máximo. Gracias por todo y por todos los momentos que vamos a pasar juntos.

Contents

List of Tables

List of Figures

List of Algorithms

Acronyms

FDA Functional Data Analysis.

MAD Median Absolute Deviation.

MSE Mean Square Error.

MSPE Mean Square Prediction Error.

PCR Principal Components Regression.

PLS Partial Least Squares.

SVD Singular Value Decomposition.

1 INTRODUCTION

1.1 Motivation

Nowadays, machine learning, deep learning, or data science and analytics are topics where both industry and academy are investing resources in. Although data is normally massive, it is important and profitable for companies. In this light, they face problems such as data heterogeneity (i.e. data is available in different schemes: images, time series, categorical variables, etc.), data dimensionality (i.e. data contains thousands of variables, sometimes even more than the number of available samples) and collinearity and dependence (i.e. correlation among variables is too high or variables are redundant).

Consequently, statistics must deal with several challenges to analyze this data. Many problems in statistics are formalized in terms of two spaces: the sample space $\mathcal{X}$ and the parameter space Θ. When the sample size (hereinafter n) is much bigger than the dimension of the both $\mathcal{X}$ (hereinafter d) and Θ (hereinafter m), classic multivariate statistics techniques usually apply.

Nevertheless, current problems go far beyond that and we run into a critical obstacle: d and m can be bigger than n and even in some cases, the dimension of these spaces might be infinity. This is the case of both high dimensional problems ($n < d$) and Functional Data Analysis (FDA) ($\mathcal{X}$ is a function space such as $L^2([0,1])$, e.g. samples are signals or time series). Table 1.1 summarizes the differences between univariate statistics, multivariate statistics, high dimensional problems, and functional data. Also, FDA copes with other issues due to the structure of the sample curves. For instance, if sample curves are continuous, we expect that the value of the curve at two close points must be similar, i.e. they are highly correlated.

To deal with the aforementioned modern challenges, regularization and smoothing techniques, shrinkage methods, and variable selection methods cope with some issues such as dimensionality and collinearity. These methods reduce the dimension avoiding over-fitting, ill-posed or unsolvable problems due to the dimensions of Θ and $\mathcal{X}$.

	Univariate	Multivariate	High dimensional	Functional data
Model	Random variable	Random vector	Random vector	Stochastic process
Samples	$\mathcal{X} = \mathbb{R}$	$\mathcal{X} = \mathbb{R}^d$ $(d \ll n)$	$\mathcal{X} = \mathbb{R}^d$ $(d > n)$	$\mathcal{X} = L^2([0,1])$
Parameters	$\Theta \subset \mathbb{R}^m$	$\Theta \subset \mathbb{R}^m$ $(m \ll n)$	$\Theta \subset \mathbb{R}^m$	$\Theta \subset \mathbb{R}^m$ or $L^2([0,1])$

Table 1.1: Comparison of FDA with univariate statistics, multivariate statistics and high dimensional data.

This work covers a shrinkage technique, Partial Least Squares (PLS), for the linear regression model with functional data, where both $\mathcal{X}$ and Θ belong to $L^2([0,1])$. We have focused in only one particular method and problem, but there are many more open problems that could be studied in this setting. The technique we have chosen illustrates how the application of a certain method may strongly depend on the type of data, finite dimensional or functional. The decision of choosing PLS over other techniques has a particular reason behind: state of the art of PLS is confusing and often hard to understand. Thus, one of the goals of this work is to clarify several aspects of PLS both in the finite dimensional and the functional cases.

1.2 State of the art

The topic of PLS has been studied both in the multivariate case and in the functional case. PLS was originally proposed in [Wold, 1966] and even the name was different: projection to latent structures. The proposed algorithm in this paper is usually referred as NIPALS [Wold, 1975] and its applications were focused on sociology. Nowadays, it is most widely used in chemometrics and also in other areas such as bioinformatics, medicine or neuroimaging [Krishnan et al., 2011].

Several PLS versions have been developed throughout the years, with different properties. The version known as SIMPLS [de Jong, 1993] is one of the most popular ones because it is equivalent to NIPALS if the response is scalar and it is significantly faster than NIPALS. In particular, S. de Jong is a prolific author in this topic, [Phatak and de Jong, 1997], [de Jong, 1995] and [Jong, 1993] are remarkable examples.

Many different flavours of PLS have been documented in the state of the art. For instance, [Andersson, 2019] compares nine PLS algorithms including SIMPLS and NIPALS. This comparison ranks algorithms with different criteria such as numerical stability or speed. On the other side, we focus on an agnostic point of view as long as algorithms solve the same problem, i.e. the obtained result is the same under infinite precision arithmetic.

Although the state of the art of PLS in the multivariate case is rich, the situation in FDA is totally different. One the most cited and relevant books on FDA is [Ramsay and Silverman, 1997]. In particular, this book covers the functional regression problem both with scalar and functional responses using different methods. Among these methods, the most relevant ones are: regularization, based on a penalized least squares regression, and basis smoothing or truncating, based on a truncated basis expansion. Although it is

a key reference of basis fitting and a wide variety of topics, PLS is not one of these topics.

Moreover, [Ferraty and Vieu, 2006] is a popular book in this area that covers selected topics such as classification or regression asymptotic results. Furthermore, a few comments can be found about PLS, but not directly on the functional regression problem but on the construction of a semi-metric, i.e. tools to show the curves in a reduced dimensional space.

Aside from books, [Preda and Saporta, 2005] proposed a version of PLS for FDA inspired in finite dimensional approaches. They claim the consistency and convergence with the number of components, although proofs of these facts can not be found in the paper.

On the other hand, [Delaigle and Hall, 2012] proposed a different approach called alternative PLS (APLS) based on a nonorthogonal basis. For this procedure, they were able to prove both consistency under certain regularity conditions and a convergence rate for the algorithm. They insisted on the simplicity of the algorithm compared to standard PLS. Their results require highly technical proofs that go step by step proving the consistency of several intermediate estimated operators or quantities.

Furthermore, [Febrero-Bande et al., 2017] performed an interesting comparison among different regularization techniques for the functional regression problem: PCR, PLS and Ridge regression. They provided an approach to PLS similar to [Preda and Saporta, 2005] and the present work contains some simulations fairly similar to the ones they described.

1.3 Organization of the document

This document is divided in six chapters. First, this chapter provides a introduction to PLS and functional data to put everything in a context. Chapter 2 adds some preliminary results that are useful to understand the document. Chapter 3 address several definitions of PLS for the finite dimension model, whereas chapter 4 explores same concepts for functional data. Next, results and simulations are performed and results are shown in chapter 5. Finally, chapter 6 is reserved for final comments and conclusions.

PRELIMINARY RESULTS

2

2.1 Introduction

This chapter presents a collection of results that are useful for the complete understanding of the whole document. These results are classified into several categories: stochastic process theory, functional data analysis, reproducing kernel Hilbert spaces, and other results.

2.2 Stochastic Process Theory

This section covers some required aspects about the covariance function, integrability and Gaussian processes.

2.2.1 Covariance function and stationary processes

Several of the operators that are commonly used in functional data statistics are related to the covariance function and its properties. Continuity, symmetry, normality, and other properties play an important role and can help to obtain stronger results.

Definition 2.2.1 (Covariance function). *Let $\{X(t); t \geq 0\}$ be an L^2 process taking complex values. The covariance function of this process is defined as*

$$K(s,t) = \text{Cov}(X(s), X(t)) = E\left[(X(s) - m(s))\overline{(X(t) - m(t))}\right],$$

where $m(t) = E(X(t))$.

Some useful properties are:

1. $K(s,t) = E((X(s) - m(s))\overline{(X(t) - m(t))}) = E(X(s)\overline{X(t)}) - m(s)\overline{m(t)}$
2. $|K(s,t)|^2 \leq \|X(s) - m(s)\|^2 \|X(t) - m(t)\|^2 = K(s,s)K(t,t)$
   (Cauchy-Schwarz inequality in $L^2(\Omega)$)

3. $K(s,t) = \overline{K(t,s)}$ (Hermitian or symmetric in real-valued spaces)

Although the results also hold in a complex-valued linear space of functions, we are only interested in real-valued spaces of functions.

Definition 2.2.2 (Stationary or stationary in a weak sense)**.** *Let $\{X(t); t \in T\}$ be an L^2 process. Let $m(t) = E(X(t))$ and let K be its covariance function. It is said to be stationary in a weak sense if:*

1. *$m(t) = m(t+h)$ for all $t, t+h \in T$. (So m is constant.)*
2. *$K(s+h, t+h) = K(s,t)$ for all $s, t, s+h, t+h \in T$.*

In other words, the covariance function is a one variable function since it only depends on the difference between s and t, $K(s,t) = K(s-t, 0)$. Abusing the notation, $K(t) := K(s+t, s)$. By applying the Cauchy-Schwarz inequality,

$$|K(t)| = |K(s+t,s)| \leq \sqrt{K(s+t,s+t)K(s,s)} = K(0) = \|X(s) - m\|^2 .$$

Definition 2.2.3 (Strictly stationary)**.** *Let $\{X(t); t \in T\}$ be an L^2 process. Let $F_{t_1,...,t_n}(x_1, ..., x_n) = P(X(t_1) \leq x_1, ..., X(t_n) \leq x_n)$. It is said to be strictly stationary if*

$$F_{t_1,...,t_n} = F_{t_1+h,...,t_n+h},$$

for all $t_1 < t_2 < ... < t_n$ and $t_i, t_i + h \in T$, and for all $i = \{1, ..., n\}$.

This means that the joint distribution of $X(t_1), ..., X(t_n)$ does only depend on the differences $t_2 - t_1$, ..., $t_n - t_{n-1}$.

Proposition 2.2.4. *Let $\{X(t); t \in T\}$ be an L^2 process. If it is strictly stationary, then it is stationary in a weak sense.*

Proof.

$$\langle X(s), X(t) \rangle = \iint xy \; dF_{s,t}(x,y) = \iint xy \; dF_{s+h,t+h}(x,y) = \langle X(s+h), X(t+h) \rangle$$

$$E((X(t))) = \int x dF_t(x) = \int x dF_{t+h}(x) = E((X(t+h)))$$

□

The proof shows that strict stationarity is stronger than weak stationarity, since stationarity in the weak sense only place restrictions on the first two moments. The converse is clearly false.

2.2.2 Gaussian processes and continuity

Brownian Motion is a really common Gaussian process that appear in many applications in physics or economics. We described briefly some properties of Brownian Motion and Gaussian processes.

Definition 2.2.5 (Gaussian process)**.** *Let $\{X(t); t \in T\}$ be an L^2 process. It is said to be a Gaussian process if for every $n \in \mathbb{N}$ and any finite sequence $t_1 \leq t_2 ... \leq t_n$, the vector [$X(t_1)$, ..., $X(t_n)$] is a multidimensional normal vector.*

Definition 2.2.6 (Brownian Motion)**.** *A Brownian Motion $\{B(t); t \geq 0\}$ is a Gaussian process with $E(B(t)) = 0$ and $K(s,t) = \sigma^2 \min(s,t)$ with $t, s \geq 0$ and $\sigma > 0$*

Proposition 2.2.7. *Let $\{B(t);\ t \geq 0\}$ be a stochastic process. The following are equivalent:*

1. *It is a Brownian Motion.*
2. *It is a process with independent increments such that $B(0) = 0$ and $B(s) - B(t) \sim N(0, \sigma^2|s-t|)$.*

Also, some useful properties are:

1. $B(0) = 0$ a.s.
2. Let $0 \leq t_1 \leq t_2 ... \leq t_{2n}$. [$B(t_2) - B(t_1)$, ... , $B(t_{2n}) - B(t_{2n-1})$] are uncorrelated. Since they are a multidimensional normal distribution with no correlation, they are also independent. This is the key idea to prove Proposition 2.2.7.
3. $\{B(t+h) - B(t); t \geq 0\}$ is a stationary process. It is also said that $\{B(t); t \geq 0\}$ has stationary increments.

2.2.3 Second order calculus

In next section, the continuity of the process is important to ensure the continuity of the covariance function and of some operators associated with it. Also, it is necessary to define briefly the meaning of L^2-integrability.

Definition 2.2.8 (L^2-continuous process)**.** *Let $\{X(t);\ t \in T\}$ be an L^2 process. It is said to be L^2-continuous if $X(t+h) \xrightarrow{L^2} X(t)$ as $h \to 0$.*

This subsection just recall some basic results of functional analysis in L^2. This first lemma follows from the continuity of inner product.

Lemma 2.2.9. *If $Y_n \xrightarrow{L^2} Y$ and $X_n \xrightarrow{L^2} X$. Then $\langle Y_n, X_n\rangle \to \langle Y, X\rangle$.*

Lemma 2.2.10. *Let $\{X(t);\ t \in T\}$ be an L^2 process. Let $t_0 \in T$. Then the following are equivalent:*

1. *There exists a $X \in L^2$ such that $X(t) \xrightarrow{L^2} X$ as $t \to t_0$.*
2. *There exists a constant L such that for all sequences t_n and t'_n that tend to t_0, $\langle X(t_n), X(t'_n)\rangle \to L$ as $n, m \to \infty$.*

Continuity of processes is necessary for some results and, in some cases, stronger results can be given if the process is stationary.

Theorem 2.2.11. *Let $\{X(t);\ t \in T\}$ be an L^2 process with $m(t) = E(X(t))$ a continuous function and K the covariance function. Then the following are equivalent:*

1. *The process is L^2-continuous at r.*
2. *K is continuous at (r, r).*

Corollary 2.2.11.1. *Let $\{X(t);\ t \in T\}$ be an L^2 process. If it is continuous at (r, r) for every r, it is continuous at (s, t) for every s, t.*

Corollary 2.2.11.2. *Let $\{X(t);\ t \in T\}$ be an L^2 stationary process. If it is continuous at r for some r, it is continuous at 0. If it is continuous at 0, it is continuous everywhere*

Theorem 2.2.12. *Let $\{X(t);\ t \in T\}$ be an L^2 stationary process with covariance function K. If it is differentiable everywhere, then K is twice differentiable and $\{X'(t);\ t \in T\}$ is a stationary process with covariance $-K''$.*

We will consider in many points of this work the integral of an stochastic process multiplied by a function. Therefore, it is necessary to properly state the definition of this integral:

Definition 2.2.13 (L^2 integral)**.** *Let $\{X(t);\ a \le t \ge b\}$ be an L^2 process with covariance K and mean m. Let g a complex-valued function. Let Δ be a partition $a = t_0 \le t_1 ... \le t_n = b$ and $|\Delta| = \max_{1\le i\le n}(t_i - t_{i-1})$. We define I as follows:*

$$I(\Delta) = \sum_{k=1}^{n} g(t_k)X(t_k)(t_k - t_{k-1}).$$

If $I(\Delta)$ converges in L^2 to some random variable I as $|\Delta| \to 0$, then it is said that $g(t)X(t)$ is L^2-integrable on $[a, b]$ and

$$I = \int_a^b g(t)X(t)dt.$$

As mentioned before, continuity and integrability are related and, as long as the process is continous (m and K are continous), the process is L^2-integrable.

Theorem 2.2.14 (L^2-integrability sufficient condition)**.** *Let $\{X(t);\ a \leq t \geq b\}$ be an L^2 process with covariance function K and mean m. Let g be a continuous functions on $[a, b]$. If K is continuous on $[a, b] \times [a, b]$ and m is continuous on $[a, b]$, then $g(t)X(t)$ is L^2-integrable on $[a, b]$.*

Also, the zero mean property can lead to interesting results on $E[\langle g, X\rangle]$ and $E[\langle g, X\rangle\langle h, X\rangle]$. In particular, the zero mean of $\langle g, X\rangle$ is interesting for our regression problem and allows simplifications if both X and Y are centered.

Theorem 2.2.15. *Let $\{X(t);\ a \leq t \geq b\}$ be an L^2 process with covariance function K and mean $m = 0$. Let g, h continuous functions and K continuous. Then the following equalities hold:*

$$E\left[\int_a^b g(s)X(s)ds\right] = E\left[\int_a^b h(t)X(t)dt\right] = 0,$$

and

$$E\left[\int_a^b g(s)X(s)ds\overline{\int_a^b h(t)X(t)dt}\right] = \int_a^b \int_a^b g(s)\overline{h(t)}K(s,t)dsdt.$$

These results are important since they mean that if X has zero mean, the regression predictor $\langle g, X\rangle$ also has zero mean and the covariance of $\langle g, X\rangle$ with $\langle h, X\rangle$ is:

$$\begin{aligned}
\mathrm{Cov}(\langle g, X\rangle, \langle h, X\rangle) &= E\left[(\langle g, X\rangle - E(\langle g, X\rangle))(\langle h, X\rangle - E(\langle h, X\rangle))\right] \\
&= E\left[\langle g, X\rangle\langle h, X\rangle\right] && \text{(Zero mean)} \\
&= \int_a^b \int_a^b g(s)\overline{h(t)}K(s,t)dsdt
\end{aligned}$$

Theorem 2.2.16. *Let $\{X(t);\ a \leq t \geq b\}$ be an L^2 process with covariance function K and mean $m = 0$. Let g be a continuous functions and K continuous. Then,*

$$E\left[X(s)\overline{\int_a^b g(t)X(t)dt}\right] = \int_a^b \overline{h(t)}K(s,t)dt.$$

Hereinafter, we will be assuming these results. Consequently, we will assume that $m = E(X(t))$ and $K(t,s)$ are continuous functions.

2.3 Functional Data Analysis

This section covers some results about the eigenvalues of the covariance operator and how this can be applied to decompose the stochastic process.

Theorem 2.3.1 (Mercer's theorem)**.** *Let K be a continuous symmetric nonnegative definite kernel. Then there is an orthonormal basis $\{e_n\}_{n\geq 1}$ of $L^2[a,b]$ consisting of eigenfunctions with nonnegative eigenvalues $\{\lambda\}_{i\geq 1}$ such that K has the following representation:*

$$K(s,t) = \sum_{j=1}^{\infty} \lambda_j e_j(s)\overline{e_j(t)},$$

where the convergence of the series is both absolute and uniform.

This decomposition leads to the Karhunen-Loève expansion.

Theorem 2.3.2 (Karhunen-Loève Expansion)**.** *Let $\{X(t);\ a \leq t \geq b\}$ be an L^2 process with continuous covariance function K and mean $m = 0$. Let $\{e_n\}_{n\geq 1}$ be an orthonormal basis for the space of eigenfunctions with nonzero eigenvalues of the kernel K and $\{\lambda_n\}_{n\geq 1}$ the corresponding eigenvalues. Then $X(t)$ can be expressed as*

$$X(t) = \sum_{n=1}^{\infty} Z_n e_n(t), \qquad a \leq t \leq b$$

where $Z_n = \langle X(t), e_n(t)\rangle$ are orthogonal random variables with $E(Z_n) = 0$ and $E(|Z_n|^2) = \lambda_n$ and this series converges in L^2 uniformly in t.

Also, if process is Gaussian, it leads to a stronger result.

Corollary 2.3.2.1. *Let $\{X(t);\ a \leq t \geq b\}$ be an L^2 process with continuous covariance function K and mean $m = 0$. If the process is Gaussian, $[Z_1, ...Z_k]$ are jointly Gaussian for every k with $\mathrm{Cov}(Z_i, Z_j) = \delta_{i,j}$ and therefore they are independent.*

2.4 Regularization techniques

This section describes other regularization techniques that are related to PLS.

First, we define the principal component regression, i.e. the regression over the principal components of the process. These components are known to be direction in which the variance is maximized.

Definition 2.4.1 (Principal Component Regression, PCR)**.** *Let X be a stochastic process under the assumptions of Theorem 2.3.2. The principal components of X are the $e_n(t)$ of Theorem 2.3.2, and we define the principal component regression estimates:*

$$\hat{\beta}_k^{(PCR)} = \underset{\beta \in span\{e_1,\dots,e_k\}}{\arg\min} E(\langle X, \beta\rangle - Y)^2.$$

In the definition above, Y stands for the (scalar) response variable in the regression model (to be defined more precisely later on).

Another possibility to avoid overfitting and bias is to add a penalty to the least squares expression:

Definition 2.4.2 (Ridge Regression, RR)**.** *Let X be an L^2 process. We define the ridge regression estimate:*

$$\hat{\beta}_k^{(RR)} = \underset{\beta \in L^2}{\arg\min} \left[E(\langle X, \beta\rangle - Y)^2 + k \left\|\beta\right\|^2\right],$$

where $\left\|\beta\right\|^2 = \int_a^b \beta^2(t)dt$.

There are many more estimates on this topic, but here we only covered the ones that we mention on the document. See chapters 3 and 5 of [Hastie et al., 2001] to find more alternatives for the linear regression model.

2.5 Reproducing Kernel Hilbert spaces

First, we state the definition of a Reproducing Kernel Hilbert Space (RKHS):

Definition 2.5.1 (RKHS)**.** *Let X be a topological space and $\mathcal{H}$ be a Hilbert space of functions from X to $\mathbb{R}$ or $\mathbb{C}$. $\mathcal{H}$ is called a Reproducing Kernel Hilbert Space if the evaluation functional $\delta_x(f) := f(x)$ in every $x \in X$ is continuous in $\mathcal{H}$, i.e. for any $x \in X$ we have that there exists $M > 0$ such that*

$$|\delta_x(f)| \leq M \left\|f\right\|_{\mathcal{H}}.$$

As a remark, this can be expressed as

$$\delta_x \in \mathcal{H}^* = \mathcal{B}(\mathcal{H}, \mathbb{C}),$$

where $\mathcal{H}^*$ is the dual space of $\mathcal{H}$, i.e. the bounded applications from $\mathcal{H}$ to $\mathbb{R}$.

Corollary 2.5.1.1 (Convergence in $\mathcal{H} \implies$ pointwise convergence)**.** *If $\mathcal{H}$ is RKHS and $f_n \xrightarrow{H} f$, then $f_n(x) \to f(x)$ for any $x \in X$.*

This corollary shows that RKHS are not trivial spaces and, indeed, if we have convergence in the RKHS, we have pointwise convergence. Next, we have to explore the concept of kernel that will help us to prove that a space is an RKHS.

Definition 2.5.2 (Kernel, Feature map and Feature space)**.** *Let X be a non-empty set. $k : X \times X \to \mathbb{C}$ is a kernel if there exists a Hilbert space $\mathcal{G}$ and a map $\phi : X \to \mathcal{G}$ so that*

$$k(x, x') = \langle \phi(x'), \phi(x) \rangle_{\mathcal{G}}.$$

ϕ is called feature map and $\mathcal{G}$ feature space.

It can be shown that kernel functions have two main properties:

Corollary 2.5.2.1 (Kernel properties)**.** *If k is a kernel, then the following properties hold:*

1. *(symmetry, hermitian) $k(x, y) = \overline{k(y, x)}$.*
2. *(positive definite) for any $n \in \mathbb{N}$, $\alpha_i \in \mathbb{C}$ and $x_i \in X$, $i = 1, \ldots, n$.*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$

Nevertheless, it is not simple to associate RKHSs to kernels. The concept that it is closely related to RKHS is reproducing kernel.

Definition 2.5.3 (Reproducing kernel)**.** *Let X be a non-empty set. $k : X \times X \to \mathbb{C}$ is a **reproducing kernel** of $\mathcal{H}$ (Hilbert space) if*

- *(Canonical feature map) $\phi(x) = k(\cdot, x) \in \mathcal{H}$.*
- *(Reproducing property) $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$ for all $x \in X$ and $f \in \mathcal{H}$.*

Corollary 2.5.3.1 (Reproducing kernels are kernels)**.** *Taking $\phi = k(\cdot, x) \in \mathcal{H}$ and $\mathcal{G} = \mathcal{H}$, k is a kernel.*

Then, we have that:

$$\text{Reproducing kernel} \implies \text{kernel} \implies \text{positive definite and symmetric.}$$

But how is this related to RKHS?

Proposition 2.5.4 (RKHS and Reproducing kernels)**.**

$$\mathcal{H} \text{ is a RKHS} \iff \mathcal{H} \text{ admits a reproducing kernel.}$$

Moreover, this correspondence is unique.

Proof. $\Longrightarrow$)

By Riesz Representation Theorem, $f(x) = \delta_x(f) = \langle f, K_x \rangle$. We have just to define $k(x, y) = \langle K_y, K_x \rangle$, and k is clearly a reproducing kernel.

$\Longleftarrow$) Just by the reproducing property and Cauchy-Schwarz inequality, we obtain

$$|f(x)| = |\langle f, k(\cdot, x)\rangle| \leq \|f\| \, \|k(\cdot, x)\| \leq M \|f\| .$$

□

To end this section, we conclude that indeed it is enough to show that k is positive definite and symmetric.

Theorem 2.5.5 (Moore–Aronszajn)**.** *Let $k : X \times X \to \mathbb{C}$ be positive-definite and hermitian. Then, k is a reproducing kernel and there exists $\mathcal{H}$, RKHS, whose kernel is k.*

This means that it is enough to analyze the properties of k to prove that there exists an underlying RKHS for which k is a reproducing kernel.

2.6 Other results

To end this chapter, we reserve this space to other results on topics that can not be easily classified into the previous categories.

Lemma 2.6.1. *Let A be a symmetric, positive definite matrix $n \times n$ matrix and $b \in \mathbb{R}^n$. The following problems are equivalent:*

1. $x = \arg\min_{x \in \mathbb{R}^n} \left(\frac{1}{2} x' A x - b' x\right)$.
2. *Find x such that $Ax = b$.*

Proof. Let $f(x) = \frac{1}{2} x' A x - b' x$. To find the minimum, we calculate the differential of f:

$$Df_x(v) = \frac{1}{2} v' A x + \frac{1}{2} x' A v - b' v.$$

Since the scalar product and A are symmetric, we have that

$$Df_x(v) = v' A x - v' b = v'(Ax - b).$$

If x is a minimum, $Df_x(v) = 0$ for any $v \in \mathbb{R}^n$, so $Ax - b = 0$ must be zero. Consequently, $Ax = b$ must hold.

□

3 PARTIAL LEAST SQUARES IN FINITE DIMENSION

3.1 Introduction

The goal of this chapter to clarify the different approaches that can be found in the state of the art of PLS estimation in the case of the finite dimensional linear regression model. Although some implementations will be referenced, the computational and numerical aspects of the algorithm are out of the scope of this chapter and this project. The objectives are:

1. Summarize the different methodologies employed to define PLS.
2. Study the properties that are derived from each definition.
3. Prove or disprove the equivalence among different definitions.

Prior to go deep into PLS, we formulate the linear regression model in the finite dimensional situation. Let $\mathcal{X}$ be a random vector of size d and $\mathcal{Y}$ be a random variable, for each observation of $\mathcal{X}$ and $\mathcal{Y}$, called X_i and Y_i respectively, the linear regression model is:

$$Y_i = X_i'\beta + \epsilon_i,$$

where β is a constant vector of size d and all the ϵ_i form a random vector ϵ of dimension n, the number of observations, which has multivariate normal distribution with zero mean and a covariance matrix $\Sigma = \sigma I_n$ (identity matrix of size n).

3.2 Definition 1: Minimization in Krylov Spaces

Krylov spaces play an outstanding role in the PLS regularization technique. Next section covers the definition of our minimization problem based on Krylov

spaces and also includes some useful properties. Note that the complete notion of Krylov space and the implications behind it are covered throughout the whole chapter.

3.2.1 Krylov spaces and definition of PLS

This subsection contains the first definition of PLS. This definition was chosen as the main one since other regularization techniques (e.g. Principal Components Regression (PCR), Ridge regression) can be also expressed in similar terms, that is, as a least squares minimization problem subject to appropriate restrictions. Moreover, using Krylov spaces to define PLS makes this technique clearer than other approaches that will be considered in subsequent subsections.

Before defining PLS, Krylov spaces must be properly defined.

Definition 3.2.1 (Krylov space of order q)**.** *Let A be a positive definite, symmetric $d \times d$ matrix and $b \in \mathbb{R}^d$, $b \neq 0$. The Krylov space of order q defined by A and b is:*

$$\mathscr{K}_q(A, b) = \mathscr{K}_q := span(b, Ab, A^2b, \ldots, A^{q-1}b). \tag{3.1}$$

Hereinafter, we will denote by X and Y the data matrices, which are $n \times d$ and $n \times 1$ respectively. This is, we are denoting by n the sample size and by d the dimension of the regressor.

PLS estimators are just least squares estimators constrained to Krylov spaces:

Definition 3.2.2 (PLS Version 1)**.** *The PLS estimator with q components is defined as*

$$\hat{\beta}_{PLS}^{(q)} := \underset{\beta \in \mathscr{K}_q}{\arg\min} \|X\beta - Y\|^2 = \underset{\beta \in \mathscr{K}_q}{\arg\min} \left(\sum_{i=1}^{n} (x_i'\beta - y_i)^2 \right), \tag{3.2}$$

where $\mathscr{K}_q = \mathscr{K}_q(A, b)$ with $A = X'X$ and $b = X'Y$.

Note that this problem can be solved explicitly and a closed formula of the estimator $\hat{\beta}_{\text{PLS}}^{(q)}$ of β can be obtained. Given an orthonormal basis of $\mathscr{K}_q(A, b)$, $\hat{\beta}_{\text{PLS}}^{(q)}$ can be computed in the same way we calculate the ordinary least squares linear estimator:

Proposition 3.2.3. *According to Definition 3.2.2, if $R_q'X'XR_q$ is invertible, then*

$$\hat{\beta}_{PLS}^{(q)} = R_q\hat{a}, \tag{3.3}$$

where $\hat{a} = (R_q'X'XR_q)^{-1}R_q'X'Y$ and R_q is a $d \times q$ matrix whose columns conform an orthonormal basis of $\mathscr{K}_q(A, b)$.

Proof. Consider (3.2) and express the squared norm as a scalar product:

$$\|X\beta - Y\|^2 = (X\beta - Y)'(X\beta - Y) = \beta'X'X\beta - 2Y'X\beta + Y'Y.$$

Any term that does not depend on β can be ignored. Moreover, multiplying by a positive constant does not change the solution of a minimization problem. Thus,

$$\hat{\beta}_{\text{PLS}}^{(q)} = \underset{\beta \in \mathscr{K}_q}{\arg\min} \left(\frac{1}{2} \beta' X' X \beta - Y' X \beta \right). \tag{3.4}$$

We consider R_q, an orthonormal basis of $\mathscr{K}_q$, and express any β in terms of that basis, i.e. $\beta = R_q \alpha$. Then, the minimization problem can be written without constraints as follows:

$$\hat{a} = \underset{\alpha \in \mathbb{R}^q}{\arg\min} \left(\frac{1}{2} \alpha' R_q' X' X R_q \alpha - Y' X R_q \alpha \right).$$

Using Lemma 2.6.1, this problem is equivalent to solving the equation

$$R_q' X' X R_q \hat{a} = R_q' X' Y,$$

and its solution is

$$\hat{a} = (R_q' X' X R_q)^{-1} R_q' X' Y.$$

□

Observation: *Is it always $R_q' X' X R_q$ invertible?* Only if $X'X$ is invertible, note that R_q is a orthonormal matrix so it is invertible.

3.2.2 The idea behind Krylov spaces

In this section we explain why considering Krylov spaces is suitable in the context of least squares estimation.

It is useful for the results that follow to consider the polynomial of a matrix, i.e. if $P(x) = \sum_{i=0}^{m} d_i x^i$, then $P(A) = \sum_{i=0}^{m} d_i A^i$ where A is a squared matrix and $A^0 = I$, the identity matrix of the same size.

The following result of linear algebra provides a representation of the inverse matrix in terms of the powers of the matrix, this is, as an element of a Krylov space:

Proposition 3.2.4. *Let A be a $d \times d$ symmetric square matrix with $det(A) \neq 0$. Then, there exists a polynomial P of degree $m-1$ that satisfies $A \cdot P(A) = I$, where $m \leq d$ is the number of distinct eigenvalues of A.*

Proof. Let $Q^*(\lambda) = \det(A - \lambda I)$ be the characteristic polynomial of A. Then $Q(A) = 0$ where $Q(x)$ is the minimal polynomial of Q^*, i.e. the monic polynomial of minimum degree that have the same roots of Q^*, which exists because A is symmetric and all eigenvalues are real. Also, $\det(A) \neq 0$ says that $Q(0) \neq 0$. If $Q(A) = c_0 I + c_1 A + \cdots + c_m A^m = 0$, then $c_0 I = -A(c_1 + c_2 A + \cdots +$

$c_m A^{m-1}$). If we define $P(A) = -\frac{1}{c_0}(c_1 + c_2 A + \cdots + c_d A^{m-1})$, then we have that $AP(A) = I$.

□

Put it in another way, this proposition means that $A^{-1} = P(A) = \sum_{i=1}^{m-1} \alpha_i A^i$. Recalling the expression for least squares estimator, $\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'Y$, this means that $\hat{\beta}_{\text{OLS}} = \sum_{i=1}^{m-1} \alpha_i A^i b$, with $A = X'X$ and $b = X'Y$. Therefore, $\hat{\beta}_{\text{OLS}} \in \mathscr{K}_m(A, b)$. This is particularly convenient when we have a small number of distinct eigenvalues of A. Thus, if the eigenvalues of A are concentrated around a small number of values, PLS provides a good approximation to OLS. Indeed, $\hat{\beta}^{(q)}_{\text{PLS}} \to \hat{\beta}_{\text{OLS}}$ as $q \to m \leq d$. Krylov spaces are not well documented in the literature of statistics. Indeed, this concept arose in numerical iterative methods that were employed to obtain the eigenvalues of linear operators or solving linear systems with large matrices avoiding product of matrices.

It is important to point out this relationship between numerical methods and PLS, so that the following sections explore different ideas based on numerical analysis techniques.

3.3 Definition 2: Partial Conjugate Gradient

In this section, conjugate direction methods are studied in detail. Although they arise as a way to accelerate methods for solving linear systems and to solve quadratic or nearly quadratic optimization problems, our goal here is to clarify the relation of these methods with PLS and Krylov spaces.

3.3.1 Introduction

Conjugate direction methods were proposed to deal with the following quadratic optimization problem:

$$\min_{x \in \mathbb{R}^n} \left(\frac{1}{2} x'Ax - b'x \right), \tag{3.5}$$

where A is an $n \times n$ symmetric positive definite matrix and $b \in \mathbb{R}^n$ a constant. Note that solving (3.5) is equivalent to find the solution of the linear system $Ax = b$. This result was already proved in Lemma 2.6.1. Next, we define the conjugate gradient algorithm that allows us to define a second version of PLS.

Proposition 3.3.1. *(Conjugate Gradient Algorithm)*

Let $x_0 \in \mathbb{R}^n$ be an initial approximation vector. Define $v_0 = -g_0 = b - Ax_0$. The following iterative method converge to the solution of problem (3.5) in at most n steps:

$$x_{k+1} = x_k + \alpha_k v_k, \tag{3.6}$$

where

$$g_k = Ax_k - b,$$

$$\alpha_k = -\frac{g_k' v_k}{v_k' A v_k},$$

$$v_{k+1} = -g_{k+1} + \gamma_k v_k,$$

and

$$\gamma_k = \frac{g_{k+1}' A v_k}{v_k' A v_k}.$$

Subsequent sections will cover the main properties of the conjugate gradient algorithm. Proof of the convergence will be covered at the end of the section.

3.3.2 Conjugate directions

The concept of conjugate vectors is similar to orthogonality but replacing the euclidean scalar product by the inner product defined by a matrix:

Definition 3.3.2. *(Conjugate with respect to A) Given a symmetric $d \times d$ symmetric matrix A, two elements $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^d$ are said to be conjugate with respect to A or A-orthogonal if $x'Ay = 0$.*

Nevertheless, for many interesting properties we require that A is positive definite. From now on, A will be a symmetric positive definite matrix.

Proposition 3.3.3. *Let A be a symmetric positive definite matrix and $\{v_i\}_{i=1}^d$ a set of conjugate directions with respect to A. If v_i is not zero for every $i \in \{1, ..., d\}$, then the vectors v_i are linearly independent.*

Proof. Let $v = \sum_{j=1}^n \alpha_j v_j = 0$. Then, $v_i' A v = v_i' A(\sum_{j=1}^n \alpha_j v_j) = 0$. Using that the vectors $\{v_j\}_{j=1}^d$ are conjugated, this yields $\alpha_i v_i' A v_i = 0$. Since A is positive definite, $v_i' A v_i > 0$. Therefore, $\alpha_i = 0$ for every $i = 1, \ldots, d$.

□

As a consequence, if d is the dimension of the space, then $\{v_1, \ldots, v_d\}$ form a basis of the whole space.

Theorem 3.3.4. *(Conjugate Direction Theorem) Let $\{v_i\}_{i=0}^{n-1}$ be a sequence of nonzero conjugated (with respect to A) vectors of $\mathbb{R}^n$ and let $x_0 \in \mathbb{R}^n$ be an initial approximation vector. Then the following iterative method converge to the solution*

of problem (3.5) in n steps, that is, $x_n = x^*$, where x^* is the solution of (3.5). For $k = 1, 2 \dots, n-1$,

$$x_{k+1} = x_k + \alpha_k v_k, \tag{3.7}$$

where

$$g_k = Ax_k - b,$$

and

$$\alpha_k = -\frac{g_k' v_k}{v_k' A v_k}.$$

Proof. Using proposition 3.3.3, the $\{v_i\}_{i=0}^{n-1}$ are linearly independent so that they form a basis of $\mathbb{R}^n$. Thus, $x^* - x_0$ can be expressed in that basis: $x^* - x_0 = \sum_{i=0}^{n-1} \gamma_i v_i$. Note that the coefficients in this expression are

$$\gamma_k = \frac{v_k' A(x^* - x_0)}{v_k' A v_k} = \frac{v_k' b - v_k' A x_0}{v_k' A v_k}$$

. Indeed,

$$\begin{aligned}
x^* - x_0 &= \sum_{i=0}^{n-1} \gamma_i v_i, \\
A(x^* - x_0) &= \sum_{i=0}^{n-1} \gamma_i A v_i, && \text{(Apply } A \text{ and linearity)} \\
v_k' A(x^* - x_0) &= \sum_{i=0}^{n-1} \gamma_i v_k' A v_i && \text{(Multiply by } v_k\text{)} \\
&= \gamma_k v_k' A v_k. && (A - \text{orthogonality})
\end{aligned}$$

Also, applying the definition of x_k recursively, we obtain

$$x_k = \alpha_{k-1} v_{k-1} + x_{k-1} = \cdots = \sum_{i=0}^{k-1} \alpha_i v_i + x_0,$$

and following the same reasoning:

$$\begin{aligned}
x_k - x_0 &= \sum_{i=0}^{k-1} \alpha_i v_i, \\
A(x_k - x_0) &= \sum_{i=0}^{k-1} \alpha_i A v_i, && \text{(Apply } A \text{ and linearity)} \\
v_k' A(x_k - x_0) &= \sum_{i=0}^{k-1} \alpha_i v_k' A v_i && \text{(Multiply by } v_k\text{)} \\
&= 0. && (A - \text{orthogonality})
\end{aligned}$$

then, $v_k' A x_0 = v_k' A x_k$. Substituting into the expressión of γ_k, we get

$$\gamma_k = \frac{v_k' b - v_k' A x_0}{v_k' A v_k} = \frac{-v_k'(A x_k - b)}{v_k' A v_k} = \frac{-g_k' v_k}{v_k' A v_k}.$$

In other words, $\gamma_i = \alpha_i$ for $i = 1, \ldots, k-1$, which implies that $x_n = x_0 + \sum_{i=0}^{n-1} \alpha_i v_i = x^*$.

□

As a corollary, the gradients are g_k are orthogonal to the space span$\{v_0, \ldots, v_{k-1}\}$.

Corollary 3.3.4.1. *Under the hypothesis of theorem 3.3.4, we have that for $k = 1, \ldots, d$ and $j = 0, \ldots, k-1$, $v_j' g_k = 0$.*

Proof. For $k = 1, \ldots, d$,

$$\begin{aligned} g_{k+1} &= g_k + \alpha_k A v_k \\ &= g_k - \frac{g_k' v_k}{v_k' A v_k} A v_k. \end{aligned}$$

Then,

$$v_k' g_{k+1} = v_k' g_k - \frac{v_k' g_k}{v_k' A v_k} v_k' A v_k = 0. \tag{3.8}$$

Reasoning by induction, we have already proved it for $k = 1$, i.e. $g_1' v_0 = 0$. Assuming it is true the property for g_k, we write g_{k+1}

$$v_j' g_{k+1} = v_j' g_k - \frac{v_k' g_k}{v_k' A v_k} v_j' A v_k.$$

For $j = 1, \ldots, k-1$, the first term cancels due to the induction hypothesis and the second one because the vectors are conjugate. For $j = k$, we checked that it is zero in (3.8). This is, for $j = 0, \ldots, k$,

$$v_j' g_{k+1} = 0.$$

□

3.3.3 Conjugate Gradient Algorithm

Any conjugate direction method converges to the solution x^* of the linear equation $Ax = b$ and also solves the minimization problem (3.5). It can also

be proved a richer result, ensuring that x_k, $k = 1, 2, \dots d$, is the solution to the problem $\min_{x \in x_0 + \mathbb{B}_k} f(x)$ where $\mathbb{B}_k = \text{span}\{v_0, v_1, ..., v_{k-1}\}$ and $f(x) = \frac{1}{2}x'Ax - b'x$.

Corollary 3.3.4.2. *(Expanding Subspace) With the same notation and assumptions of Theorem 3.3.4, x_k is the solution to the problem:*

$$\min_{x \in \mathbb{B}_k} \left(\frac{1}{2} x'Ax - b'x \right),$$

where $\mathbb{B}_k = span(v_0, v_1, ..., v_{k-1})$.

Proof. Defining $f(x) = \frac{1}{2}x'Ax - b'x$, it is clear that f is a strictly convex function if A is positive definite, so we just need to show that the gradient of f at x_k is orthogonal to the space $\mathbb{B}_{k-1}$. Since $\nabla f(x_k) = Ax_k - b$, it is enough to show that $Ax_k - b \perp \mathbb{B}_{k-1}$.

Furthermore, we have proved in Corollary 3.3.4.1 that $Ax_k - b$ is orthogonal to $v_0, ..., v_{k-1}$. Then $Ax_k - b \perp \mathbb{B}_{k-1}$.

□

With all these results, the second definition of PLS can be stated.

Definition 3.3.5 (PLS Version 2). *Let $x_0 = 0$, $A = X'X$ and $b = X'Y$. We define $\hat{\beta}^{(q)}_{PLS} = x_q$, where x_q is the value obtained after q steps of the conjugate gradient algorithm.*

Since we want to show the equivalence between definition 3.2.2 and definition 3.3.5, it is required that $x_q \in \mathscr{K}_q(A, b)$, the Krylov space of dimension q. The following proposition verifies this and other properties about the conjugate gradient algorithm:

Proposition 3.3.6. *Let x^* be the true solution to the optimization problem (3.5). If $x_k \neq x^*$, the iterations of the algorithm described in Theorem 3.3.1 satisfy the following properties:*

(a) $span\{g_0, \dots, g_k\} = span\{v_0, \dots, v_k\} = \mathscr{K}_{k+1}(A, g_0) = span\{g_0, Ag_0..., A^k g_0\}$.

(b) v_k is conjugate to any v_i with $i < k$.

Proof. (a) First, we prove $\text{span}\{g_0, \dots, g_k\} = \mathscr{K}_{k+1}(A, g_0)$ By induction, it is trivial for $k = 0$, since $v_0 = -g_0 \in \text{span}\{g_0\}$. Suppose it is true for k, then for $k + 1$ we have that:

$$g_{k+1} = Ax_{k+1} - b = \underbrace{Ax_k - b}_{g_k} + \alpha_k Av_k = g_k + \alpha_k Av_k.$$

By the hypothesis, $g_k \in \mathscr{K}_{k+1}(A, g_0) \subset \mathscr{K}_{k+2}(A, g_0)$, so the only thing to prove is that $Av_k \in \mathscr{K}_{k+2}(A, g_0)$. Since $v_k \in \mathscr{K}_{k+1}(A, g_0)$, this means that $v_k = \sum_{i=0}^{k} c_i A^i g_0$. Consequently, $Av_k = \sum_{i=0}^{k} c_i A^{i+1} g_0$ and thus $Av_k \in \mathscr{K}_{k+2}(A, g_0)$. Moreover, g_{k+1} is a new vector that expands the Krylov space, i.e. $g_{k+1} \notin \mathscr{K}_{k+1}(A, g_0)$. This a consequence of the Corollary 3.3.4.1, $g_{k+1} \perp \text{span}\{v_0, \ldots, v_k\} = \mathscr{K}_{k+1}(A, g_0)$.

Then, we have obtained that $\text{span}\{g_0, \ldots, g_{k+1}\} = \mathscr{K}_{k+2}(A, g_0)$.

Also, $\text{span}\{v_0, \ldots, v_{k+1}\} = \mathscr{K}_{k+2}(A, g_0)$ since $v_{k+1} = -g_{k+1} + \gamma_k v_k$ where we have already checked that $g_{k+1} \in \mathscr{K}_{k+2}(A, g_0)$ and $v_k \in \mathscr{K}_{k+1}(A, g_0) \subset \mathscr{K}_{k+2}(A, g_0)$. Also, since v_{k+1} is conjugate to the previous ones, it is also independent from any of them, therefore it expands the Krylov Space and again: $\text{span}\{v_0, \ldots, v_{k+1}\} = \mathscr{K}_{k+2}(A, g_0)$.

(b) The last thing to check is that v_{k+1} is conjugate to any v_i with $i < k+1$. For $i = k$, the scalar product is zero due to the definition of $\gamma_k = \frac{g'_{k+1} A v_k}{v'_k A v_k}$.

$$\begin{aligned} v'_{k+1} A v_k &= (-g_{k+1} + \gamma_k v_k)' A v_k \\ &= \frac{g'_{k+1} A v_k}{v'_k A v_k} v'_k A v_k - g'_{k+1} A v_k \\ &= 0. \end{aligned}$$

If $i < k$,

$$\begin{aligned} v'_{k+1} A v_i &= (-g_{k+1} + \gamma_k v_k)' A v_i \\ &= \frac{g'_{k+1} A v_k}{v'_k A v_k} v'_k A v_i - g'_{k+1} A v_i \\ &= 0. \end{aligned}$$

The first term vanish due to the induction hypothesis. The second term vanish because g_{k+1} is orthogonal to $\text{span}\{v_0, \ldots, v_{i+1}\}$ due to corollary 3.3.4.1 and $Av_i \in \text{span}\{v_0, \ldots, v_{i+1}\}$. □

Finally, we are able to state the main result that shows the equivalence between the two definitions of PLS.

Theorem 3.3.7 (Equivalence of PLS version 1 and PLS version 2)**.** *Definition 3.2.2 and definition 3.3.5 are equivalent.*

Proof. In proposition 3.3.6, we proved that the directions $\{v_i\}_{i=1}^{q}$ obtained in the conjugate gradient are conjugate vectors of $\mathscr{K}_q(A, b)$. Then, corollary 3.3.4.2 implies that x_q, iteration of the conjugate gradient algorithm solves

$$x_k = \underset{x \in \mathscr{K}_q(A,b)}{\arg\min} \left(\frac{1}{2} x'Ax - b'x\right) = \underset{\beta \in \mathscr{K}_q(A,b)}{\arg\min} \left(\frac{1}{2} \beta' X'X\beta - Y'X\beta\right).$$

Equation (3.4) shows a representation of $\hat{\beta}^{(q)}_{\text{PLS}}$ that is the same as the one above, which proves that both methods solve the same minimization problem.

□

Once we proved that the two previous definition are the same, we end the section with a result that will be used in the sequel:

Proposition 3.3.8. *Let $u \in \mathscr{K}_q(A, b)$. Then, there exists a polynomial of degree at most q, $P(\cdot)$, such that $P(0) = 1$ and the following equality holds:*

$$x^* - u = P(A)x^*,$$

where x^* fulfills $Ax^* = b$.

Proof. Let $P(t) = \sum_{i=0}^{q} \lambda_i t^i$. It is quite trivial to find the coefficients λ_i by just imposing the conditions over P. First, $P(0) = 1$ implies that $\lambda_0 = 1$, i.e. $P(A)x^* = x^* + \sum_{i=1}^{q} \lambda_i A^q x^*$. Since $Ax^* = b$, $P(A)x^* = x^* + \sum_{i=1}^{q} \lambda_i A^{q-1} b$, and $x^* - P(A)x^* \in \mathscr{K}_q(A, b)$. Recalling that $u \in \mathscr{K}_q(A, b)$, i.e. $u = \sum_{i=0}^{q-1} \alpha_i A^i b$, we just have to choose $\lambda_i = \alpha_{i-1}$, for $i = 1, \ldots, q$. □

3.4 Definition 3: Filter factors

3.4.1 Singular Value Decomposition

First, we recall that the Singular Value Decomposition (SVD) of a rectangular matrix X is:

$$X = U\Sigma V' \tag{3.9}$$

where

1. X is the $n \times d$ data matrix,
2. U is an orthogonal $n \times d$ matrix,
3. Σ is a non-negative $d \times d$ diagonal matrix. The elements of the diagonal are ordered ($\sigma_1 \geq \ldots \geq \sigma_d$),
4. V is an orthogonal $d \times d$ matrix.

Some useful properties makes this decomposition really convenient for the regression analysis [Mandel, 1982]:

Proposition 3.4.1. *Let X be a $n \times p$ real matrix. The following statements are satisfied:*

1. *The singular value decomposition of X always exists.*
2. *U is a matrix whose columns are the eigenvectors of XX' with eigenvalues not equal to zero.*
3. *V is a matrix whose columns are the eigenvectors of $X'X$ with eigenvalues not equal to zero.*
4. *The elements of the diagonal of Σ are the square root of the eigenvalues of $X'X$.*
5. *(Moore-Penrose pseudoinverse) $\hat{\beta}_{OLS} = X^{\dagger}Y$, where $X^{\dagger}$ is known as the Moore-Penrose pseudoinverse and it is just an approximation to the inverse in terms of the SVD, i.e. $X^{\dagger} = V\Sigma^{-1}U'$, where Σ^{-1} denotes to the matrix with non-null elements inverted.*

Further information about this result and many more that might be useful can be found in [Golub and Reinsch, 1970].

3.4.2 Relationship with regression analysis

In particular, from property 5, an interesting formula can be given:

$$\hat{\beta}_{\text{OLS}} = V\Sigma^{-1}U'Y = \sum_{i=1}^{d} \frac{u_i'Y}{\sigma_i} v_i, \tag{3.10}$$

where u_i and v_i are the column vector of U and V respectively and σ_i is the ith element of the diagonal of Σ. Note that the singular values σ_i are connected to the v_i in a way that is really similar to PCA. Observing the equations, it is simple to see that PCR (definition 2.4.1) has the following formula:

$$\hat{\beta}_{\text{PCR}}^{(m)} = \sum_{i=1}^{m} \frac{u_i'Y}{\sigma_i} v_i = \sum_{i=1}^{d} \mathbb{1}_{[0,m]}(i) \frac{u_i'Y}{\sigma_i} v_i.$$

Less obvious but also true is that Ridge Regression (definition 2.4.2) has also a formula in terms of the SVD:

$$\hat{\beta}_{\text{RR}}^{(k)} = \sum_{i=1}^{d} \frac{\sigma_i^2}{\sigma_i^2 + k} \frac{u_i'Y}{\sigma_i} v_i.$$

Since the proof for Ridge Regression is completely unrelated to any other incoming results, the proof is elided.

Definition 3.4.2 (Filter factors). *If we express an arbitrary regression estimator as*

$$\hat{\beta} = \sum_{i=1}^{p} w_i \frac{u_i' Y}{\sigma_i} v_i, \tag{3.11}$$

the weights w_i are called filter factors. Equivalently,

$$\hat{\beta} = V D^{-1} F w,$$

where F is a $d \times d$ matrix whose diagonal is the vector $U'Y$, D is a $d \times d$ matrix whose diagonal is the vector $(\sigma_1, \ldots, \sigma_d)$ and V is the orthogonal basis of the SVD.

With the aforementioned expressions in mind, it is simple to see that the filter factors of $\hat{\beta}_{\text{OLS}}$, $\hat{\beta}^{(m)}_{\text{PCR}}$ and $\hat{\beta}^{(k)}_{\text{RR}}$ are

$$w_i^{(\text{OLS})} = 1,$$
$$w_i^{(\text{PCR(m)})} = \mathbb{1}_{[0,m]}(i),$$

and

$$w_i^{(\text{RR(k)})} = \frac{\sigma_i^2}{\sigma_i^2 + k}.$$

The next question is: can $\hat{\beta}^{(q)}_{\text{PLS}}$ be expressed in the same way? Then, which are the filter factors of $\hat{\beta}^{(q)}_{\text{PLS}}$? To answer this question, we must pay attention to Krylov spaces. They are connected to SVD. Let $\mathscr{K}_M$ the maximal Krylov space, i.e. $\mathscr{K}_{M-1} \subsetneq \mathscr{K}_M = \mathscr{K}_{M+1}$, then the following result provides the dimension of the maximal Krylov space:

Lemma 3.4.3. *Let $\mathscr{K}_M$ be the maximal Krylov space, Λ be the set of all eigenvalues of $X'X$ and P_λ be the orthogonal projector onto the eigenspace generated by the eigenvector whose corresponding eigenvalue is λ. Then:*

- *$\{P_\lambda X'Y\}_{\lambda \in \Lambda}$ is a basis of $\mathscr{K}_M$.*
- *$P_\lambda X'Y = \sum_{i=1}^{p} \mathbb{1}_{\{\lambda_i = \lambda\}} \sigma_i (u_i' Y) v_i$.*

Therefore, M is the number of eigenvalues λ for which $u_i' Y \neq 0$.

The proof of this lemma can be found at [Parlett, 1998]. It leads itself to important consequences, such as knowing the number of iterations of the Conjugate Gradient to reach $\hat{\beta}_{\text{OLS}}$.

The key question of this section is: Can $\hat{\beta}^{(q)}_{\text{PLS}}$ be expressed in (3.11) for some suitable filter factors? By applying the previous lemma, we see that the answer to this question is affirmative.

First, a simple form that proof its existence is given, although a more convenient result drive us to many more properties:

Theorem 3.4.4. *Let F be the $d \times d$ matrix whose diagonal is the vector $U'Y$ and* $\mathbf{w}$ *the column vector of the filter factors. Then,*

$$\hat{\beta}_{PLS}^{(q)} = VD^{-1}F\mathbf{w},$$

given by the solution to the system where $\mathbf{w}$ *is*

$$F\mathbf{w} = DV'\hat{\beta}_{PLS}^{(q)} = U'\mathrm{P}Y,$$

where $\mathrm{P} = XR_q(R_q'X'XR_q)^{-1}(XR_q)'$ *and R_q is a basis of $\mathscr{K}_q$*

Proof. Assuming $\hat{\beta}_{\mathrm{PLS}}^{(q)} = VD^{-1}F\mathbf{w}$, we have

$$F\mathbf{w} = DV'\hat{\beta}_{\mathrm{PLS}}^{(q)}.$$

Using definition 1 of PLS, i.e. equation (3.2) and also $U'U = I$:

$$DV'\hat{\beta}_{\mathrm{PLS}}^{(q)} = U'(UDV'R_q(R_q'X'XR_q)^{-1}R_q'X'Y).$$

Using the SVD and the definition of P, we finally obtain:

$$F\mathbf{w} = U'(\underbrace{XR_q(R_q'X'XR_q)^{-1}(XR_q)'}_{\mathrm{P}}Y) = U'\mathrm{P}Y.$$

□

As it was anticipated, another characterization of w_i can be given. This characterization follows a numerical approach to solve the problem. First, the Ritz values are defined:

Definition 3.4.5 (Ritz values)**.** *For any orthonormal basis of $\mathscr{K}_q(A, b)$ given as a matrix R_q of column vectors, the eigenvalues*

$$\theta_1 \geq \theta_2 \ldots \geq \theta_m$$

of $R_q'AR_q$ are called Ritz values. The corresponding eigenvectors are called Ritz vectors.

If θ were an eigenvalue with eigenvector v, we would have that $Av - \theta v = 0$. Being a Ritz value is weaker.

The connection between PLS and these Ritz values is behind the conjugate gradient method. The following lemma associates the Ritz values to conjugate gradient.

Lemma 3.4.6. *Let $x_0 = 0$ and let x_k be the k iteration of conjugate gradient. Assuming the common notation $A = X'X$, $b = X'Y$ and x^* the solution of $Ax = b$. The following equality holds:*

$$x^* - x_q = \mathscr{R}_q(A)x^*, \tag{3.12}$$

where $\mathscr{R}_q(A) = \prod_{i=1}^{q} \frac{\theta_i^{(q)} I - A}{\theta_i^{(q)}}$ and $\{\theta_i^{(q)}\}_{i=1}^{q}$ are the Ritz values corresponding to $\mathscr{K}_q(A, b)$.

Proof. Step 1: Prove that the Ritz vectors and b are not orthogonal

Let $\theta_i^{(q)}$ be a Ritz value and let v_i be the corresponding Ritz vector. Also, consider $B_q = (v_1, \dots, v_q)$ a matrix whose columns form an orthonormal basis of $\mathscr{K}_q(A, b)$. Note that for $k < q$ we have that

$$\begin{aligned} v_i'(B_q'AB_q)^k b &= (B_q'AB_q)'v_i'(B_q'AB_q)^{k-1}b && (B_q'AB_q \text{ is self-adjoint}) \\ &= \theta_i^{(q)} v_i'(B_q'AB_q)^{k-1}b && (v_i \text{ is a Ritz vector}) \\ &= \dots && (k \text{ iterations}) \\ &= (\theta_i^{(q)})^k v_i'b. \end{aligned}$$

This implies that if $v_i \perp b$, then $v_i \perp (B_q'AB_q)^k b$ for any $k < q$. Note that $(B_q'AB_q)^k = B_q'A^kB_q = A_k$ because $A_k \in \mathscr{K}_q(A, b)$. Hence, $v_i \perp \mathscr{K}_q(A, b)$. But, this is itself a contradiction. If $v_i \perp \mathscr{K}_q(A, b)$, this means that v_i is 0 since $v_i \in \mathscr{K}_q(A, b)$. This is a contradiction and thus we have $v_i'b \neq 0$.

Step 2: Prove that a polynomial exists

In proposition 3.3.8, we proved that if $x_q \in \mathscr{K}_q(A, b)$, then there exists p polynomial of degree q such that $x^* - x_q = p(A)x^*$.

This simplifies a lot our effort, since we have already proved the existence, we only need to prove that $R_q = p$.

Step 3: Check that polynomials p and R_q have the same roots

It is clear that $x^* - x_k$ is A-orthogonal to all the vectors in $\mathscr{K}_q(A, b)$, implying that $p(A)x^*$ is also A-orthogonal. If $p(A)x^*$ is A-orthogonal, $p(A)b$ is orthogonal to $\mathscr{K}_q(A, b)$, i.e. for any $v \in \mathscr{K}_q(A, b)$:

$$v'Ap(A)x^* = v'p(A)Ax^* = v'p(A)b = 0.$$

It holds that $B_q'p(A)bB_q = 0$ because the columns of B_q belong in $\mathscr{K}_q(A, b)$. Doing some computations, we have:

$$\begin{aligned} 0 =& B_q'p(A)bB_q \\ =& B_q'p(A)B_qB_q'bB_q && (B_qB_q' = I_d, \text{ identity of size d}) \\ =& B_q'p(A)B_qb && (b \in \mathscr{K}_q(A, b)) \\ =& p(B_q'AB_q)b && (B_q'A^iB_q = (B_q'AB_q)^i \text{ and linearity}) \end{aligned}$$

Consequently, if we consider the diagonalization of $B_q'AB_q = UDU^{-1}$, where $D = \text{diags}(\theta_1^{(q)}, \ldots, \theta_q^{(q)})$ and U a matrix whose columns are the eigenvectors of $B_q'AB_q$, i.e. the Ritz vectors. With this, we can conclude that:

$$\begin{aligned} 0 =& p(B_q'AB_q)b \\ =& p(UDU^{-1})b \qquad\qquad \text{(Diagonalization)} \\ =& Up(D)U^{-1}b, \end{aligned}$$

i.e. either $v'b = 0$ or $p(D) = 0$. Since we already prove that $v'b \neq 0$, the $p(D) = 0$, that is, $p(\theta_i^{(q)}) = 0$ for $i = 1, \ldots, q$.

□

This means that the conjugate gradient iteration can be expressed as

$$x_q = (1 - \mathscr{R}_q(A))x^*.$$

Thanks to theorem 3.3.7 that exposes the equivalence between PLS definition and Conjugate Gradient, this can be expressed in terms of the notation of the problem we want to solve, i.e.

$$\hat{\beta}_{\text{PLS}}^{(q)} = (I - \mathscr{R}_q(X'X))\hat{\beta}_{\text{OLS}} \tag{3.13}$$

Notice that $\mathscr{R}_q(\theta_i^{(q)}) = 0$ for any Ritz value $\theta_i^{(q)}$ and $R_q(0) = 1$. With all these ingredients, it is possible to proof the following result

Definition 3.4.7 (PLS Version 3). *Let $X = U\Sigma V'$ be the SVD of the data and let $\{\theta_i^{(q)}\}_{i=1}^q$ the Ritz values associated to $\mathscr{K}_q(X'X, X'Y)$. Also, let $\lambda_i = \sigma_i^2$ be the eigenvalues of $X'X$. Then, the estimator of PLS as a filter factors expression is:*

$$\hat{\beta}_{PLS}^{(q)} = \sum_{i=1}^{p} w_i \frac{u_i'Y}{\sigma_i} v_i, \tag{3.14}$$

with $w_i = 1 - \mathscr{R}_q(\lambda_i)$.

Theorem 3.4.8. *Definition 3.2.2 and definition 3.4.7 are equivalent.*

Proof. Starting from (3.13), we couple in $\hat{\beta}_{\text{OLS}}$

$$\hat{\beta}_{\text{PLS}}^{(q)} = (I - \mathscr{R}_q(X'X))\hat{\beta}_{\text{OLS}} = (I - \mathscr{R}_q(X'X))V\Sigma^{-1}U'Y.$$

Now, we want to express $\mathscr{R}_q(X'X)$ in terms of the SVD. Easily, $X'X = V\Sigma^2V'$, where Σ^2 is a $p \times p$ diagonal matrix whose entries are the eigenvalues λ_i of $X'X$ in decreasing order. Writting $\mathscr{R}_q$, we are able to see that (3.14) arises.

$$\mathscr{R}_q(V\Sigma^2V') = \prod_{i=1}^{q} \frac{\theta_i^{(q)} I - V\Sigma^2V'}{\theta_i^{(q)}} = \prod_{i=1}^{q} \frac{\theta_i^{(q)} VV' - V\Sigma^2V'}{\theta_i^{(q)}} = V \prod_{i=1}^{q} \frac{\theta_i^{(q)} V' - \Sigma^2V'}{\theta_i^{(q)}}$$

$$R_q(V\Sigma^2V') = V \prod_{i=1}^{q} \frac{\theta_i^{(q)} V' - \Sigma^2V'}{\theta_i^{(q)}} = V \underbrace{\left(\prod_{i=1}^{q} \frac{\theta_i^{(q)} I - \Sigma^2}{\theta_i^{(q)}} \right)}_{\Sigma^*} V'.$$

We notice that Σ^* is a diagonal matrix whose elements are $\sigma_i^* = \frac{\theta_i^{(q)} - \lambda_i}{\theta_i^{(q)}} = \mathscr{R}_q(\lambda_i)$, i.e. $\Sigma^* = \mathscr{R}_q(\Sigma^2)$. Coupling this expression in the first formula of $\hat{\beta}_{\text{OLS}}$ and using V is orthogonal ($VV' = V'V = I$), we arrive at

$$\hat{\beta}_{\text{PLS}}^{(q)} = (VV' - V\Sigma^*V')V\Sigma^{-1}U'Y = V(I - \Sigma^*)V'V\Sigma^{-1}U'Y = V(I - \Sigma^*)\Sigma^{-1}U'Y.$$

Once again, $(I - \Sigma^*)\Sigma^{-1}$ is a diagonal matrix whose elements are $\frac{1-\mathscr{R}_q(\lambda_i)}{\sigma_i}$. This means then that the filter factors are $1 - \mathscr{R}_q(\lambda_i)$ and thus the PLS estimator can be written as follows:

$$\hat{\beta}_{\text{PLS}}^{(q)} = V(I - \Sigma^*)\Sigma^{-1}U'Y = \sum_{i=1}^{p} w_i \frac{u_i'Y}{\sigma_i} v_i = \sum_{i=1}^{p} (1 - \mathscr{R}_q(\lambda_i)) \frac{u_i'Y}{\sigma_i} v_i.$$

□

3.4.3 Connection to Lanczos and Arnoldi methods

The main reason to study the Lanczos method in this context is the possibility of provide convergence results of $\theta_i^{(q)}$, Ritz value, to λ_i, real eigenvalue of A.

Originally proposed in [Lanczos, 1950], Lanczos method is an adaption of the power methods to find the m biggest eigenvalues and eigenvectors of hermitian matrices. On the other hand, Arnoldi method works similarly but with all kind of matrices. Both algorithms compute the approximation to the eigenvectors (i.e. the Ritz vectors) by decomposing the matrix and using the QR decomposition to find them.

Definition 3.4.9 (Lanczos method)**.** *Let A be a hermitian matrix of size $d \times d$. Let $v_1 \in \mathbb{R}^d$ with norm 1, then the following iterative method is defined for $j = 1, \ldots, q \leq m$:*

- *If* $j = 1$*:*
  - $b_1 = Av_1$,
  - $\alpha_1 = b_1' v_1$,
  - $w_1 = b_1 - \alpha_1 v_1$.
- *For* $j = 2, \ldots, m$*:*
  - $\gamma_j = \|w_{j-1}\|$.
  - $v_j = \begin{cases} w_j/\gamma_j & \text{if } \gamma_j \neq 0 \\ \text{any orthonormal vector to } v_1, \cdots, v_{j-1} & \text{otherwise} \end{cases}$
  - $b_j = Av_j$,
  - $\alpha_j = b_j' v_j$,
  - $w_j = b_j - \alpha_j v_j$.

The result is a matrix V *whose columns are the vectors* $\{v_j\}_{j=1}^m$ *and a tridiagonal matrix*

$$T = \begin{pmatrix} \alpha_1 & \gamma_2 & & & & 0 \\ \gamma_2 & \alpha_2 & \gamma_3 & & & \\ & \gamma_3 & \alpha_3 & \ddots & & \\ & & \ddots & \ddots & \gamma_{m-1} & \\ & & & \gamma_{m-1} & \alpha_{m-1} & \gamma_m \\ 0 & & & & \gamma_m & \alpha_m \end{pmatrix}.$$

The eigenvalues of T *are the Ritz values of order* m*, i.e.* $\{\theta_j^{(m)}\}_{j=1}^m$

Note first that if $\gamma_j = 0$ for some j, it means that $Av_j - \alpha_j v_j = 0$, i.e. v_j is an eigenvector and α_j is an eigenvalue. Kaniel–Paige convergence theory states how the eigenvalues of T are related to the eigenvalues of A. The following theorem presents a results related to eigenvalues and Ritz values.

Theorem 3.4.10. *Let* A *be a* $d \times d$ *symmetric matrix and* T *be the tridiagonalization obtained by Lanczos algorithm after* q *steps. Let* λ_1 *be the biggest eigenvalue of* A *and* $\theta_1^{(q)}$ *be the biggest eigenvalue of* T*. Then,*

$$\lambda_1 - \theta_1^{(q)} \leq 4 \frac{1 - |d_1|^2}{|d_1|^2} (\lambda_1 - \lambda_n) R^{-2(q-1)},$$

where $\{d_j\}_{j=1}^d$ *are the coefficients of* v_1 *in the eigenvector basis, i.e.* $v_1 = \sum_{j=1}^d z_j d_j$*,* z_j *eigenvector of* A*; and* $R = \exp(\operatorname{arccosh}(1 + 2\rho))$ *with* $\rho = \frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n}$.

This fact shows that Lanczos method, Ritz values and PLS are closely related. An improved result that deals not only with the largest Ritz value can be found in theorem 12.4.1 of [Parlett, 1998]. Furthermore, Kaniel-Paige convergence theory provides results of convergence that might be useful to investigate the structure

of $R_q(x) = \prod_{i=1}^{q} \frac{\theta_i^{(q)} - x}{\theta_i^{(q)}}$. Nevertheless, this already yields out of the scope of this project and only some few results [Phatak and de Hoog, 2002] can be found in the state of the art.

3.5 Definition 4: statistical criterion

The preceding definitions are based on optimization and linear algebra concepts but they lack a clear statistical interpretation, despite the connections we have established between PLS and PCR and Ridge/Lasso in the previous section. Consequently, this section aims at giving a statistical meaning for PLS.

No matter which flavour of PLS is consulted, all versions aggrees on using conjugate directions that maximize the covariance. There are two particular versions strongly used in the state of the art: NIPALS [Wold, 1966] and SIMPLS [de Jong, 1993]. Indeed, for the vast majority of modifications of PLS, if Y is scalar, they are equivalent. For the sake of simplicity, we will employ NIPALS version:

Definition 3.5.1 (PLS Version 4)**.** *Let X be the centered $n \times d$ data matrix and Y the response vector.*

1. *Initialization:* $X_0 = X$
2. *Iteration: for* $i = 1, \dots, q$*:*
   a) $w_i = \frac{X_{i-1}'Y}{\left\|X_{i-1}'Y\right\|}$,
   b) $t_i = \frac{X_{i-1}'w_i}{\left\|X_{i-1}'w_i\right\|}$,
   c) $p_i = X_{i-1}'t_i$,
   d) $X_i = X_{i-1} - t_i p_i'$,
3. *Result: the sequences* $\{w_i\}_{i=1}^q$, $\{t_i\}_{i=1}^q$ *and* $\{p_i\}_{i=1}^q$

Given the following matrices $W_q = (w_1 \dots w_q)$, $T_q = (t_1 \dots t_q)$ *and* $P_q = (p_1 \dots p_q)$*, we have that*

$$\hat{\beta}_{PLS}^{(q)} = W_q(P_q'W_q)^{-1}T_qY. \tag{3.15}$$

Next, we want to state a result relating this definition to Krylov spaces.

Proposition 3.5.2. *The vectors in NIPALS algorithm satisfy the following properties:*

1. $\{w_i\}_{i=1}^q$ *is a orthonormal sequence.*
2. *span*$(\{w_i\}_{i=1}^q) = \mathscr{K}_q(X'X, X'Y)$.

3. $\{t_i\}_{i=1}^q$ *is a orthonormal sequence.*

4. $span(\{t_i\}_{i=1}^q) = \mathscr{K}_q(XX', XX'Y)$.

The proof of this proposition can be found in [Eldén, 2004] (proposition 3.1). As we see, this proposition together with (3.15) provides an easy way of seeing that $\hat{\beta}_{\text{PLS}}^{(q)} \in \mathscr{K}_q(X'X, X'Y)$. The only thing left to prove is that this minimizes a least squares expression. Indeed, this is immediate consequence of the algorithm, since each iteration applies regression to the residuals.

Theorem 3.5.3. *Definition 3.2.2 and Definition 3.5.1 are equivalent.*

This is a classic result that appears in many reviews and books on this topic, such as [Tenenhaus, 1998] and [Eldén, 2004] (proposition 3.1 and 3.2). Also, another proof that relies on using Definition 3.4.7 is available at [de Jong, 1995].

The relevant question in this part is: which is the difference between different PLS algorithms? Most of the times, the difference relies on the basis of the Krylov space that is considered and on numerical issues. Thus, we should not see any difference in the results under infinite precision arithmetic. On the other side, alternative approaches can be proposed and they might not be equivalent to NIPALS, SIMPLS or one of our definitions. In such cases, we will not consider them as PLS. As a commentary, discrepancies of the PLS definition are common for the case of multivariate Y, for instance NIPALS and SIMPLS lead to different results. Nevertheless, they are unusual for the case of Y being a scalar.

4 PARTIAL LEAST SQUARES FOR FUNCTIONAL DATA

After reviewing the different techniques that can be applied to the finite-dimensional linear model, we want to explore the possible extensions for functional data.

4.1 The Functional Linear Model

The first section of this chapter aims at reproducing the basic theory for finite-dimensional linear regression models in the functional setting. In order to properly separate the regularization techniques from concepts and limitations of the linear models, this section will not cover any regularization method. Regularization aspects will be studied deeply in next section.

4.1.1 Definition of the model

First of all, it is required to define formally the functional linear model. Compared to finite-dimensional linear models, here the covariates X is not in $\mathbb{R}^p$ but $X(\omega, \cdot) \in L^2(\mathbf{T})$, a separable Hilbert space of functions for some domain $\mathbf{T}$. Depending on the characteristic of the response, two kinds of regression models can be established:

1. Functional regression with scalar response, i.e. the response $Y \in \mathbb{R}$.

$$y_i = \alpha + \int_{\mathbf{T}} x_i(t)\beta(t)dt + \epsilon_i.$$

2. Functional regression with functional response, i.e. the response $Y \in L^2(\mathbf{T})$. This can be a concurrent linear model (linear regression on each time t)

$$y_i(t) = \alpha + \beta(t)x_i(t) + \epsilon_i,$$

or a fully functional model

$$y_i(t) = \alpha(t) + \int_{\mathbf{T}} \beta(t,s)x_i(s)ds + \epsilon_i.$$

Since each approach leads to a rather different situation, we focus in only one model. For the sake of simplicity, only the first model will be considered, i.e.

$$Y = \int_{\mathbf{T}} \beta^*(t)X(t)dt + \epsilon \tag{4.1}$$

where X is an L^2 process with $E\left\|X_i\right\|^2 < \infty$, ϵ is a random variable with $E(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = E(\epsilon^2) = \sigma^2$ and $E(X(t)\epsilon) = 0$ and $\beta^* \in L^2(\mathbf{T})$.

4.1.2 Issues in the infinite-dimensional setting

The finite-dimensional linear regression model has been deeply studied through the years and has been used extensively in many practical applications. This is due to the fact that this model is rich in nice properties. Nevertheless, linear regression model in the functional setting loses some important properties that must be highlighted before starting our study of PLS:

- The covariance operator, analogous to $X'X$, is no longer invertible. This is due to the fact that the covariance operator is compact and compact operators in infinite-dimensional Hilbert spaces are not invertible.
- When we present the β_{OLS} estimator, we write it as $\beta_{\text{OLS}} = (X'X)^{-1}X'Y$. This is known to be the best linear unbiased estimator (Gauss-Markov theorem). Due to the high dimension of the space, in the functional setting there are not direct equivalents to β_{OLS} and thus, we have not an optimality result such as Gauss-Markov theorem. We usually build estimators using a basis representation and truncating it, that is, some kind of regularization is unavoidable.

4.2 Extensions of PLS for functional data

4.2.1 Krylov spaces in the functional data context

To begin with, Krylov spaces must be defined in the functional setting:

Definition 4.2.1 (Krylov space)**.** *Let X be an $L^2(\mathbf{T})$-integrable stochastic process and let R be the corresponding covariance integral operator of $L^2(\mathbf{T})$ defined by*

$$Rf(t) = \int_{\mathbf{T}} \rho(t,s)f(s)ds,$$

where $\rho(t,s) = \mathrm{Cov}(X(t), X(s))$. The Krylov space of dimension q is defined as follows:

$$\mathscr{K}_q(R,f) = span\{f, Rf, \ldots, R^{q-1}f\},$$

where $f \in L^2(\mathbf{T})$.

The following result gives a sufficient condition for the continuity of the operator R, which is often required:

Proposition 4.2.2. *Let $R\beta = \int_{\mathbf{T}} \rho(t,s)\beta(s)ds$. If $\rho \in L^2(\mathbb{R}^2)$, then R is a continuous operator in $L^2(\mathbb{R})$ and $\|R\| \leq \|\rho\|$*

Proof can be found in chapter 8 of [Heil, 2018] (see theorem 8.2.1). Note that we are requiring that $\|\rho\|^2 = \int_{\mathbf{T}} \int_{\mathbf{T}} \rho(t,s)^2 dtds$ is bounded. We check it:

$$\begin{aligned}
\|\rho\|^2 &= \int_{\mathbf{T}} \int_{\mathbf{T}} \left[E(X(t)X(s))\right]^2 dtds \\
&\leq \int_{\mathbf{T}} \int_{\mathbf{T}} E(X(t)^2)E(X(s)^2)dtds \qquad \text{(Cauchy-Schwarz in } L^2(\Omega)) \\
&= \int_{\mathbf{T}} E(X(t)^2)dt \int_{\mathbf{T}} E(X(s)^2)ds \\
&= \left[\int_{\mathbf{T}} E(X(t)^2)dt\right]^2 \\
&= \left[E\left(\int_{\mathbf{T}} X(t)^2 dt\right)\right]^2 \qquad \text{(Fubini's theorem)} \\
&= \left[E(\|X\|^2)\right]^2 < \infty,
\end{aligned}$$

because we imposed the hypothesis $E\|X\|^2 < \infty$, i.e. R is always continuous under our assumptions.

Now, we relate the operator R with the parameter β. We start from the model equation

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

multiply by X and take expectation at both sides:

$$E(X(t)Y) = E(X(t)\langle X, \beta \rangle) + E(X(t)\epsilon).$$

Since $E(X(t)\epsilon) = 0$, it follows

$$E(X(t)Y) = E(X(t)\langle X, \beta \rangle).$$

Using Fubini and taking into account the model, we can rewrite this in terms of the operator R

$$\begin{aligned}
E(X(t)\langle X, \beta \rangle) &= \int_{\Omega} X(t) \int_{\mathbf{T}} X(s)\beta(s)dsdP \\
&= \int_{\mathbf{T}} \int_{\Omega} X(t)X(s)dP\beta(s)ds \\
&= \int_{\mathbf{T}} \rho(t,s)\beta(s)ds \\
&= R\beta.
\end{aligned}$$

Consequently, our problem is now:

$$E(X(t)Y) = \mathrm{Cov}(X(t), Y) = R\beta,$$

and it is equivalent to solve the following minimization problem that becomes our definition of FPLS, as we did in the finite dimensional case.

Definition 4.2.3 (FPLS Version 1). *Let X be a (centered) L^2-integrable stochastic process and Y an $L^2(\Omega)$ random variable, the (centered) response. Then, $\beta_{PLS}^{(q)} \in L^2(\mathbf{T})$ is defined as follows*

$$\beta_{PLS}^{(q)} := \underset{\beta \in \mathscr{K}_q}{\arg\min}\, E(\langle X, \beta \rangle - Y)^2, \tag{4.2}$$

where $\mathscr{K}_q = \mathscr{K}_q(R, b)$ is determined by $Rf(t) = \int_{\mathbf{T}} \rho(t,s) f(s) ds$ and $b(t) = \mathrm{Cov}(X(t), Y)$.

The minimization problem is simpler than it seems because $\mathscr{K}_q(R, b)$ is a finite dimensional space, so the problem is equivalent to find the coefficients that minimize $\left\| \sum_{i=0}^{q-1} c_i \langle X, R^i f \rangle - Y \right\|_{L^2(\Omega, P)}$.

4.2.2 Functional conjugate gradient

In the previous chapter, we covered the conjugate gradient algorithm and show that it is equivalent to other PLS definitions. The advantage of using the conjugate gradient instead of approaching directly the problem is that this method is numerically stable, in other words, it was thought to be implemented in a computer so that the error in each step is not accumulated.

Definition 4.2.4 (FPLS Version 2). *Let R be the operator $Rf = Rf(t) = \int_{\mathbf{T}} \rho(t,s) f(s) ds$ and let $b(t) = E(X(t)Y)$. Let x_0 be an initial approximation that can be 0 and $v_0 = -g_0 = b - Rx_0$. The solution for x of the system $Rx = b$ can be computed using the conjugate gradient algorithm as follows:*

$$x_{k+1} = x_k + \alpha_k v_k, \tag{4.3}$$

where

$$g_k = Rx_k - b,$$

$$\alpha_k = -\frac{\langle g_k, v_k \rangle}{\langle v_k, Rv_k \rangle},$$

$$v_{k+1} = -g_{k+1} + \gamma_k v_k,$$

and

$$\gamma_k = \frac{\langle g_{k+1}, Rv_k \rangle}{\langle v_k, Rv_k \rangle}.$$

In next steps, we will reproduce the same results that led us to the equivalence between the two definitions in the finite-dimensional case. Whenever needed, stronger hypothesis are imposed.

Theorem 4.2.5. *(Functional Conjugate Direction Theorem) Let R be a continuous linear operator of $L^2(\mathbf{T})$. Let $\{v_i\}_{i=0}^{\infty}$ be a sequence of nonzero conjugated (with*

respect to R) functions of $L^2(\mathbf{T})$ and $x_0 \in L^2(\mathbf{T})$ an initial approximation. Suppose that the solution β^* to the problem $R\beta^* = b$ exists and it can be expressed in terms of the sequence $\{v_i\}_{i=0}^{\infty}$, i.e. $\beta^* = x_0 + \sum_{i=0}^{\infty} \gamma_i v_i$ Then the following iterative method converge to the solution of problem $R\beta = b$

$$x_{k+1} = x_k + \alpha_k v_k, \quad k \geq 0 \tag{4.4}$$

where

$$g_k = Rx_k - b$$

and

$$\alpha_k = -\frac{g_k' v_k}{v_k' R v_k}.$$

Proof. As for the finite case, we compute an expression for the coefficient γ_i:

$$\begin{aligned}
\beta^* - x_0 &= \sum_{i=0}^{\infty} \gamma_i v_i, \\
R(\beta^* - x_0) &= R\left(\sum_{i=0}^{\infty} \gamma_i v_i\right) && \text{(Apply operator } R\text{)} \\
&= \sum_{i=0}^{\infty} \gamma_i R v_i, && \text{(Continuity and linearity)} \\
\langle v_k, R(\beta^* - x_0)\rangle &= \langle v_k, \sum_{i=0}^{\infty} \gamma_i R v_i\rangle && \text{(Multiply by } v_k\text{)} \\
&= \sum_{i=0}^{\infty} \gamma_i \langle v_k, R v_i\rangle && \text{(Continuity of scalar product)} \\
&= \gamma_k \langle v_k, R v_k\rangle. && \text{(Conjugated functions)}
\end{aligned}$$

This leads to the expression

$$\gamma_k = \frac{\langle v_k, R(\beta^* - x_0)\rangle}{\langle v_k, R v_k\rangle} = \frac{\langle v_k, b - Rx_k\rangle}{\langle v_k, R v_k\rangle}.$$

For x_k, we apply (4.4) recursively:

$$x_k = x_{k-1} + \alpha_{k-1} v_{k-1} = \cdots = x_0 + \sum_{i=0}^{k-1} \alpha_i v_i,$$

what leads to the useful equality $\langle v_k, Rx_0\rangle = \langle v_k, Rx_k\rangle$. Then,

$$\gamma_k = \frac{\langle v_k, b - Rx_k\rangle}{\langle v_k, R v_k\rangle}$$

$$
\begin{aligned}
x_k - x_0 &= \sum_{i=0}^{k-1} \alpha_i v_i, \\
R(x_k - x_0) &= \sum_{i=0}^{k-1} \alpha_i R(v_i), && \text{(Apply operator R and linearity)} \\
\langle v_k, R(x_k - x_0)\rangle &= \sum_{i=0}^{k-1} \alpha_i \langle v_k, R(v_i)\rangle && \text{(Multiply by } v_k \text{ and linearity)} \\
&= 0. && \text{(Conjugated functions)}
\end{aligned}
$$

Applying this last equality in the expression of γ_k, we get:

$$\gamma_k = \frac{\langle v_k, b - Rx_k)\rangle}{\langle v_k, Rv_k\rangle} = -\frac{\langle v_k, g_k\rangle}{\langle v_k, Rv_k\rangle} = \alpha_k,$$

which implies that $\lim_{k\to\infty} x_k = \beta^*$. □

As a corollary, we get that g_k is orthogonal to span$\{v_i\}_{i=0}^{k-1}$.

Corollary 4.2.5.1. *Under the hypothesis of the previous theorem, g_k is orthogonal to span$\{v_i\}_{i=0}^{k-1}$.*

Proof. By induction, the result is clear for $k = 0$ because the space is empty. Assume the result is true for k and consider the expression of g_{k+1}.

$$
\begin{aligned}
g_{k+1} &= g_k + \alpha_k R v_k, \\
\langle v_i, g_{k+1}\rangle &= \langle v_i, g_k\rangle + \alpha_k \langle v_i, Rv_k\rangle && \text{(Scalar product with } v_k) \\
&= 0.
\end{aligned}
$$

Both terms vanish if $i < k$, the first one due to the induction hypothesis, and the second one because the sequence $\{v_i\}_{i=0}^{k-1}$ is conjugate to v_k. If $i = k$, we have just to recall the expression of α_k and check that it is also 0.

$$
\begin{aligned}
\langle v_k, g_{k+1}\rangle &= \langle v_k, g_k\rangle + \alpha_k \langle v_k, Rv_k\rangle \\
\langle v_k, g_{k+1}\rangle &= \langle v_k, g_k\rangle - \frac{\langle v_k, g_k\rangle}{\langle v_k, Rv_k\rangle}\langle v_k, Rv_k\rangle \\
&= 0
\end{aligned}
$$

□

Once we proved that, under certain hypothesis, the algorithm must converge, we need to prove that the sequence $\{v_i\}_{i\in\mathbb{N}}$ generated by the algorithm of functional conjugate gradient is indeed a sequence of conjugated functions.

Theorem 4.2.6. *Consider the algorithm of definition 4.2.4. If $x_k \neq x_{k+1}$, the following properties hold:*

(a) span$\{g_0, \ldots, g_k\}$ = span$\{v_0, \ldots, v_k\}$ = $\mathscr{K}_{k+1}(R, g_0)$ = span$\{g_0, Rg_0..., R^k g_0\}$.

(b) v_k is conjugate to any v_i, i.e. $\langle v_k, Rv_i\rangle$, with $i < k$.

Proof. The proof of (a) is by induction. For $k = 0$, we have span$\{g_0\}$ = span$\{v_0\}$ = $\mathscr{K}_1(R, g_0)$ = span$\{g_0\}$. Assume the equalities hold for k, and consider the expressions for $k + 1$:

$$g_{k+1} = g_k + \alpha_k R v_k.$$

It is clear that $g_{k+1} \in \mathscr{K}_{k+2}(R, g_0)$, because $g_k \in \mathscr{K}_{k+1}(R, g_0) \subset \mathscr{K}_{k+2}(R, g_0)$ and $v_k \in \mathscr{K}_{k+1}(R, g_0)$ (i.e. $Rv_k \in \mathscr{K}_{k+2}(R, g_0)$).

Moreover, $g_{k+1} \notin \mathscr{K}_{k+1}(R, g_0)$. Because, by induction hypothesis, $\mathscr{K}_{k+1}(R, g_0)$ = span$\{v_0, \ldots, v_k\}$ and g_{k+1} is orthogonal to span$\{v_0, \ldots, v_k\}$ due to corollary 4.2.5.1. Consequently, it can be concluded that:

$$\text{span}\{g_0, \ldots, g_{k+1}\} = \text{span}\{g_0, Rg_0, \ldots, R^{k+1}g_0\}.$$

To prove the second equality, we write the formula for v_{k+1},

$$v_{k+1} = -g_{k+1} + \gamma_k v_k,$$

which implies that $v_{k+1} \in \mathscr{K}_{k+2}(R, g_0)$ and span$\{v_0, \ldots, v_{k+1}\}$ = span$\{g_0, Rg_0, \ldots, R^{k+1}g_0\}$. To prove (b), we need to show that $\langle v_{k+1}, Rv_i\rangle = 0$ for $i < k$. For $i = k$, the scalar product is zero due to the definition of $\gamma_k = \frac{\langle g_{k+1}, Rv_k\rangle}{\langle v_k, Rv_k\rangle}$.

$$\begin{aligned}\langle v_{k+1}, Rv_k\rangle &= \langle -g_{k+1} + \gamma_k v_k, Rv_k\rangle \\ &= \frac{\langle g_{k+1}, Rv_k\rangle}{\langle v_k, Rv_k\rangle}\langle v_k, Rv_k\rangle - \langle g_{k+1}, Rv_k\rangle \\ &= 0.\end{aligned}$$

If $i < k$, first term vanish due to the induction hypothesis. The second term vanish because g_{k+1} is orthogonal to span$\{v_0, \ldots, v_{i+1}\}$ due to corollary 4.2.5.1.

$$\begin{aligned}\langle v_{k+1}, Rv_i\rangle &= \langle -g_{k+1} + \gamma_k v_k, Rv_i\rangle \\ &= \frac{\langle g_{k+1}, Rv_k\rangle}{\langle v_k, Rv_k\rangle}\langle v_k, Rv_i\rangle - \langle g_{k+1}, Rv_i\rangle \\ &= 0.\end{aligned}$$

□

4.2.3 Extended statistical criterion

Several authors, such as [Delaigle and Hall, 2012, Febrero-Bande et al., 2017] and [Preda and Saporta, 2005], have defined PLS for functional data from a statistical point of view. [Febrero-Bande et al., 2017, Preda and Saporta, 2005] present the algorithm that we will be employing in this work and a simple convergence result.

Although the method in [Delaigle and Hall, 2012] achieves both convergence and convergence rates, the method relies on explicit construction of non-orthogonal PLS directions. This means that the method might be slightly different from the standard definition in numerical experiments and consequently, we choose as a reference the one presented in [Febrero-Bande et al., 2017].

Definition 4.2.7 (FPLS Version 3). *Let $y_0 = y - E(Y)$ and $X_0 = X(t) - E(X(t))$*

For $l = 0, 1, \ldots,$ the following iteration is defined:

- *Let $\phi_{l+1} \in L^2(\mathbf{T})$ such that it maximizes $\mathrm{Cov}^2(y_l, \langle X_l, \phi_{l+1}\rangle)$. Hereinafter, $p_{l+1} = \langle X_l, \phi_{l+1}\rangle$. This function can be explicitly expressed as:*

$$\phi_{l+1} = \frac{\mathrm{Cov}(y_l, X_l(t))}{\|\mathrm{Cov}(y_l, X_l(t))\|}.$$

- *Let $y_{l+1} = y_l - v_{l+1}p_{l+1}$ where v_{l+1} is*

$$v_{l+1} = \frac{\mathrm{Cov}(y_l, p_{l+1})}{\mathrm{Var}(p_{l+1})}.$$

- *Let $X_{l+1}(t) = X_l(t) - p_{l+1}\varrho_{l+1}(t)$ where $\varrho_{l+1}(t)$ is*

$$\varrho_{l+1}(t) = \frac{\mathrm{Cov}(X_l(t), p_{l+1})}{\mathrm{Var}(p_{l+1})}.$$

Given the q-th iteration, we have that

$$\beta_{PLS}^{(q)}(t) = \sum_{l=1}^{q} v_l \varphi_l(t),$$

where $\varphi_l(t) = \phi_l(t) - \sum_{1<j<l} \langle \varrho_j, \phi_l \rangle \varphi_j$.

The following result presents the principal statistical properties of this version of PLS:

Proposition 4.2.8. *Using the notation of Definition 4.2.7, the following properties hold:*

- $\{p_l\}_{l\in\mathbb{N}}$ *is an orthogonal basis of* $L^2(X)$*, such that*

$$X(t) = E(X(t)) + \sum_{l=1}^{\infty} p_l \varrho_l(t).$$

- $Y = E(Y) + \sum_{l=1}^{\infty} v_l p_l + e$.
- $\beta(t) = \sum_{l=1}^{\infty} v_l \varphi_l(t)$ *where* $\varphi_l(t) = \phi_l(t) - \sum_{1<j<l} \langle \varrho_j, \phi_l \rangle \varphi_j$.
- *The coefficient of determination can be computed as follows:*

$$R^2 = \sum_{l=1}^{\infty} \mathrm{Corr}^2(y, p_l).$$

Also, if there exists a $\beta^* \in L^2(\mathbf{T})$ *such that* $b = E(X(t)Y) = R\beta^*$*. Let* $Y_{PLS}^{(q)}$ *be the response estimated by PLS of order* q*, then* $Y_{PLS}^{(q)} = E(Y) + c_1 t_1 + \cdots + c_q t_q$ *and*

$$\lim_{q\to\infty} E(|Y_{PLS}^{(q)} - \langle X, \beta^* \rangle|^2) = 0.$$

The proof can be found in [Preda and Saporta, 2005]

4.3 Properties of PLS

4.3.1 Structure of RKHS

The following results establish some relationships between Krylov spaces and RKHS.

Theorem 4.3.1 (Kernel operator)**.** *Let* $\rho(t,s) = \mathrm{Cov}(X(t), X(s))$ *be the covariance function. Define* $k : L^2(\mathbf{T}) \times L^2(\mathbf{T}) \to \mathbb{R}$ *as*

$$k(f,g) = \int_{\mathbf{T}} \int_{\mathbf{T}} \rho(t,s) f(t) g(s) dt ds. \tag{4.5}$$

Then, k is a kernel (that is, it is symmetric and positive definite).

Proof. The operator is clearly symmetric because $\rho(t,s)$ is symmetric. Let $f_i \in L^2(\mathbf{T})$ and let $\alpha_i \in \mathbb{R}$ for $i = 1, \ldots, M$. Also, without loss of generality, we assume $EX(t) = 0$ and we use Fubini

$$\begin{aligned} k(f,g) &= \int_{\mathbf{T}} \int_{\mathbf{T}} \rho(t,s) f(t) g(s) dt ds \\ &= \int_{\mathbf{T}} \int_{\mathbf{T}} E(X(t), X(s)) f(t) g(s) dt ds \\ &= E\left(\int_{\mathbf{T}} \int_{\mathbf{T}} X(t) X(s)) f(t) g(s) dt ds \right) \qquad \text{(Fubini's theorem)} \\ &= E\left(\int_{\mathbf{T}} X(t) f(t) dt \int_{\mathbf{T}} X(s) g(s) ds \right) \\ &= E\left(\langle X, f \rangle \langle X, g \rangle \right) \\ &= \operatorname{Cov}\left[\langle X, f \rangle, \langle X, g \rangle \right]. \end{aligned}$$

Hence, we have that

$$\begin{aligned} \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_i \alpha_j k(f_i, f_j) &= k\left(\sum_{i=1}^{M} \alpha_i f_i, \sum_{j=1}^{M} \alpha_j f_j \right) \qquad \text{(Bilinearity of } k\text{)} \\ &= \operatorname{Var}\left[\sum_{i=1}^{M} \alpha_i \langle X, f_i \rangle \right] \geq 0 \end{aligned}$$

This proves that k is positive definite and then, a kernel. □

Observe that due to the Moore-Aroszajn theorem (Theorem 2.5.5), there exists a unique RKHS for which k is a reproducing kernel.

Corollary 4.3.1.1 (Operator R and kernel operator)**.** *If k is the aforementioned kernel and $Rf(\cdot) = \int_{\mathbf{T}} \rho(t, \cdot) f(t) dt$, then*

$$\langle f, Rg \rangle = \langle Rf, g \rangle = k(f,g).$$

In particular, R is self-adjoint.

Proof. It is an immediate consequence of the symmetry of ρ

$$\begin{aligned}\langle Rf, g\rangle &= \int_{\mathbf{T}}\int_{\mathbf{T}} \rho(t,s)f(t)dtg(s)ds \\ &= \int_{\mathbf{T}}\int_{\mathbf{T}} \rho(t,s)f(t)g(s)dtds \\ &= \int_{\mathbf{T}}\int_{\mathbf{T}} \rho(s,t)f(t)g(s)dsdt \\ &= \langle f, Rg\rangle.\end{aligned}$$

□

Note that since the conjugate gradient and some other implementations of PLS use sequences of R-orthogonal vectors, this means that these sequences are orthogonal in the RKHS defined by the kernel k. Nevertheless, this RKHS has a special property with respect to uncorrelation.

Theorem 4.3.2. *Let k be the kernel asociated to ρ as defined above. Also, let $f, g \in L^2(\mathbf{T})$ deterministic functions. Then, k is given by*

$$k(f,g) = \mathrm{Cov}(\langle X, f\rangle, \langle X, g\rangle), \tag{4.6}$$

and $k(f,g) = 0$ means that $\langle X, f\rangle$ and $\langle X, g\rangle$ are uncorrelated random variables.

Proof. Without loss of generality, assume X is centered. Then, using Fubini (everything is L^2, so product is in L^1), we can write that

$$\begin{aligned}k(f,g) &= \int_{\mathbf{T}}\int_{\mathbf{T}} \rho(t,s)f(t)g(s)dtds \\ &= \int_{\mathbf{T}}\int_{\mathbf{T}} E(X(t)X(s))f(t)g(s)dtds \\ &= E\left[\int_{\mathbf{T}} f(t)X(t)dt \int_{\mathbf{T}} g(s)X(s)ds\right] \\ &= E\left[\langle X, f\rangle\langle X, g\rangle\right].\end{aligned}$$

Note that $E\left[\langle X, f\rangle\right]$ is 0 because X is centered

$$\begin{aligned}E\left[\langle X, f\rangle\right] &= E\left[\int_{\mathbf{T}} X(t)f(t)dt\right] \\ &= \int_{\mathbf{T}} E\left[X(t)\right]f(t)dt \\ &= 0.\end{aligned}$$

Thus, the expression before is the $\mathrm{Cov}(\langle X, f\rangle, \langle X, g\rangle)$, i.e.

$$k(f,g) = \mathrm{Cov}(\langle X, f\rangle\langle X, g\rangle).$$

□

This result is key in understanding what happens behind the PLS algorithm, because it allowed us to link sequences of R-orthogonal functions to sequences of uncorrelated predictions. Unfortunately, we can not ensure this is connected to independence at all unless we have strong assumptions on the distribution of X.

Theorem 4.3.3. *Let X be a Gaussian process with 0 mean and continuous covariance function ρ. Let k be the kernel asociated to ρ as defined above. Then, if $k(f,g) = 0$, $\langle X, f\rangle$ and $\langle X, g\rangle$ are independent random variable with normal distribution.*

Proof. Under these hypothesis, X can be decomposed using the Karhunen-Loève expansion: (Theorem 2.3.2)

$$X(t) = \sum_{i=1}^{\infty} Z_i e_i(t),$$

where $Z_i = \langle X, e_i\rangle$ are jointly Gaussian (Corollary 2.3.2.1) and $\{e_i\}_{i\in\mathbb{N}}$ form a basis of $L^2(\mathbf{T})$. Let $f = \sum_{i=1}^{\infty}\alpha_i e_i(t)$ and let $g = \sum_{i=1}^{\infty}\gamma_i e_i(t)$. Then, the bidimensional vector of scalar products of X with f and g can be written as follows:

$$\begin{aligned}
[\langle X, f\rangle, \langle X, g\rangle] &= [\langle\sum_{i=1}^{\infty} Z_i e_i(t), \sum_{i=1}^{\infty}\alpha_i e_i(t)\rangle, \langle\sum_{i=1}^{\infty} Z_i e_i(t), \sum_{i=1}^{\infty}\gamma_i e_i(t)\rangle] \\
&= [\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} Z_i\alpha_j\langle e_i, e_j\rangle, \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} Z_i\gamma_j\langle e_i, e_j\rangle] \\
&= [\sum_{i=1}^{\infty}\alpha_i Z_i, \sum_{i=1}^{\infty}\gamma_i Z_i] \\
&= \sum_{i=1}^{\infty}[\alpha_i Z_i, \gamma_i Z_i].
\end{aligned}$$

This means, $[\langle X, f\rangle, \langle X, g\rangle]$ is a series of independent normal bivariate random vectors. To ensure that this series converges to a normal bivariate random vector, we use the characteristic functions of a Gaussian:

$$\begin{aligned}
\prod_{i=1}^{\infty}\exp\Big(\boldsymbol{\mu_i}'\mathbf{t} + \tfrac{1}{2}\mathbf{t}'\boldsymbol{\Sigma_i}\mathbf{t}\Big) &= \exp\Big(\sum_{i=1}^{\infty}\boldsymbol{\mu_i}'\mathbf{t} + \tfrac{1}{2}\sum_{i=1}^{\infty}\mathbf{t}'\boldsymbol{\Sigma_i}\mathbf{t}\Big) \\
&= \exp\left(\left(\sum_{i=1}^{\infty}\boldsymbol{\mu_i}'\right)\mathbf{t} + \tfrac{1}{2}\mathbf{t}'\left(\sum_{i=1}^{\infty}\boldsymbol{\Sigma_i}\right)\mathbf{t}\right).
\end{aligned}$$

Given that $\sum_{i=1}^{\infty} \boldsymbol{\mu_i}$ and $\sum_{i=1}^{\infty} \boldsymbol{\Sigma_i}$ are convergent series, by Lévy's continuity theorem, the series converges in distribution to a normal bivariate random vector. Consequently, if $\langle X, f\rangle$ and $\langle X, g\rangle$ are uncorrelated, this means that they are also independent.

□

In conclusion, this section connected k and R to the well-known idea of maximizing covariance of PLS. Also, stronger hypothesis lead to stronger results such as independence.

4.4 Residuals of gradient methods

Proving the equivalence of definitions 4.2.3 and 4.2.4 relies on different arguments in the infinite dimensional situation compare to those in the finite-dimensional case. The following proposition shows that conjugate gradient is the solution to a minimization problem:

Proposition 4.4.1. *Let x_q be the q-th iteration of the conjugate gradient algorithm. Assume that there exists $\beta^* \in L^2(\mathbf{T})$ such that $R\beta^* = b$. Then, x_q is a solution to the minimization problem*

$$x_q = \underset{\beta \in \mathscr{K}_q(R,b)}{\arg\min}\, \mathscr{E}(\beta) = \underset{\beta \in \mathscr{K}_q(R,b)}{\arg\min}\, k(\beta^* - \beta, \beta^* - \beta).$$

Proof. Let $z \in \mathscr{K}_q(R, b)$, we compute $\mathscr{E}(z) - \mathscr{E}(x_q)$:

$$\begin{aligned}
\mathscr{E}(z) - \mathscr{E}(x_q) &= \mathscr{E}(x_q - (x_q - z)) - \mathscr{E}(x_q) \\
&= k(\beta^* - x_q + (x_q - z), \beta^* - x_q + (x_q - z)) - k(\beta^* - x_q, \beta^* - x_q) \\
&= \langle \beta^* - x_q + (x_q - z), R(\beta^* - x_q + (x_q - z))\rangle - \langle \beta^* - x_q, R(\beta^* - x_q)\rangle \\
&= 2\langle \beta^* - x_q, R(x_q - z)\rangle + \langle x_q - z, R(x_q - z)\rangle,
\end{aligned}$$

where in the last equality we have applied Corollary 4.3.1.1. Next, $\langle \beta^* - x_q, R(x_q - z)\rangle$ can be easily seen to vanish:

$$\begin{aligned}
\langle \beta^* - x_q, R(x_q - z)\rangle &= \langle R(\beta^* - x_q), x_q - z\rangle && (R \text{ is self-adjoint}) \\
&= \langle b - Rx_q, x_q - z\rangle && (R\beta^* = b) \\
&= -\langle g_q, x_q - z\rangle && (g_q = Rx_q - b) \\
&= 0. && (z, x_q \in \mathscr{K}_q(R, b) \text{ and } g_q \perp \mathscr{K}_q(R, b))
\end{aligned}$$

This implies that:

$$\begin{aligned}
\mathscr{E}(z) - \mathscr{E}(x_q) &= \langle x_q - z, R(x_q - z)\rangle \\
&= k(x_q - z, x_q - z) > 0, && (k \text{ is positive definite})
\end{aligned}$$

and consequently, x_q is a minimum of $\mathscr{E}$. □

Another ingredient to achieve the proof of the equivalence is that the problem in the last result is equivalent to minimization of squared residuals in a Krylov space:

Corollary 4.4.1.1. *Let x_q as in the previous theorem. Let β^* be such that $Y = \langle X, \beta^* \rangle + \epsilon$ for some ϵ random variable with zero mean and finite variance that is independent of X. Then, x_q also solves the following minimization problem:*

$$x_q = \underset{\beta \in \mathcal{K}_q(R,b)}{\arg\min}\ E(Y - \langle X, \beta \rangle)^2.$$

Proof. Without loss of generality, assume $E(X(t)) = 0$. Let $x_q = \arg\min_{\beta \in \mathcal{K}_q(R,b)} \mathscr{E}(\beta) = \arg\min_{\beta \in \mathcal{K}_q(R,b)} k(\beta^* - \beta, \beta^* - \beta)$. Then:

$$\begin{aligned}
x_q &= \underset{\beta \in \mathcal{K}_q(R,b)}{\arg\min}\ k(\beta^* - \beta, \beta^* - \beta) \\
&= \underset{\beta \in \mathcal{K}_q(R,b)}{\arg\min}\ \langle \beta^* - \beta, R(\beta^* - \beta) \rangle \\
&= \underset{\beta \in \mathcal{K}_q(R,b)}{\arg\min}\ E\left((\langle X, \beta^* \rangle - \langle X, \beta \rangle)^2 \right) && \text{(Fubini's theorem)} \\
&= \underset{\beta \in \mathcal{K}_q(R,b)}{\arg\min}\ E\left((Y - \epsilon - \langle X, \beta \rangle)^2 \right) && \text{(Relation between } \beta^* \text{ and } Y) \\
&= \underset{\beta \in \mathcal{K}_q(R,b)}{\arg\min}\ E\left((Y - \langle X, \beta \rangle)^2 + \epsilon^2 - 2\epsilon(Y - \langle X, \beta \rangle) \right) && (a-b)^2 = a^2 + b^2 - 2ab \\
&= \underset{\beta \in \mathcal{K}_q(R,b)}{\arg\min}\ E\left((Y - \langle X, \beta \rangle)^2 - 2\epsilon(Y - \langle X, \beta \rangle) \right) && (E(\epsilon^2) \text{ is a constant}) \\
&= \underset{\beta \in \mathcal{K}_q(R,b)}{\arg\min}\ E\left((Y - \langle X, \beta \rangle)^2 \right). && (\epsilon \text{ is independent of } X(t))
\end{aligned}$$

□

Note that the last two results, give us the equivalence between definitions 4.2.3 and 4.2.4 under certain assumptions. Recall that $\beta^* \in L^2(\mathbf{T})$ is such that $Y = \langle X, \beta^* \rangle + \epsilon$, where ϵ is a random variable independent from X, with zero mean and finite variance. Recall also that $R\beta^* = b$:

$$\begin{aligned}
R\beta^* &= \int_{\mathbf{T}} \rho(t,s)\beta^*(t)dt \\
&= \int_{\mathbf{T}} E(X(t)\beta^*(t)X(s))dt \\
&= E(\langle X, \beta^* \rangle X(s)) && \text{(Fubini's theorem)} \\
&= E(YX(s)) - E(\epsilon X(s)) \\
&= b. && \text{(Independence)}
\end{aligned}$$

Next, we want to investigate the convergence of x_q when $q \to \infty$, that is the dimension of the Krylov space get larger.

Theorem 4.4.2. *Under the assumptions of Corollary 4.4.1.1, let*

$$x_q = \underset{\beta \in \mathscr{K}_q(R,b)}{\arg\min}\, E\left((Y - \langle X, \beta\rangle)^2\right)$$

and

$$x_\infty = \underset{\beta \in \overline{\bigcup_{q=1}^{\infty} \mathscr{K}_q(R,b)}}{\arg\min}\, E\left((Y - \langle X, \beta\rangle)^2\right).$$

Then, $\|x_q - x_\infty\|^2_{L^2(\mathbf{T})} \to 0$ *when* $q \to \infty$.

Proof. First, if there exists a maximal Krylov space ($\mathscr{K}_Q(R,b) = \overline{\bigcup_{q=1}^{\infty} \mathscr{K}_q(R,b)}$), the result is clearly true since $x_i = x_Q$ for $i > Q$. Assume $\overline{\bigcup_{q=1}^{\infty} \mathscr{K}_q(R,b)}$ is a infinite dimensional space, i.e. there is no maximal Krylov space.

Let $B = \{b, Rb, R^2b, \dots\}$ be a basis of $\bigcup_{q=1}^{\infty} \mathscr{K}_q(R,b)$. Performing a Gramm-Schmidt orthonormalization, we build an orthonormal basis $\mathbf{B} = \{v_1, v_2, \dots\}$. Let $\mathbf{B}_q = \{v_1, v_2, \dots v_q\}$ be a orthonormal basis of $\mathscr{K}_q(R,b)$.

Next, we write x_q and x_∞ in terms of these basis: $x_q = \sum_{i=1}^{q} \alpha_i v_i$ and $x_\infty = \sum_{i=1}^{\infty} \alpha_i v_i$. Coefficients $\{\alpha_i\}$ happen to be the same because regression is performed in orthogonal components, i.e. X can be decomposed $X_0, X_1, X_2, \dots$ where $X_0 \in \overline{\bigcup_{q=1}^{\infty} \mathscr{K}_q(R,b)}^{\perp}$ and X_i is the projection of X onto v_i for $i \in \mathbb{N}$, and regression can be performed on each component separately.

We write then the difference

$$\begin{aligned} x_q - x_\infty &= \sum_{i=1}^{q} (\alpha_i v_i(t)) - \sum_{i=1}^{\infty} (\alpha_i v_i(t)) \\ &= \sum_{i=q}^{\infty} \alpha_i v_i(t) \end{aligned}$$

and we have just to use Parseval's identity to compute the L^2-norm

$$\|x_q - x_\infty\|^2_{L^2(\mathbf{T})} = \sum_{i=q}^{\infty} \alpha_i^2 \to 0,$$

since $\sum_{i=q}^{\infty} \alpha_i^2$ is the tail of the convergent series $\sum_{i=1}^{\infty} \alpha_i^2 = \|x_\infty\|^2$. □

As a corollary, we can write that x_q converges to β^* in the L^2 sense.

Corollary 4.4.2.1. *Let x_q be the $q-th$ conjugate gradient iteration. Let $\beta^* \in L^2(\mathbf{T})$ such that $\beta^* \in \overline{\bigcup_{i=1}^{\infty} \mathscr{K}_i(R,b)}$ and $R\beta^* = b$. Then,*

$$\|x_q - \beta^*\|^2_{L^2(\mathbf{T})} \to 0,$$

when $q \to \infty$.

Proof. Since $\beta^* = \arg\min_{\beta \in L^2(\mathbf{T})} E\left((Y - \langle X, \beta\rangle)^2\right)$ and $\beta^* \in \overline{\bigcup_{i=1}^{\infty} \mathscr{K}_i(R,b)}$, $x_\infty = \arg\min_{\beta \in \overline{\bigcup_{i=1}^{\infty} \mathscr{K}_i(R,b)}} E\left((Y - \langle X, \beta\rangle)^2\right) = \beta^*$. Applying previous theorem, we have that $\|x_q - x_\infty\|^2_{L^2(\mathbf{T})} = \|x_q - \beta^*\|^2_{L^2(\mathbf{T})} \to 0$. □

Note that the hypothesis of $\beta^* \in \overline{\bigcup_{i=1}^{\infty} \mathscr{K}_i(R,b)}$ can be replaced by the stronger assumption:

$$\overline{\bigcup_{i=1}^{\infty} \mathscr{K}_q(R,b)} = L^2(\mathbf{T}).$$

An interesting open question is to study conditions (probably in terms of the eigenvalues of R) such that the last equality holds. Under such conditions the conjugate gradient we have described would converge to β^*, for any $\beta^* \in L^2(\mathbf{T})$.

4.5 Computational aspects

4.5.1 Calculating scalar products and basic operations

It is important to remark which operations must be performed over functional observations to ensure the feasibility of the algorithm in a computer. These operations can be summarized into these basic computations:

- Multiply a functional observation by a constant,
- Add two functional observations,
- Compute the scalar product between two functional observations,
- Compute the integral operator R.

Because of the difference in complexity between the first three operations and the last one, the last one will be covered in the next subsection. As a remark, one must notice that the first two properties aim at computing linear combinations of functional observations, so that the mean or our conjugate direction methods can be performed.

For the scalar product problem, we need to consider a basis, preferably an orthonormal basis. Since the computer can not deal with infinitely-many elements, the basis also should have some properties to represent properly the functional samples. Through the adoption of one basis instead of another, one can change the way the scalar products are computed and the final result of the

Algorithm 4.1 Implementation of integral operator R

```
R <- function(fd, g, inprod = fda::inprod) {
  N = length(fd$coefs[1,])
  s = 0 * g
  innerProducts = inprod(fd,g)
  for (i in 1:N) {
      x = fd[i]
      s = s + innerProducts[i] * x
  }
  return(s * (1/N))
}
```

process. Consequently, several basis are tested in order to ensure the reliability of the scalar product computations.

4.5.2 Calculating the kernel operator R and the transformation of the response

The formula of the kernel operator is not directly computable in a computer

$$Rf(t) = \int_{\mathbf{T}} \rho(t,s)f(s)ds$$

but we recall that we proved that

$$Rf(t) = E(X(t)\langle X, f\rangle).$$

Note that this expression does not explicitly includes either the integral operator or the covariance operator. Also, it was computed in terms of scalar products and linear combinations of functional observations. In a similar way, we can compute the transformation of the response

$$E(X(t)Y) \approx \frac{1}{N}\sum_{i=1}^{N} X_i(t)Y_i$$

which is computed again in terms of linear combinations of observations. Algorithm 4.1 and Algorithm 4.2 contains both implementations in R. Note that the function SUM does not work as intended but returns a value that it is not the sum of the functional observations.

Algorithm 4.2 Implementation of operator $b(t) = E(X(t)Y)$

```
YX_t <- function(fd, y) {
  N = length(fd$coefs[1,])
  s = 0 * fd[1]
  for (i in 1:N) {
    x = fd[i]
    s = s + y[i] * x
  }
  return(s * (1/N))
}
```

5 APPLICATIONS AND PERFORMANCE COMPARISONS

5.1 Introduction

This chapter presents a collection of results in both simulations and real-world datasets to benchmark PLS and PCR to check whether there are some improvements and in which situations we expect to see them.

5.2 Algorithms and implementations

5.2.1 Partial Least Squares

We have studied in previous chapters several ways of implementing PLS. For instance, [Febrero-Bande et al., 2017] implements PLS through a estimation of the PLS directions and doing regression on these directions. [Delaigle and Hall, 2012] implements PLS in a similar way but using nonorthogonal components. We think that this method can be ill-conditioned under some cases, whereas Conjugate Gradient is a simple, geometric algorithm that allows us to solve the system in a consistent and stable way. Algorithm 5.1 shows the implementation of the algorithm we want to use for the experiments and hereinafter will be our implementation of PLS.

5.2.2 Principal Components Regression

On the other hand, PCR relies on a simpler algorithm that just consists on:

1. Compute the (orthonormal) principal components ψ_k and the scores a_{ik}, i.e. the ith-sample $x_i(t) = \sum_{k=1}^{\infty} a_{ik}\psi_k$. We truncate this expression to retain K terms.

2. We compute the least-squares estimate Z of the regression $AZ = Y$, where $A = (a_{ik})_{k=1,\dots,K;i=1,\dots,n}$ and $Z = (z_i)_{i=1,\dots,K}$.

3. We define $\hat{\beta}_{\text{PCR}}^{(K)} = z_0 + \sum_{i=1}^{K} z_i\psi_k$

Algorithm 5.1 Functional Conjugate Gradient implementation of PLS

```
library("fda")
# Solve the system R x(t) = b(t)
FCG <- function(Afd, bfd, k = 3) {
  # Initial values
  d = bfd
  g = bfd * (-1)
  x = 0 * bfd
  sols <- list()
  for (j in 1:k) {
    Ad = R(Afd, d)
    alpha = - inprod(g, d) / inprod(d, Ad)
    x = x + alpha * d
    sols[[j]] <- x
    g = R(Afd, x) - bfd
    beta = inprod(g, Ad) / inprod(d, Ad)
    d = beta * d - g
  }
  return(sols)
}

FPLS <- function(Xfd, Y, order=3) {
  # Y = \int_t X \beta dt + eps
  # E(Y X(s)) = E (X(s)<X,beta>) = \int_T \int_\Omega X(s)X(t) beta(t)
  b = YX_t(Xfd, Y)
  A = Xfd
  return(FCG(A, b, k = order))
}
```

5.3 Simulations and synthetic data

This section covers the process of evaluation of our model with Monte Carlo methods. These experiments were carried out in R, using the fda package [Ramsay et al., 2018] and the fda.usc [Febrero-Bande and Oviedo de la Fuente, 2012] package. The objective is to provide a wide summary of the performance of the aforementioned techniques in differents cases.

5.3.1 Design of the experiments

The design of an experiment is usually critical, specially when providing feedback about which method is the most suitable depending the situation. Thus, we reproduce the experiments of another authors that previously investigated regularization topics on regression.

We fix the time-domain $\mathbf{T} = [0, 1]$. The experiments are based on the following stochastic process:

$$X(t, \omega) = \sum_{k=1}^{\infty} g_k Z_k(\omega) \psi_k(t), \tag{5.1}$$

where $\{Z_i\}_{i \in \mathbb{N}}$ is a sequence independent and identically distributed normal random variables with mean 0 and variance 1, $\{\psi_i\}_{i \in \mathbb{N}}$ is the sequence of eigenfunctions of R and $\{g_i\}_{i \in \mathbb{N}}$ is the sequence of the square roots of the magnitude of the eigenvalues.

Recalling that our model is $Y_i = \langle X_i, \beta \rangle + \epsilon_i$, we should also define $\{\epsilon_i\}_{i \in \mathbb{N}}$. As usual, this will be Gaussian error of mean zero and variance $\sigma_\epsilon^2 = 0.1$.

Moreover, some practical issues are not yet solved. First, we should truncate the series that defined X to 50 elements. Therefore, we will consider $\{g_i\}_{i \in \mathbb{N}}$ to be decreasing, so that we always take the most significant eigenfunctions. Second, we need to fix a basis for computing the scalar product. For this purpose, we choose a basis of B-splines with 20 elements of order 6, which fits the data well. Lastly, we fix sample sizes n varying from 50 to 200 for training and tests sets. The number of Monte Carlo simulations will be fixed for all experiments to be 500.

For the performance evaluation, we always choose the best parameter among a range of reasonable ones, i.e. we try different number of principal components, differents orders for PLS. To compare them, we choose two metrics:

1. Mean Square Error (MSE) of β

$$\text{MSE} = \left\| \beta - \hat{\beta} \right\|_{L^2(\mathbf{T})}^2 = \int_{\mathbf{T}} \left(\beta - \hat{\beta} \right)^2 dt$$

2. Mean Square Prediction Error (MSPE)

$$\text{MSPE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Note that for the MSPE, the test set must be employed whereas MSE does not depend on it. Next, we present five scenarios where PLS and PCR are compared. Some cases are known to be worst-case scenarios for PCR and we want to know whether PLS has similar issues or outperforms PCR.

5.3.1.1 Scenario 1

The first scenario is based on [Cardot et al., 2003]. We take as eigenfunction $\psi_k = \sqrt{2}\sin((k - 0.5)\pi t)$, eigenvalues $g_k = \frac{1}{(k-0.5)\pi}$ and $\beta(t) = \sqrt{2}\psi_1 + 2\sqrt{2}\psi_2 + \frac{5\sqrt{2}}{2}\psi_3$, i.e. a finite linear combination of the first eigenfunctions.

Table 5.1 shows comparative results. Figure 5.1 displays boxplots for MSE and MSPE for the 500 Monte-Carlo experiments of size $n = 50$. It is clear than for this

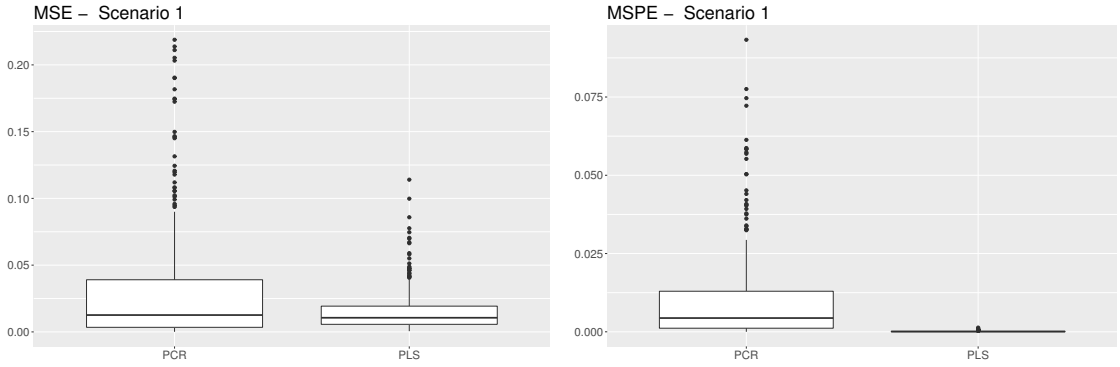

Figure 5.1: Boxplots for MSE and MSPE for 500 simulations of scenario 1

		PCR		PLS	
Metric	n	Mean (std)	Median (mad)	Mean (std)	Median (mad)
MSE	50	$2.719 \cdot 10^{-2}$	$1.243 \cdot 10^{-2}$	$1.474 \cdot 10^{-2}$	$1.125 \cdot 10^{-2}$
		$(1.375 \cdot 10^{-3})$	$(1.673 \cdot 10^{-2})$	$(1.685 \cdot 10^{-4})$	$(9.520 \cdot 10^{-3})$
	100	$1.515 \cdot 10^{-2}$	$6.118 \cdot 10^{-3}$	$6.272 \cdot 10^{-3}$	$4.956 \cdot 10^{-3}$
		$(4.194 \cdot 10^{-4})$	$(8.293 \cdot 10^{-3})$	$(3.061 \cdot 10^{-5})$	$(4.013 \cdot 10^{-3})$
	200	$6.709 \cdot 10^{-3}$	$3.007 \cdot 10^{-3}$	$3.209 \cdot 10^{-3}$	$2.435 \cdot 10^{-3}$
		$(1.102 \cdot 10^{-4})$	$(4.030 \cdot 10^{-3})$	$(8.277 \cdot 10^{-6})$	$(1.906 \cdot 10^{-3})$
MSPE	50	$8.902 \cdot 10^{-3}$	$3.954 \cdot 10^{-3}$	$1.182 \cdot 10^{-4}$	$6.859 \cdot 10^{-5}$
		$(1.644 \cdot 10^{-4})$	$(5.374 \cdot 10^{-3})$	$(1.870 \cdot 10^{-8})$	$(7.215 \cdot 10^{-5})$
	100	$5.082 \cdot 10^{-3}$	$2.038 \cdot 10^{-3}$	$1.128 \cdot 10^{-4}$	$6.045 \cdot 10^{-5}$
		$(4.883 \cdot 10^{-5})$	$(2.690 \cdot 10^{-3})$	$(2.082 \cdot 10^{-8})$	$(6.674 \cdot 10^{-5})$
	200	$2.282 \cdot 10^{-3}$	$1.053 \cdot 10^{-3}$	$1.052 \cdot 10^{-4}$	$4.973 \cdot 10^{-5}$
		$(1.144 \cdot 10^{-5})$	$(1.342 \cdot 10^{-3})$	$(1.817 \cdot 10^{-8})$	$(5.894 \cdot 10^{-5})$

Table 5.1: Mean, standard deviation, median and MAD of MSE and MSPE for scenario 1

simple case, PLS outperforms PCR, as shown in Figure 5.1. Table 5.1 shows the same consequences for different sizes of n.

5.3.1.2 Scenario 2

The second scenario is based again on [Cardot et al., 2003]. We take as eigenfunction $\psi_k = \sqrt{2}\sin((k-0.5)\pi t)$, eigenvalues $g_k = \frac{1}{(k-0.5)\pi}$ and $\beta(t) = \log(1.5t^2+10) + \cos(4\pi t)$, i.e. a infinite linear combination of the eigenfunctions.

Table 5.2 shows comparative results. Figure 5.2 displays boxplots for MSE and MSPE. For this scenario, we observe a surprising phenomenon: although PLS MSE is worse than PCR MSE, the prediction error is better for the PLS estimates. This provides an example of why we consider both metrics at the same time. Besides, the metrics for PCR in both scenarios have several outliers were as PLS metrics seem to be more stable with less noticeable outliers.

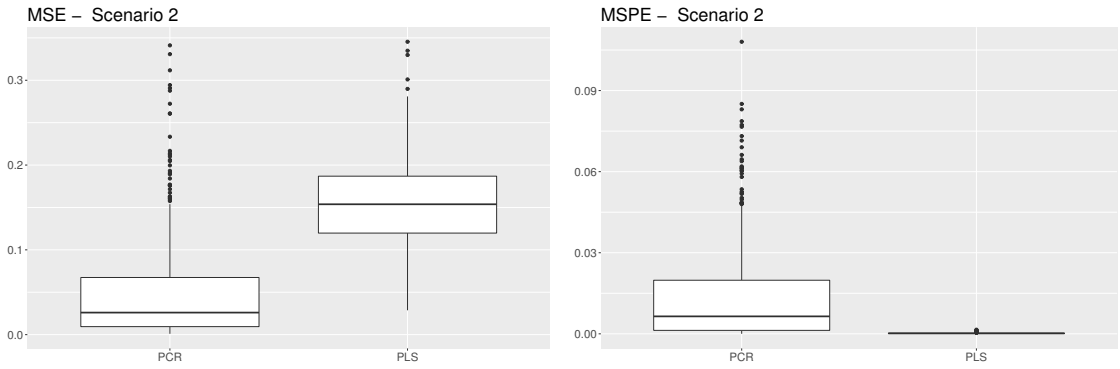

Figure 5.2: Boxplots for MSE and MSPE for 500 simulations of scenario 2

		PCR		PLS	
Metric	n	Mean (std)	Median (mad)	Mean (std)	Median (mad)
MSE	50	$4.084 \cdot 10^{-2}$	$2.150 \cdot 10^{-2}$	$1.576 \cdot 10^{-1}$	$1.542 \cdot 10^{-1}$
		$(2.580 \cdot 10^{-3})$	$(2.190 \cdot 10^{-2})$	$(2.490 \cdot 10^{-3})$	$(5.036 \cdot 10^{-2})$
	100	$2.399 \cdot 10^{-2}$	$1.430 \cdot 10^{-2}$	$1.601 \cdot 10^{-1}$	$1.605 \cdot 10^{-1}$
		$(7.024 \cdot 10^{-4})$	$(1.301 \cdot 10^{-2})$	$(1.724 \cdot 10^{-3})$	$(4.136 \cdot 10^{-2})$
	200	$1.293 \cdot 10^{-2}$	$8.390 \cdot 10^{-3}$	$1.666 \cdot 10^{-1}$	$1.671 \cdot 10^{-1}$
		$(1.536 \cdot 10^{-4})$	$(5.325 \cdot 10^{-3})$	$(1.303 \cdot 10^{-3})$	$(3.857 \cdot 10^{-2})$
MSPE	50	$1.147 \cdot 10^{-2}$	$5.313 \cdot 10^{-3}$	$2.088 \cdot 10^{-4}$	$1.719 \cdot 10^{-4}$
		$(2.601 \cdot 10^{-4})$	$(7.173 \cdot 10^{-3})$	$(2.313 \cdot 10^{-8})$	$(9.825 \cdot 10^{-5})$
	100	$6.157 \cdot 10^{-3}$	$3.100 \cdot 10^{-3}$	$2.059 \cdot 10^{-4}$	$1.635 \cdot 10^{-5}$
		$(6.794 \cdot 10^{-5})$	$(3.977 \cdot 10^{-3})$	$(2.003 \cdot 10^{-8})$	$(8.792 \cdot 10^{-5})$
	200	$2.800 \cdot 10^{-3}$	$1.374 \cdot 10^{-3}$	$2.182 \cdot 10^{-4}$	$1.712 \cdot 10^{-4}$
		$(1.504 \cdot 10^{-5})$	$(1.670 \cdot 10^{-3})$	$(2.263 \cdot 10^{-8})$	$(9.730 \cdot 10^{-5})$

Table 5.2: Mean, standard deviation, median and MAD of MSE and MSPE for scenario 2

5.3.1.3 Scenario 3

The third scenario is based on [Hall and Hosseini-Nasab, 2006]. We take as eigenfunction $\psi_k = \sqrt{2}\cos(k\pi t)$, eigenvalues $g_k = \frac{1}{k}$ and $\beta(t) = \pi^2(t^2 - \frac{1}{3})$, i.e. a infinite linear combination of the eigenfunctions.

Table 5.3 shows comparative results. Figure 5.3 displays boxplots for MSE and MSPE. For this scenario, we observe the reversed effect: in this case PLS has a clearly better MSE but a slightly worse. In contrast to previous case, the difference in the MSPE is not extremely high and both methods perfom in a similar way.

This scenario represents a simple case of next ones, that will deal with cases where eigenvalues are well-spaced or closely spaced and β is a infinite linear combination with decreasing coefficients. It is known that these scenarios are more complicated for principal components techniques and we want to check whether PLS can perform better.

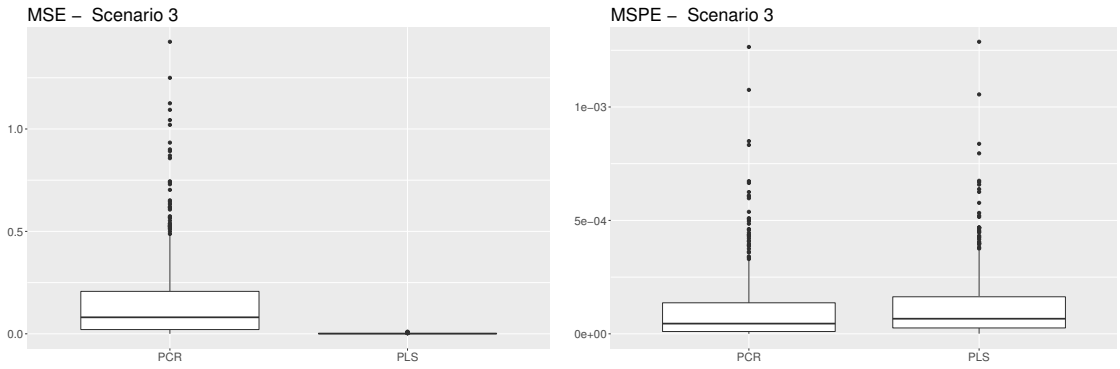

Figure 5.3: Boxplots for MSE and MSPE for 500 simulations of scenario 3

		PCR		PLS	
Metric	n	Mean (std)	Median (mad)	Mean (std)	Median (mad)
MSE	50	$1.627 \cdot 10^{-1}$ $(4.758 \cdot 10^{-2})$	$7.196 \cdot 10^{-2}$ $(1.004 \cdot 10^{-1})$	$1.407 \cdot 10^{-3}$ $(1.790 \cdot 10^{-6})$	$9.867 \cdot 10^{-4}$ $(8.160 \cdot 10^{-4})$
	100	$7.326 \cdot 10^{-2}$ $(8.948 \cdot 10^{-3})$	$3.433 \cdot 10^{-2}$ $(4.714 \cdot 10^{-2})$	$8.770 \cdot 10^{-4}$ $(5.496 \cdot 10^{-7})$	$6.696 \cdot 10^{-4}$ $(5.472 \cdot 10^{-4})$
	200	$3.820 \cdot 10^{-2}$ $(2.844 \cdot 10^{-3})$	$1.703 \cdot 10^{-2}$ $(2.288 \cdot 10^{-2})$	$6.922 \cdot 10^{-4}$ $(2.335 \cdot 10^{-7})$	$5.797 \cdot 10^{-4}$ $(3.932 \cdot 10^{-4})$
MSPE	50	$9.312 \cdot 10^{-5}$ $(1.744 \cdot 10^{-8})$	$5.418 \cdot 10^{-5}$ $(5.418 \cdot 10^{-5})$	$1.121 \cdot 10^{-4}$ $(1.827 \cdot 10^{-8})$	$6.271 \cdot 10^{-5}$ $(6.165 \cdot 10^{-5})$
	100	$9.887 \cdot 10^{-5}$ $(1.794 \cdot 10^{-8})$	$4.740 \cdot 10^{-5}$ $(6.367 \cdot 10^{-5})$	$1.103 \cdot 10^{-4}$ $(1.830 \cdot 10^{-8})$	$6.136 \cdot 10^{-5}$ $(6.842 \cdot 10^{-5})$
	200	$9.437 \cdot 10^{-5}$ $(3.864 \cdot 10^{-8})$	$3.864 \cdot 10^{-5}$ $(5.402 \cdot 10^{-5})$	$1.027 \cdot 10^{-4}$ $(1.685 \cdot 10^{-8})$	$4.709 \cdot 10^{-5}$ $(5.523 \cdot 10^{-5})$

Table 5.3: Mean, standard deviation, median and MAD of MSE and MSPE for scenario 3

5.3.1.4 Scenario 4

The fourth scenario is based on [Hall and Horowitz, 2007]. We take as eigenfunction $\psi_1 = 1$, $\psi_k = \sqrt{2}\cos(k\pi t)$ for $k \geq 2$, eigenvalues $g_k = (-1)^{k+1}\frac{1}{k}$ and $\beta(t) = \sum_{k=1}^{\infty} b_k \psi_k$, where b_k is defined as $b_1 = 0.3$ and $b_k = 4(-1)^{k+1}\frac{1}{k^2}$. This case, compared to next one, has well-spaced eigenvalues, which benefits PCR.

Table 5.4 shows comparative results. Figure 5.4 displays boxplots for MSE and MSPE. For this scenario, we observe that in both metrics PLS outperforms PCR. Nevertheless, next scenario is the most complicated one for PCR since eigenvalues are not well-spaced.

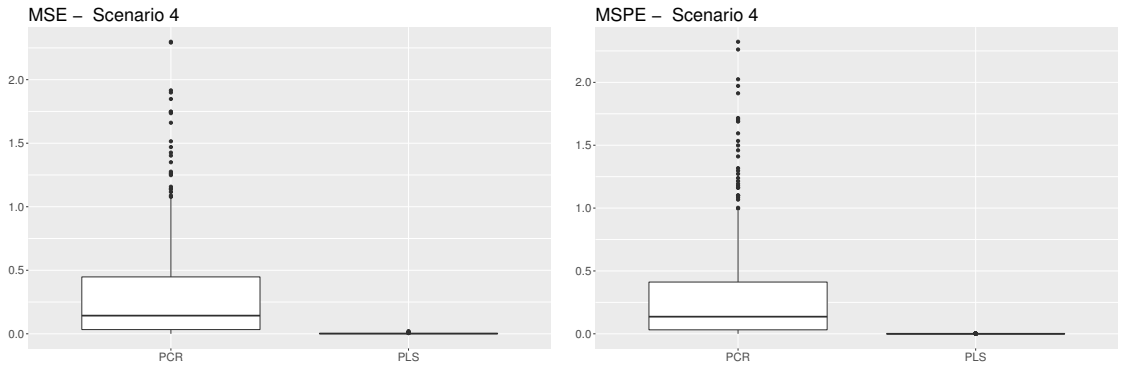

Figure 5.4: Boxplots for MSE and MSPE for 500 simulations of scenario 4

		PCR		PLS	
Metric	n	Mean (std)	Median (mad)	Mean (std)	Median (mad)
MSE	50	$2.757 \cdot 10^{-1}$	$1.186 \cdot 10^{-1}$	$2.281 \cdot 10^{-3}$	$1.561 \cdot 10^{-3}$
		$(1.747 \cdot 10^{-1})$	$(1.642 \cdot 10^{-1})$	$(6.759 \cdot 10^{-6})$	$(1.308 \cdot 10^{-3})$
	100	$1.604 \cdot 10^{-1}$	$7.469 \cdot 10^{-2}$	$1.128 \cdot 10^{-3}$	$8.292 \cdot 10^{-4}$
		$(5.509 \cdot 10^{-2})$	$(1.026 \cdot 10^{-1})$	$(9.005 \cdot 10^{-7})$	$(6.591 \cdot 10^{-4})$
	200	$9.376 \cdot 10^{-2}$	$4.436 \cdot 10^{-2}$	$8.213 \cdot 10^{-4}$	$6.821 \cdot 10^{-4}$
		$(4.436 \cdot 10^{-2})$	$(5.779 \cdot 10^{-2})$	$(3.532 \cdot 10^{-7})$	$(4.740 \cdot 10^{-4})$
MSPE	50	$2.648 \cdot 10^{-1}$	$1.139 \cdot 10^{-1}$	$1.232 \cdot 10^{-4}$	$7.816 \cdot 10^{-5}$
		$(1.497 \cdot 10^{-1})$	$(1.587 \cdot 10^{-1})$	$(1.713 \cdot 10^{-8})$	$(7.511 \cdot 10^{-5})$
	100	$1.596 \cdot 10^{-1}$	$7.085 \cdot 10^{-2}$	$1.069 \cdot 10^{-4}$	$5.855 \cdot 10^{-5}$
		$(5.295 \cdot 10^{-2})$	$(9.817 \cdot 10^{-2})$	$(1.870 \cdot 10^{-8})$	$(6.367 \cdot 10^{-5})$
	200	$9.357 \cdot 10^{-2}$	$4.380 \cdot 10^{-2}$	$1.009 \cdot 10^{-4}$	$4.777 \cdot 10^{-5}$
		$(1.933 \cdot 10^{-2})$	$(5.826 \cdot 10^{-2})$	$(1.708 \cdot 10^{-8})$	$(5.641 \cdot 10^{-5})$

Table 5.4: Mean, standard deviation, median and MAD of MSE and MSPE for scenario 4

5.3.1.5 Scenario 5

The last scenario is based on [Hall and Horowitz, 2007]. We take as eigenfunctions $\psi_1 = 1$, $\psi_k = \sqrt{2}\cos(k\pi t)$ for $k \geq 2$, eigenvalues

$$g_k = \begin{cases} 1 & \text{if } k = 1 \\ 0.2(-1)^{k+1}(1 - 0.0001k) & \text{if } k = 2, 3, 4 \\ 0.2(-1)^{k+1}(\frac{1}{5n} - 0.0001j) & \text{if } k = 5n + j \end{cases}$$

and $\beta(t) = \sum_{k=1}^{\infty} b_k \psi_k$, where b_k is defined as $b_1 = 0.3$ and $b_k = 4(-1)^{k+1}\frac{1}{k^2}$. As mentioned before, this case is built to disadvantage PCR because eigenvalues are not well-spaced.

Table 5.5 shows comparative results. Figure 5.5 displays boxplots for MSE and MSPE. For this scenario, we have the same consequences of scenario 4. In both metrics, PLS surpasses PCR in this case.

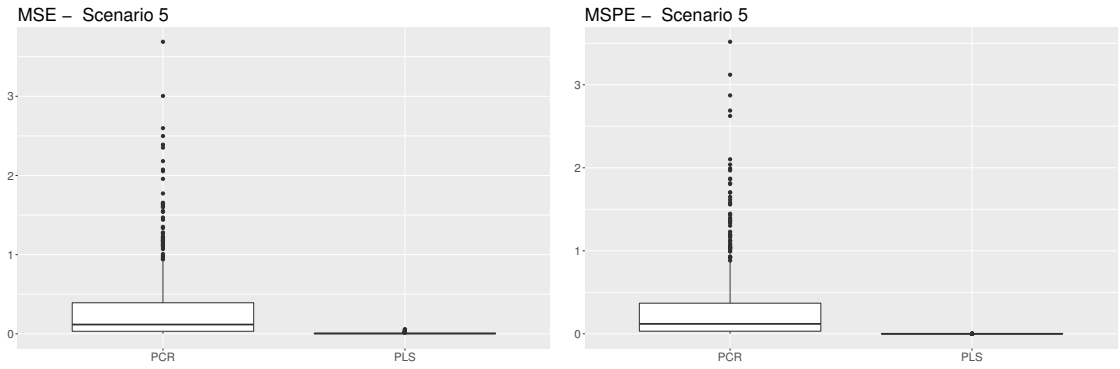

Figure 5.5: Boxplots for MSE and MSPE for 500 simulations of scenario 5

		PCR		PLS	
Metric	n	Mean (std)	Median (mad)	Mean (std)	Median (mad)
MSE	50	$2.972 \cdot 10^{-1}$	$1.411 \cdot 10^{-1}$	$5.948 \cdot 10^{-3}$	$3.541 \cdot 10^{-3}$
		$(1.899 \cdot 10^{-1})$	$(1.922 \cdot 10^{-1})$	$(4.684 \cdot 10^{-5})$	$(3.234 \cdot 10^{-3})$
	100	$1.805 \cdot 10^{-1}$	$9.348 \cdot 10^{-2}$	$3.074 \cdot 10^{-3}$	$1.903 \cdot 10^{-3}$
		$(5.580 \cdot 10^{-2})$	$(1.276 \cdot 10^{-1})$	$(1.518 \cdot 10^{-5})$	$(1.790 \cdot 10^{-3})$
	200	$7.144 \cdot 10^{-2}$	$3.116 \cdot 10^{-2}$	$1.896 \cdot 10^{-3}$	$1.209 \cdot 10^{-3}$
		$(9.241 \cdot 10^{-3})$	$(4.206 \cdot 10^{-2})$	$(5.570 \cdot 10^{-6})$	$(9.816 \cdot 10^{-4})$
MSPE	50	$2.966 \cdot 10^{-1}$	$1.421 \cdot 10^{-1}$	$1.120 \cdot 10^{-4}$	$6.197 \cdot 10^{-5}$
		$(2.233 \cdot 10^{-1})$	$(1.891 \cdot 10^{-1})$	$(2.021 \cdot 10^{-8})$	$(7.183 \cdot 10^{-5})$
	100	$1.760 \cdot 10^{-1}$	$8.777 \cdot 10^{-2}$	$1.040 \cdot 10^{-4}$	$5.022 \cdot 10^{-5}$
		$(5.290 \cdot 10^{-2})$	$(1.211 \cdot 10^{-1})$	$(1.945 \cdot 10^{-8})$	$(6.104 \cdot 10^{-5})$
	200	$7.042 \cdot 10^{-2}$	$3.192 \cdot 10^{-2}$	$1.001 \cdot 10^{-4}$	$4.835 \cdot 10^{-5}$
		$(9.055 \cdot 10^{-3})$	$(4.311 \cdot 10^{-2})$	$(2.071 \cdot 10^{-8})$	$(6.137 \cdot 10^{-5})$

Table 5.5: Mean, standard deviation, median and MAD of MSE and MSPE for scenario 5

5.4 Real data

This second section aims to describe the performance of PLS against its alternatives in real-life environments. Note that the lack of a ground truth makes some metrics such as MSE not computable, therefore, only MSPE will be employ.

5.4.1 Water, Fat and Protein content of meat samples (Tecator)

This dataset is commonly used for regression for FDA, consequently, we will use it as a reference. The dataset is composed of 250 absorbances curves, where the horizontal axis is the wavelength (mm) and the vertical axis is the absorbance. Note that in contrast to what we did in other cases, the horizontal axis does not represent time but wavelength. Anyway, the method is still applicable since it is a continuous function taking values in a continuous interval.

The response in this case is the water, fat and protein contain of the meat sample. This means that the problem we are facing is how to estimate water, fat and protein of a meat sample given its absorbance curve. Usually, the regressor

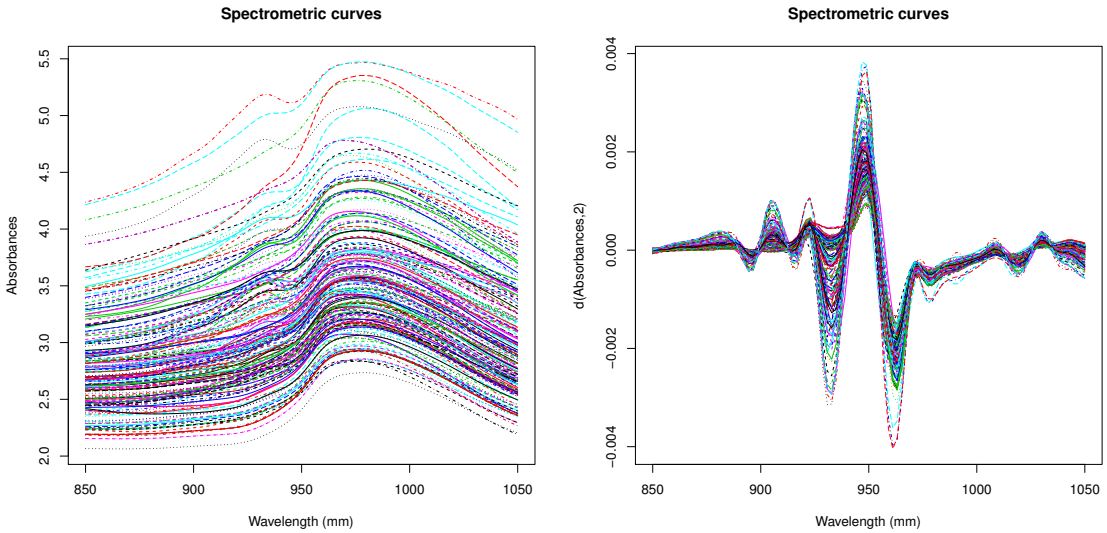

Figure 5.6: Spectometric curves and their second derivative

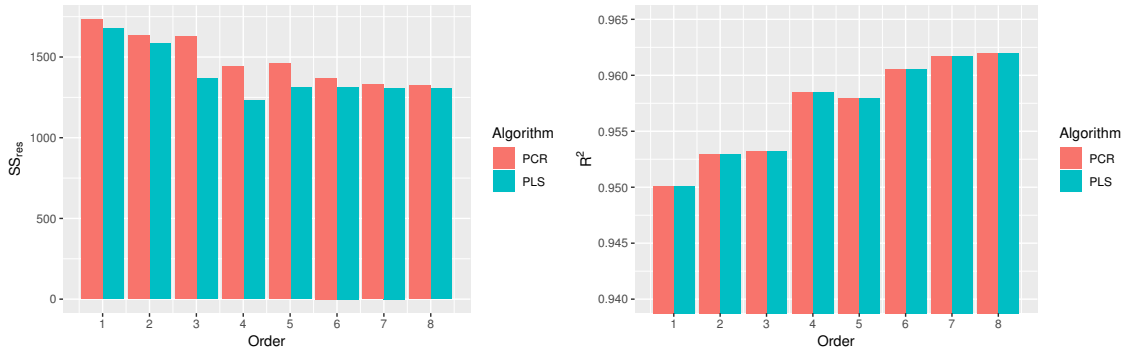

Figure 5.7: Sum of residuals and coefficient of determination of PLS and PCR for tecator dataset

is not the absorbance curve itself but the second derivative of the curves, which is computed using the basis. The shape of the curves can be observe in figure 5.6.

For this experiment, 100 samples of the 250 of the dataset are employed as training set, i.e. we use our algorithms to build $\hat{\beta}$. We use the rest of them as a test dataset, i.e. we use the $\hat{\beta}$ and the curve to compute the estimated response and compare it to the real response.

In contrast to the simulations, we want to see the differences in performance between PCR and PLS varying with the number of components or order. Therefore, figure 5.7 displays the MSPE for PLS and PCR with different order and number of components.

In the light of the results, PLS performs better than PCR and it is noticeable in the sum of the residuals. This difference is less noticeable when the order/number of components increases.

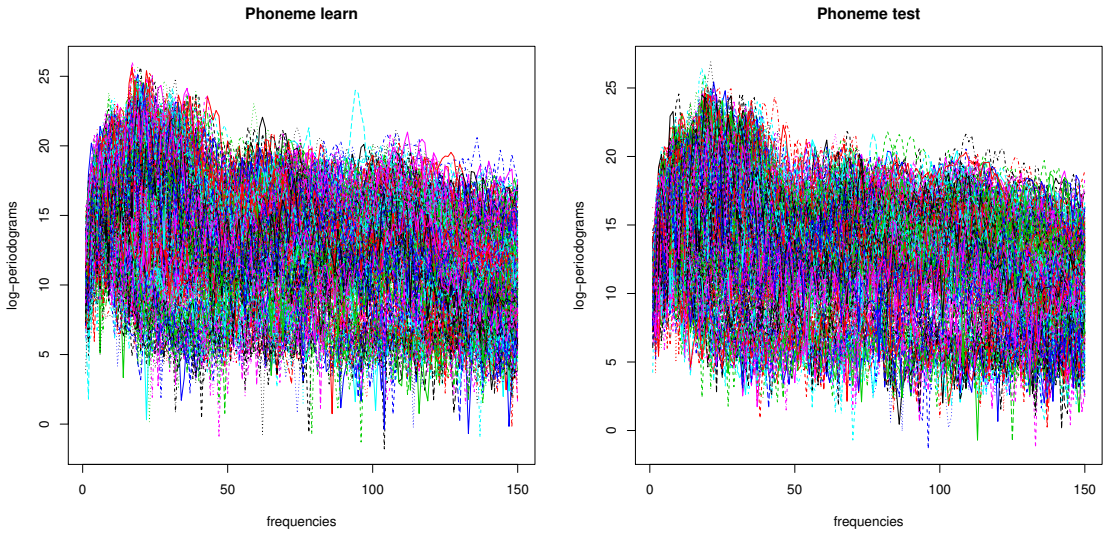

Figure 5.8: Log-periodgrams for both train and test set of curves

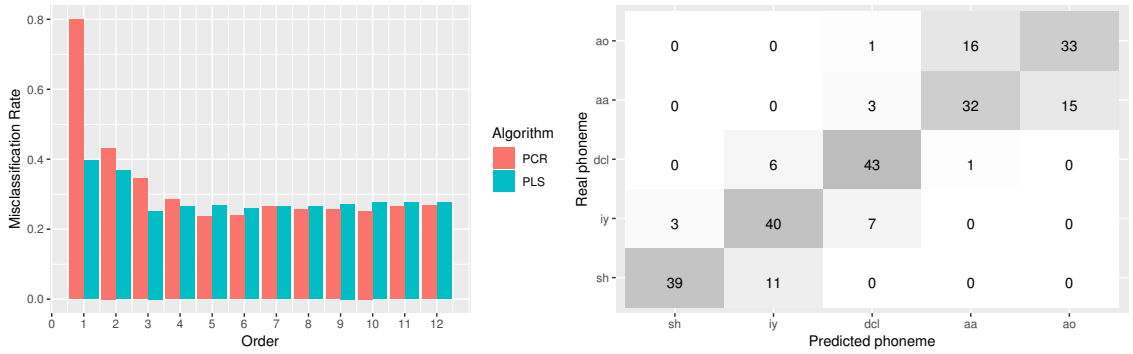

Figure 5.9: Bar plot of misclassification rate for PLS and PCR (left) and confussion matrix (right) for phoneme dataset

5.4.2 Benchmark Phoneme dataset

We consider the phoneme dataset, extracted from [Hastie et al., 2009] and available in this link[1]. Instead of a regression problem, we consider now a classification problem, i.e. the response Y is a categorical variable and takes only five values: ac, aa, dcl, iy and sh. The X curves are not time series but their Fourier transform, i.e. the x-axis represents frequencies. To consider a regression problem instead of a classification problem, we just simplify each class to a number (-2, -1, 0, 1 and 2 respectively) and estimate the class by choosing the closest integer of the regression estimation.

Figure 5.8 shows training and test curves. Figure 5.9 shows a bar plot of misclassification rate depending the order/number of components. As shown, PLS performs much better for low number of components. Nevertheless, PLS and PCR are very similar when the number of components increases. On the right-hand side of Figure 5.9, confussion matrix is displayed for the lowest

[1] https://web.stanford.edu/~hastie/ElemStatLearn/datasets/phoneme.data

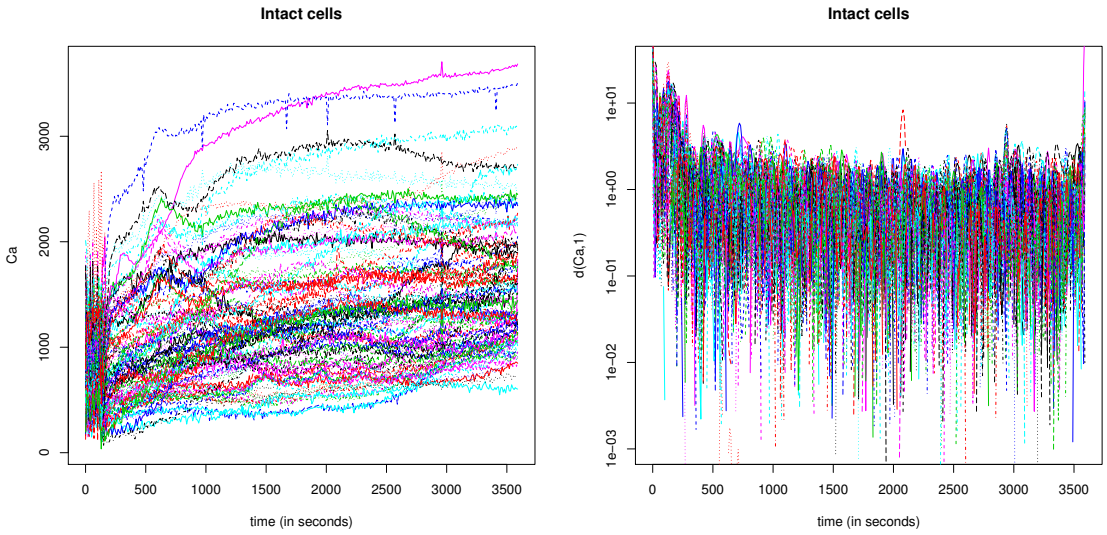

Figure 5.10: Time series of mitochondrial calcium overload and its first derivative

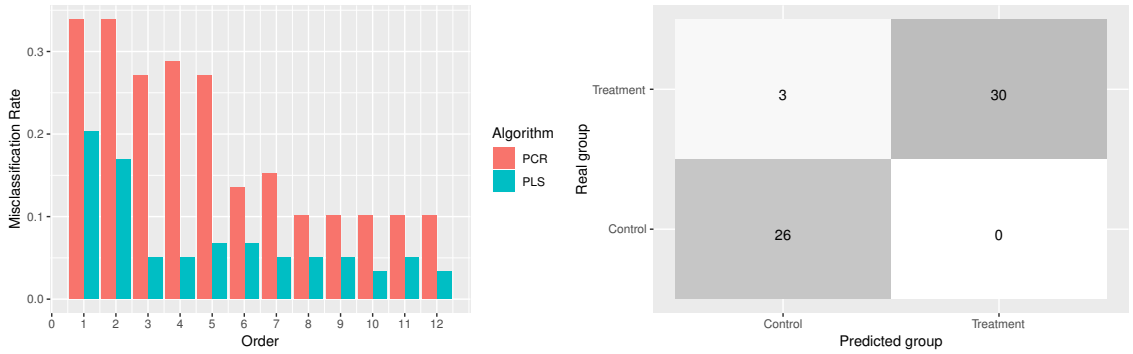

Figure 5.11: Bar plot of misclassification rate for PLS and PCR (left) and confussion matrix (right) for MCO dataset

misclassification rate of PLS. This help us to diagnose where the system performs the worst, that is, when distinguishing between phonemes ac and aa.

5.4.3 Mitochondrial calcium overload (MCO) of control and under-treatment groups

Next, we look at the MCO dataset. It can be found as part of the fda.usc package. We will study curves X that are time series of concentration of mitochondrial calcium in intact cells. The response Y is a binary variable that indicates whether the patient was under-treatment or in the control group, we will identify each group as -1 and 1 for our regression model. Figure 5.10 displays time series of mitochondrial calcium overload and their first derivative. In this case, we use the first derivative as regressor because the accomplished results are better.

Figure 5.11 shows misclassification rate for different order/number of componenets of PLS and PCR. In this case, PLS scores significantly better than PCR. We also show the confussion matrix for the best case of PLS.

Furthermore, we see that PCR is not able to match PLS even for large number of components. This was analyzed and the conclusion is that there are 5-10 curves that PCR is not able to correctly classify. Also, for small number of components such as 3 or 4, we see a huge different between PLS and PCR. As we mentioned before, PLS is a good alternative to PCR for a small number of components.

5.5 Conclusion

This chapter consider five synthetic scenarios and three real datasets. In three (1, 4 and 5) out of five synthetic scenarios, PLS clearly outperformed PCR. In the other two, we see scenario 2 where PLS has a larger MSE but a better MSPE and scenario 3 where PLS performs as good as PCR.

Furthermore, PLS scored better than PCR in the three real datasets. We distinguish two cases: first, PLS performs better than PCR for a low number of components but this difference is not relevant when the number of components increases and, second, PLS performs better than PCR in every case.

Tecator and phoneme datasets are in this first case, where PLS should be consider over PCR when a small number of components is necessary. For a large number of components, this difference does not justify the use of one algorithm over other, but PLS offers no disadvantage over PCR. For MCO, we see that PLS perfoBrms better. This was because PCR can not classify properly 5-10 samples that makes a huge difference for PLS, possibly because this data lies on one of the worst cases of PCR.

CONCLUSION

6

6.1 Summary

In this work, we provided a thorough view of PLS for functional data. First, we have tried to clarify the different approaches that lead to PLS in the finite-dimensional context. Although most of this part is not novel, it is insightful in order to extend PLS techniques for functional data. We examined four different approaches to PLS: least squares minimization with restrictions, conjugate gradient algorithms, SVD decomposition and filter factors and, finally, the statistical definition. Depending on the goals, we may prefer one definition over the rest. For the purpose of this work, we choose the first definition as main criterion since it is a simple definition that allows us to express PLS in the same form of many other approaches such as Ridge regression of PCR. Conveniently, we managed to show the equivalence among the definitions and thus, we showed many different properties of PLS.

For functional data, PLS is a developing topic of the state of the art. Consequently, bibliography are neither complete nor clear for some parts. This means that the challenge relied on adapting the different techniques and showing how they are related. Here, we proposed three definitions: least squares minimization with restrictions, conjugate gradient algorithms, and an adapted statistical approach. We showed the equivalence between the first two ones and we stated some properties of the statistical criterion that lead to partial results about the equivalence. Additionally, we provided a collection of interesting properties of PLS for functional data, such as some links with RKHS theory or convergence when the dimension of the Krylov spaces grows.

To illustrate the usefulness of our algorithm we checked also that there are many situation where it performs better than PCR. We checked out five synthetic scenarios and three real data problems. It is arguable that PLS performs better than PCR in all scenarios, but this chapter showed many examples where PLS was a fair alternative to PCR. For instance, PLS performed generally better

for a low number of components/order. On the other hand, this difference was negligible when the number of components grows. The conclusion of these experiments is that PLS showed no disadvantage over PCR and even we managed to observed some improvements.

6.2 Future work

Although we have deeply analyzed PLS, there are still open topics that can be tackled. First of all, we talked briefly about Ritz values and Kaniel-Paige convergence theory. We proved that the filter factors associated to PLS are $w_k = 1 - \mathscr{R}_q(\lambda_k) = 1 - \prod_{i=1}^{q} \frac{\theta_i^{(q)} - \lambda_k}{\theta_i^{(q)}}$ and Kaniel-Paige theory bounds the error of $\lambda_k - \theta_k^{(q)}$. This would provide a metric of how a principal component is represented in the qth order PLS estimate. This can provides a precise description of how PLS, PCR, and OLS are related and when they are similar.

Furthermore, we did not provide the equivalent of Definition 3.4.7 for functional data. SVD decompositions exists for functional data and similar approaches can be found for PCR. Even, we can guess that filter factor expression will be the aforementioned w_k. Nevertheless, the matrix-based computations associated to the proofs is not immediate. Consequently, we decided to check this definition and its equivalence in future work together with the previous research line.

Moreover, we think that Definition 3.5.1, Definition 4.2.7, and many more variants of them are usually hard to analyze and, usually, authors tend to use alternative definitions such as conjugate gradient or NIPALS. It is a common assumption that the definitions are equivalent, but we are not able to find a formal proof of the equivalence of the many different available flavours of PLS. Indeed, this is also a research challenge to categorize the available variations of PLS and see which of them are equivalent to standard definitions.

Bibliography

[Andersson, 2019] Andersson, M. (2019). A comparison of nine PLS1 algorithms. *Journal of Chemometrics*, 23(10):518–529.

[Cardot et al., 2003] Cardot, H., Ferraty, F., Mas, A., and Sarda, P. (2003). Testing Hypotheses in the Functional Linear Model. *Scandinavian Journal of Statistics*, 30(1):241–255.

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

[de Jong, 1995] de Jong, S. (1995). PLS shrinks. *Journal of Chemometrics*, 9(4):323–326.

[Delaigle and Hall, 2012] Delaigle, A. and Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics*, 40(1):322–352.

[Eldén, 2004] Eldén, L. (2004). Partial least-squares vs. Lanczos bidiagonalization—I: analysis of a projection method for multiple regression. *Computational Statistics & Data Analysis*, 46(1):11–31.

[Febrero-Bande et al., 2017] Febrero-Bande, M., Galeano, P., and 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.

[Febrero-Bande and Oviedo de la Fuente, 2012] Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical Computing in Functional Data Analysis: The {R} Package {fda.usc}. *Journal of Statistical Software*, 51(4):1–28.

[Ferraty and Vieu, 2006] Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis : theory and practice*. Springer.

[Golub and Reinsch, 1970] Golub, G. H. and Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420.

[Hall and Horowitz, 2007] Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91.

[Hall and Hosseini-Nasab, 2006] Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):109–126.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction, 2nd Edition*. Springer series in statistics. Springer.

[Heil, 2018] Heil, C. (2018). *Metrics, norms, inner products, and operator theory*. Birkhäuser Basel - Springer, Georgia, USA.

[Jong, 1993] Jong, S. D. (1993). PLS fits closer than PCR. *Journal of Chemometrics*, 7(6):551–557.

[Krishnan et al., 2011] Krishnan, A., Williams, L. J., McIntosh, A. R., and Abdi, H. (2011). Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage*, 56(2):455–475.

[Lanczos, 1950] Lanczos, C. (1950). An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45(4):255–282.

[Mandel, 1982] Mandel, J. (1982). Use of the Singular Value Decomposition in Regression Analysis. *The American Statistician*, 36(1):15.

[Parlett, 1998] Parlett, B. N. (1998). *The symmetric eigenvalue problem*. Prentice-Hall, Inc.

[Phatak and de Hoog, 2002] Phatak, A. and de Hoog, F. (2002). Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS. *Journal of Chemometrics*, 16(7):361–367.

[Phatak and de Jong, 1997] Phatak, A. and de Jong, S. (1997). The geometry of partial least squares. *Journal of Chemometrics*, 11(4):311–338.

[Preda and Saporta, 2005] Preda, C. and Saporta, G. (2005). PLS regression on a stochastic process. *Computational Statistics & Data Analysis*, 48(1):149–158.

[Ramsay and Silverman, 1997] Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Springer Series in Statistics. Springer New York, New York, NY.

[Ramsay et al., 2018] Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2018). fda: Functional Data Analysis.

[Tenenhaus, 1998] Tenenhaus, M. (1998). *La régression PLS : théorie et pratique*. Ed. Technip.

[Wold, 1966] Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In Krishnaiah, P., editor, *Multivariate Analysis*, pages 391–420. Academic press, New York.

[Wold, 1975] Wold, H. (1975). Path Models with Latent Variables: The NIPALS Approach. *Quantitative Sociology*, pages 307–357.