POLITECNICO DI MILANO

FACOLTÀ DI INGEGNERIA DEI SISTEMI

Corso di Laurea in Ingegneria Matematica

TESI DI LAUREA MAGISTRALE

Data Driven Sobolev Metrics for Functional Data Analysis: an Application to Natural Fibers

Relatore: **Dott. Simone VANTINI**
Correlatore: **Dott. Marco COMPAGNONI**

Candidata: **Marina RIABIZ**

Matricola: 767478

ANNO ACCADEMICO 2012 - 2013

Alla mia famiglia

Ringraziamenti

Al termine di questo lavoro è doveroso ringraziare chi mi ha permesso di arrivare a questo punto. Il primo grazie va a Simove Vantini per avermi dato l'opportunità di fare una tesi sui dati funzionali, per la disponibilità mostrata, nostantante i mille impegni, e per avermi fornito le conoscenze e le intuizioni in più per affrontre i problemi che si presentavano di volta in volta. Ringrazio Marco Compagnoni che, insieme a lui, mi ha mostrato che ci sono matematica e statistica ovunque, anche nei fili di lana, e per avermi concesso i dati da analizzare; contestualmente ringrazio il gruppo GAIA (Geometry - Algebra In Applications) che lavora al progetto di identificazione delle fibre in collaborazione con il DEIB (Dipartimento di Elettronica, Informzione e Bioingegneria) del Politecnico di Milano.

Il più sincero grazie va alla mia famiglia per il costante appoggio che mi ha sempre dato. A Dejan, perché cerca costantemente di insegnarmi che la vita va presa meno sul serio e che vale sempre la pena spendersi per gli altri. A Ines, perch è per me un modello per la forza di volontà e la determinazione che la contraddistinguono e la rendono una persona speciale. Ai nostri genitori: i tripli salti carpiati con avvitamento non sono nulla rispetto agli sforzi che avete fatto e che sareste ancora disposti a fare per darci il meglio. Mamma perchè mi dimostra che bisogna saper sostenere le proprie scelte ed è il vero collante della nostra famiglia, papà per la curiosità naturale verso ogni cosa: è delle poche persone con cui posso parlare di spazi di Sobolev e due minuti dopo saltare all'Eneide. Ogni mio passo in avanti è anche vostro.

Ringrazio gli amici del liceo, con cui sono rimasta in contatto, nonostante la distanza: Bea, Serena e, in particolare, Aurelia, per le lunghe passeggiate sulle Mura.
Un sentito ringraziamento va a Elena perchè senza di lei probabilmente non mi starei laureando in ingegneria matematica. Il suo sostegno, insieme a quello di Nur, è stato uno dei più preziosi in questi anni. Insieme a loro ringrazio le persone che ho conosciuto al Politecnico nelle varie ingegnerie (Enrica, Antonio, Marco, Roberta) e il gruppo di statistica di ingegneria matematica, in particolare Matteo P. e Matteo M., Marco, Mara, Valeria.

Ringrazio anche Carlo e gli amici di mio fratello e mia sorella, che sono diventati un po' anche i miei, in particolare Daniele che mi ha accompagnato ad intermittenza in questi mesi di biblioteca.

Infine ringrazio tutti i famigliari e gli amici sparsi qua là per il mondo, perché ho la certezza di avere sempre qualcuno su cui poter contare.

Marina

Contents

List of Figures

List of Tables

Abstract

This thesis work has a two-fold structure. In the first part we apply recently developed, but consolidated, techniques of functional data analysis to a problem that has been studied since the 1980s: development of an automated system for animal fibers classification, through feature extraction from electron microscope images. Our contribution was in facing this well-known issue thanks to functional principal component analysis, performed on radii of the fibers as functions of the curvilinear abscissa (dataset FIBER). *Scores* variables obtained in this way have a similar meaning to the modules of the Fourier transform of radii. We compare the performance of discriminant analysis carried out on *scores* and on other features (mean radius and standard deviation, modules of Fourier transform and their logarithms), varying the number of groups taken into consideration.
In the second part we face the problem of metric selection in the functional data analysis framework. We consider the possibility of getting a data-driven sparse Sobolev metric, that gives nonzero weights to the most statistically significant derivatives and zero to others. The procedure is a innovative extension of multivariate techniques for feature selection and penalized matrix decomposition. Some corrections are required, due to the not homogeneous nature of data; we make two proposals in this sense, responding to measure unit and normalization needs. The output is a dissimilarity matrix, that enables unsupervised classification. We run hierarchical clustering on two synthetic datasets and on FIBER data.

Keywords: Functional Data, Functional Principal Component Analysis, Fourier transform, Hierarchical Clustering, Discriminant Analysis, Gaussian Mixture Models, Sobolev Metrics, Convex Optimization, Penalized Matrix Decomposition.

Abstract

Questo lavoro di tesi ha una duplice struttura. Nella prima parte applichiamo tecniche sviluppate di recente, ma ormai consolidate, afferenti all'analisi di dati funzionali a una questione che stata studiata fin dagli anni Cinquanta: lo sviluppo di un sistema automatico per la classificazione di fibre animali, tramite l'estrazione di variabili significative da immagini al microscopio elettronico. Il nostro contributo è stato affrontare tale problema grazie all'analisi delle componenti principali funzionali, effettuata sui raggi delle fibre, guardati come funzioni dell'ascissa curvilinea (dataset FIBER). Le variabili di *scores* ottenute in questo modo hanno un significato analogo ai moduli della trasformata di Fourier dei raggi. Andiamo a confrontare la performance dell'analisi discriminate effettuata sugli *scores* e su altre variabili (raggi medi e deviazioni standard, moduli della trasformata di Fourier e loro logaritmi), variando il numero di gruppi presi in considerazione.
Nella seconda parte affrontiamo il problema della selezione della metrica nel contesto dell'analisi dei dati funzionali. Consideriamo la possibilità di ottenere una metrica di Sobolev sparsa e adattiva, che dia pesi non nulli alle derivate statisticamente significative e nulli alle altre. La procedura è un'estensione innovativa di tecniche multivariate per la selezione di variabili e la decomposizione di matrici con vincoli di penalità. Si rendono necessarie alcune correzioni, a causa della natura non omogenea dei dati; facciamo due proposte in tale direzione, per rispondere a esigenze relative a differenze nell'unità di misura e a esigenze di normalizzazione. L'output è una matrice di dissimilarità, che rende possibile la classificazione non supervisionata delle funzioni. In particolare effettuiamo il clustering gerarchico su due dataset sintetici e sui dati FIBER.

Parole Chiave: Dati Funzionali, Analisi delle Componenti Principali Funzionli, trasformata di Fourier, Clustering Gerarchico, Analisi Discriminante, Modelli di Mistura di Gaussiane, Metriche di Sobolev, Ottimizzazione Convessa, Decomposizione di Matrici con Vincoli di Penalità.

Chapter 1

Introduction

Functional Data Analysis (FDA) is a recent, but in quickly growing branch of statistics, that deals with datasets consisting of curves and surfaces, treated as realizations of random functions, i.e. random variables whose image is an infinite-dimensional functional space. This type of data appears today spontaneously in numerous domains of applications, like geophysics (satellite images), econometrics (stock indices), biomechanics (analysis of human movements), chemometrics (spectrometric curves), genetics (microarray data), medicine (electrocardiograms, electroencephalograms magnetic resonance imaging).
Although theoretical studies on infinite-dimensional random variables are dated to the early 1920s, first applications to real data start in the last decade of the same century. This gap is attributable to the deferred technology development. Only recently we succeeded in acquiring measurements almost continuous in time and/or space on the one hand, and to deal with this type of data with appropriate computational facilities on the other. Since the 1990s, research activity on functional data has grown to the point that a number of monographs have been published, concerning both theory and applications (e.g. Ramsay and Silverman 2002; Ramsay and Sylverman 2005; Ferraty and Vieu 2006). As a results of these recent developments, it is not so surprising that a well-known problem in textile industry, such the construction of an automated system for animal fibers classification, has not been analyzed through FDA techniques yet, as far as we know.

The differences in cost between different types of fibers are remarkable and subject to fluctuations related to the market demand and availability. They are also influenced by factors such the climate and the political situation in places of origin of the raw materials.
Identification of the fibers contained in a textile product is a priority in order to guarantee the quality and protection of producers and consumers. Cashmere is a luxury fiber, rare and expensive, that stimulates commercial fraud, just because of its economic

value. Increased competition and the growing demand for this fiber have brought the substitution of cashmere with other animal fibers (mainly wool) to worrying levels. This phenomenon penalizes the image of the most important exporter textile industries, which must be protected from unfair competition.
Current regulation on quality control is described in ISO 17751:2007 normative. It is based on morphological analysis of the fibers with electronic or optical microscope, providing an objective criterion for identification, based on cuticular height and on fibers diameters. Since its first formulation in the mid-80s (see the early works from Robson et al.), attempts have been made in order to define an automatic procedure for implementing fibers identification. Though also techniques of different nature have been recently explored (e.g., DNA analysis, mass spectroscopy) currently morphological analysis remains the most established and taken as a reference. In the literature, various techniques have been explored to get automatic morphological analysis, using several *image processing* procedures, several types of geometric indicators (diameter, shape and size of the scales, or more sophisticated analyses, as wavelet analysis of the texture, see Zhang et al., 2010) and several types of classifiers (e.g., Bayesian classifiers, neural networks).

In the first part of the present work we consider the problem of getting a small classification error about the fibers composing FIBER dataset, consisting of of $n = 894$ fibers gray-scale bidimensional images, belonging to $g = 9$ groups of materials: 5 kinds of cashmere and 4 kinds of wool. Our objective is to build, thanks to functional data techniques, a classifier able to discriminate at least two macro-groups (cashmere and wool). The dataset is quite complex, but we focus only on a single important factor, as the evolution of the radius along the curvilinear abscissa is. In particular, we follow the approach proposed by Ramsay and Silverman (2005), performing functional principal components analysis, that results in *scores* variables having a meaning analogous to modules of the Fourier transform coefficients.
We make some comparisons varying the number of groups and the features (scores, DFT modules and log-modules) used in building a quadratic classifier. We conclude that scores variables, with two groups (cashmere fibers and a mixture model for wool fibers) have the best performance among the various solutions taken into consideration. The percentage of not correctly classified fibers is lower even with respect to the case of using synthetic multivariate indices (longitudinal mean radius and its standard deviation).

In the second part of the thesis we deal with a problem that arises in the functional data framework: the choice of the norm (or semi-norm) used for measuring closure of the data deeply influences the results of statistical procedures, not only in terms of convergence, but especially in the possibility of identifying the real between-functions variability.

The problem is treated in recent works (see Ferraty and Vieu 2006; Ferraty et al. 2010): they stress tha for the purposes of a good statistical analysis, the functional space which data belong to have to be a metric space. But this is not sufficient: it is necessary that the chosen metric correctly represents the variability in the data. This is shown for example by Ferraty and Vieu (2002), that, in a framework of non-parametric functional regression, provide real applications and simulation studies examples of how the use of semi-metrics different from the one induced by the L^2 norm can improving the predictive power of the regression model.

Up to now statisticians are leaved two possibilities: either choosing the metric space *a priori*, depending on what they think best represents the data, or doing analyses in more than one metric space and seeing *a posteriori* which one is better. Both ways have weak points: the firs one is likely to be too subjective, the second one too expansive. A solution could be to make the choice of the metric space to become an integrated passage of statistical analysis, building a semi-metric *adapted* to functional variables (Ferraty et al. 2010), by searching at least in a certain class of metric spaces.

This is precisely the objective of the second part of this work: construction of a Sobolev data-driven semi-metric, where *adaptivity* means identifying which derivatives are more responsible for the between-functions differentiation. In order to do achieve this, we attribute a weighting coefficient to each term of the Sobolev metric (corresponding to a derivative order): such coefficients should be null if there is variability along that derivative order, null otherwise. We propose an optimization procedure aimed at finding a system of weights that meets this requirement, extending the work of Tibshirani and Witten (2010) regarding a multivariate feature selection framework. In particular we extend what they call *sparse hierarchical clustering*, which is indeed a technique that can be applied to any method that takes a dissimilarity matrix as its input. The main novelty that we introduce is in the type of data between which the dissimilarity is computed (functions and their derivatives), and not in the steps of the optimal algorithm used to construct the optimal weights.

Some corrections are necessary due to the not homogeneous nature of our data, and we make two proposals: the first one is concerned with unit measure need, the second with normalization need. We analyze their effect on two synthetic datasets and on FIBER dataset.

In details the thesis is structured as follows.

In Chapter 1, we study FIBER dataset by mens of functional data analyses. We perform functional principal component analysis on the values of the radii of fibers as functios of the curvilinear abscissa. Nine group of materials are present in the dataset. Hierarchical clustering of the Gaussian probability functions, estimated on the data, indicates that we should consider the existence of two groups: cashmere and a mixture of wool. This allows us to perform discriminant analysis on fibers of different materials.

In Chapter 2, we show the general framework for future selection proposed by Tibshirani and Witten (2010) and its extension for our purposes of data-driven Sobolev metric selection; we also discuss the meaning of complementary metrics made available by the method. The outcome are a dissimilarity matrices that enable unsupervised classification.

Chapter 3 contains simulations on two synthetic datasets (polynomial and trigonometric), that reveal that some corrections are necessary to the method proposed in Chapter 3, due to the not-homogeneous nature of data. We make two proposal and discuss them. The chapter is ended by the application to FIBER dataset.

Conclusions and future developments are drawn in Chapter 4, while Appendices report some proofs and details omitted in the central part of the thesis.

All computer analyses were made thanks to statistical software `R`.

Chapter 2

FIBERS Dataset

2.1 Introduction

FIBER dataset is based on bi-dimensional images of $n = 894$ textile fibers, belonging to $g = 9$ groups of materials: 5 kinds of cashmere and 4 kinds of wool. Percentages of the materials don't differ significantly from 1/9, since the sample sizes of the groups are as follows: $(97, 100, 100, 100, 100, 97, 100, 100,\ 100)$, two subgroups having less images, because they could not focus and were not replaced.
Through an electron microscope, a bi-dimensional array of gray-scale pixels is produced for each fiber. An euclidean skeletonisation algorithm provides the two spatial coordinates of the centerline (computed as the set of centers of maximal circumferences that can pointwise be inscribed in the fiber section), its bifurcations and the radius of the maximal inscribed circumference (Figure 2.1). Several scalar indicators are obtained, concerning the global and the internal morphology and the texture of the fibers; many of the indicators are related to the scales of the fibers. Nevertheless in the present work we investigate how to classify the fibers in the framework of *functional data analysis* (FDA), as defined below, focusing only on the evolution of the radius along the curvilinear abscissa.

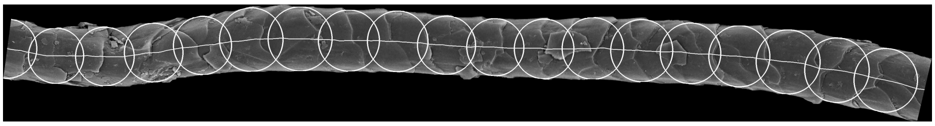


Figure 2.1: Example of a fiber image. The white lines represent the centerline and the circumferences that allow to identify the pointwise radius.

Especially in recent years, in numerous domains of applications appear spontaneously situations in which the collected data are curves or surfaces. This made arising, in modern statistics, the recent by quickly growing branch devoted to functional data analysis, that treats these kind of data as realizations of *functional variables* or random functions. To fix ideas, we can give the following general definition:

Definition 1. A random variable $\mathcal{X}(\omega)$ is called *functional variable* if it takes values in an infinite dimensional, or functional, space $(E, \mathcal{E})$. An observation χ of $\mathcal{X}(\omega)$ is called a *functional data.* Here $\mathcal{E}$ is is an appropriate σ-algebra on E.

Definition 1 is generic, in order to take into account the fact that, as anticipated, the random variable $\mathcal{X}(\omega)$ can belong to a wide class of objects: real functions, surfaces or n-dimensional (with $n \geq 2, n \in \mathbb{N}$) vector of functions. In the present work we deal with random one-dimensional curves, thus $\mathcal{X}(\omega) = \{\mathcal{X}(\omega, s) : s \in \mathcal{T}, \mathcal{T} \in \mathbb{R}\}$ and consequently $\chi = \{\chi(s) : s \in \mathcal{T}, \mathcal{T} \in \mathbb{R}\}$.

2.2 Functional Principal Component Analysis

In the rest of this chapter we assume to deal with random functions that take value in $(L^2(a,b), \mathcal{E})$: with respect to Definition 1 we have $E = L^2$, $\mathcal{T} = (a,b)$, $\mathcal{E}$ an opportune σ-algebra, for example Lebesgue σ-algebra.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathbb{P}(d\omega)$ be the Lebesgue measure. The ith fiber is represented by the function

$$\begin{array}{rcl} \mathbf{f}_i : S_i \subset \mathbb{R} & \longrightarrow & \mathbb{R}^3 \\ s & \longmapsto & \mathbf{f}_i(s) = (x_i(s), y_i(s), r_i(s))\,, \qquad 1 \leq i \leq n \end{array}$$

where each $r_i(s)$ is the radius of the maximal inscribed circumference centered in $(x_i(s), y_i(s))$ and is a realization of one of the following nine random functions (corresponding to the nine groups)

$$R_h(\omega, s) : \Omega \longrightarrow L^2(a_h, b_h), \qquad 1 \leq h \leq g.$$

When not ambiguous we will denote by $R_h(s)$ the random function $R_h(\omega, s)$, avoiding to express the dependence on the case ω. The abscissa parameter s measures the distance along the fiber, from the first section to the last (after acquisition, images are cut to have terminal sections orthogonal to the centerline). Functions $x_i(s)$ and $y_i(s)$ map s into the left-right and up-down coordinates of the corresponding point of the centerline. The nine random functions are supposed to have value in $L^2(a_h, b_h)$, thus we can analyze their mean function and their covariance function:

$$\begin{aligned} \mu_h(s) &= E\,[\,(R_h(s)\,] = \int_\Omega R_h(\omega, s)\mathbb{P}(d\omega) &\quad (2.1) \\ \Sigma_h(t,s) &= E\,[\,(R_h(t) - \mu_h(t))\,(R_h(s) - \mu_h(s))\,]\,, \qquad 1 \leq h \leq g. &\quad (2.2) \end{aligned}$$

Analogous quantities ($\mu(s)$ and $\Sigma(t,s)$) can be defined for a unique random function ($R(s)$), representing all the data, without distinction of groups (the choice of the number of groups is faced in Section 2.3).

As a matter of fact, the reconstruction algorithm provides centerlines and radius profiles only on a fine grid of points; the number of points available for each fiber ranges from 1686 to 4003, and is almost perfectly correlated to the approximate length of the reconstructed centerlines, which in turn varies from 0.17 mm to 0.46 mm. Due to the necessity of managing with valuations of the radius we interpolate data. Before interpolating and in order to have nearly the same interpolation step, we decide to cut the images, symmetrically from the center, adapting the length of their curvilinear abscissa to that of the image with the shortest length of curvilinear abscissa.

Let $\mathbf{R}$=$[\mathbf{r_1}'(s), \dots, \mathbf{r_n}'(s)]$, $1 \leq s \leq p$, denote the $n \times p$ data matrix obtained, with $n = 894$ observation and p = 1706 features, representing the value of the radius along the curvilinear abscissa, shown in Figure 2.2. Looking at the sample mean curves (in black), we can see that the cashmere groups have a mean value beneath 200 μm, while the wool groups have a mean value of about 200 μm, with the exception of wool_3, that could be better discriminated. In Section 2.3 we will show how the wool group can be treated as a mixture of wool_3 and all the the other kind of wool. Note that in the implementation we deal with the sample mean $\hat{\mu} = \overline{\mathbf{R}} = \dfrac{1}{n}\sum_{i=1}^{n} \mathbf{r_i} \in \mathbb{R}^p$, and the sample unbiased covariance matrix $\hat{\mathbf{\Sigma}} \in \mathbb{R}^{p \times p}$, $\hat{\mathbf{\Sigma}}(t,s) = \frac{1}{n-1}\sum_{i=1}^{n}\left[(\mathbf{r_i}(t) - \overline{R}(t))(\mathbf{r_i}(s) - \overline{R}(s))\right]$, $1 \leq s,\, t \leq p$. We could similarly define $\overline{\mathbf{R}}_\mathbf{h}$ and $\hat{\mathbf{\Sigma}}_\mathbf{h}$ for sample estimations in every group.

We now proceed with the the functional principal component analysis (FPCA), as described for example in Ramsay and Silverman (2005). The overall mean $\mu(s)$ is subtracted from each function $r_i(s)$ in order to center the coordinate system in zero. An analysis based on the autocovariance is preferred to the alternative analysis based on the autocorralation function because the values of the functions are homogeneous. Given the random function $R(s)$ and the deterministic function $\phi(s) \in L^2(a,b)$, the inner product $\langle \phi, R{-}\mu \rangle = \int_a^b \phi(s)(R(s){-}\mu(s))ds$ is a random function, representing the projection of $(R(s) - \mu(s))$ on $\phi(s)$, with mean function $E[\langle \phi, R - \mu \rangle] = \langle \phi, \mu - \mu \rangle = 0$ and covariance function $Var[\langle \phi, R - \mu \rangle] = \langle \phi, V_\Sigma \phi \rangle$, where V_Σ is defined as the *covariance operator*

$$\begin{aligned} V_\Sigma : L^2(a,b) &\longrightarrow L^2(a,b) \\ \phi &\longmapsto V_\Sigma(t) = \int_a^b \Sigma(t,s)\phi(s)ds. \end{aligned}$$

Details are given in the Appendix A. Note that $Var[\langle \phi, R - \mu \rangle] = Var[\langle \phi, R \rangle]$. FPCA aims then at finding a sequence of orthonormal functions $\{\phi_k(s)\}_{k=1}^{+\infty}$, called *loadings*,

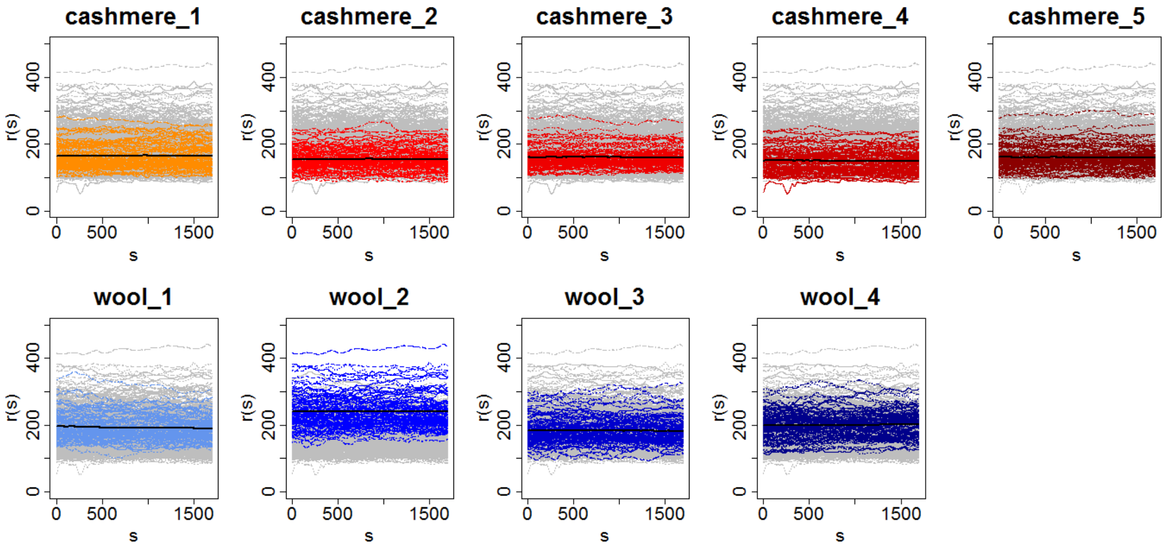


Figure 2.2: Radius of the 9 groups along the curvilinear abscissa. The black line represent the sample mean curves.

that capture the highest quantity of variance of the random function $\langle \phi, R \rangle$ in the residual orthogonal subspace at each iteration. In particular:

- the first loading is the solution of

$$\phi_1 = \underset{\substack{\phi \in L^2(a,b) \\ \|\phi\|=1}}{\operatorname{argmax}} Var[\langle \phi, R \rangle]; \tag{2.3}$$

- the k-th loading is the solution of

$$\phi_k = \underset{\substack{\phi \in L^2(a,b) \\ \|\phi\|=1 \\ \phi_k \perp \phi_i, i=1,\dots,k-1}}{\operatorname{argmax}} Var[\langle \phi, R \rangle]. \tag{2.4}$$

It can be demonstrated that, if we write the spectral decomposition of the covariance function as $\Sigma(t,s) = \sum_{k=1}^{+\infty} \lambda_k \phi_k(s) \phi_k(t)$, the eigenfuctions $\phi_k(s)$ corresponding to eigenvalues λ_k decreasing in module, according to the eigenequation $V_\Sigma \phi_k(s) = \lambda_k \phi_k(s)$, are the solution of the maximization problem above. The maximum values of the variances are realized by the eigenvalues. This result is an extension to infinite-dimensional framework of a well known fact in multivariate analysis; the proof can be found in Ramsay and Silverman (2005), Chapter 8.

Finally by *scores* or *principal components* we mean the projections of the random function $R(s) - \mu(s)$ on the eigenfunctions $\phi_k(s)$: $C_k = \langle R - \mu, \phi_k \rangle = \int_S (R(s) - \mu(s)) \phi_k ds$. The following relation holds:

$$R(s) = \mu + \sum_{k=1}^{+\infty} C_k \phi_k(s). \tag{2.5}$$

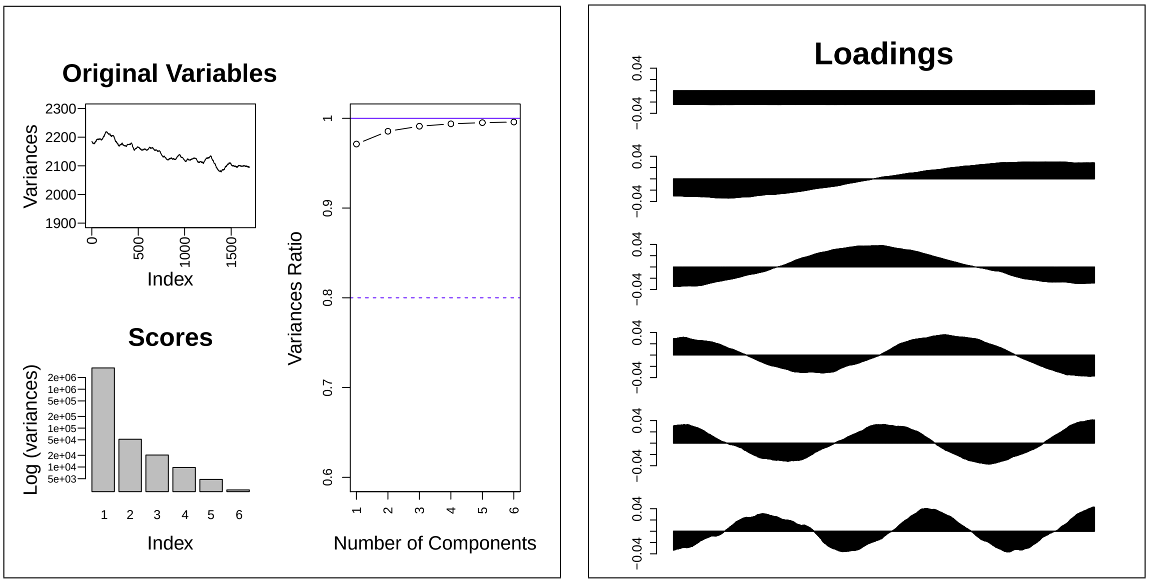


Figure 2.3: Estimate (left) of the variances of the radius along the curvilinear abscissa $R(s)$ (top-left), estimate of the variances of the first 6 scores $C_k, k = 1, \dots, 6$ in logarithmic scale (bottom-left) and estimate of the variances ratio (right). Estimate (right) of the first 6 eigenfuctions $\phi_k(s), k = 1, \dots, 6$.

Relation (2.5) is by itself a dimensional reduction, since let us pass from a non numerable infinity of random values ($R(s)$) to a numerable infinity (C_k). A further reduction is performed if the summation is truncated to m values: $\tilde{R}(s) = \mu + \sum_{k=1}^{m} C_k \phi_k(s)$. In order to chose m, we make the ratio between the explained variance and the total variance under a fixed threshold $B \in [0, 1]$, usually $B = 0.8$:

$$\frac{E[||\tilde{R} - \mu||^2]}{E[||R - \mu||^2]} = \frac{\sum_{k=1}^{m} \lambda_k}{\sum_{k=1}^{+\infty} \lambda_k} \leq B. \tag{2.6}$$

Details on the first equality of equation (A.1) are explained in the Appendix A.

In the computer implementation we can estimate only p eigenvectors $\hat{\phi}_{\mathbf{k}} \in \mathbb{R}^p$ and p eigeinvalues $\hat{\lambda}_k$, only $n-1$ of which are nonzero, since $rank(\hat{\mathbf{\Sigma}}) = \min\{n-1, p\} = n-1$. For each of the n vectors $\mathbf{r_i}(s)$, a vector of p scores $\hat{\mathbf{C}}_{\mathbf{i}}$, is calculated.
Figure 2.3 on the left shows, $\forall 1 \leq s \leq p$, $\hat{\Sigma}(s, s)$ (the estimate of the variances of the values of the radius along the curvilinear abscissa), the logarithm of the variances of the first six scores and the variances ratio in equation (A.1). It suggests a reduction only to the first score, since it explains 97.13 % of the total variance; this would lead to not multivariate analysis, thus in the present work we don't use FPCA just to perform dimensional reduction, but to interpret the meaning of the transformation of the variables that capture the highest quantity of residual variance, with the aim of building a classifier with them.
This is done looking at the eigenfunctions in the right part of Figure 2.3: they are clearly

sinusoidal functions, with decreasing period (increasing frequency). The estimate of the first loading $\hat{\phi}_1(s)$ indicates that, for each function $r_i(s)$, the first score C_1 represents its mean along the curvilinear abscissa: since $\hat{\phi}_1(s)$ is negative, high positive values are associated with narrower fibers, high negative values are associated with wider fibers. The second score C_2, corresponding to the second loading $\hat{\phi}_2(s)$, quantifies the tapering effect: higher (positive or negative) values are associated with more tapered fibers (towards one of the two extremities), lower values are associated with less tapered fibers. The third score C_3, corresponding to the third loading $\hat{\phi}_3(s)$, quantifies the narrowing or widening toward the center effect (depending on the sign of the score): higher values are associated with more deformed toward the center fibers, lower values are associated with less deformed toward the center fibers.
We can generalize, interpreting the principal components as the modules of the Discrete Fourier Transform of each fiber, shortly explained in Subsection 2.2.1; from now on, our analyses will be based on Fourier-transformed variables (more precisely on modules of their Fourier transform) and comparisons with *scores* variables will be made in Section 2.4.
The choice of the number of Fourier frequencies and associated modules taken into consideration will be explained in Section 2.4, where we perform discriminant analysis based on a Gaussian Mixture Model (GMM).

2.2.1 Discrete Fourier Transform

In this Subsection we recall the notation of the Discrete Fourier Transform (DFT) and make some considerations and comparisons with the FPCA. Typically, in functional data analysis, DFT is faced as one of the so called *not data driven* techniques for dimensional reduction, but, on the contrary, in our case FPCA, a *data driven* technique, led us toward it.

We speak of *Fourier series* of a function if we mean to expand quantities depending from a continuous variable (in our case the random function $R(s)$); its analogous for discrete samples is the DFT. Like in FPCA, the aim is to project the random function $R(s)$ (and its realizations $r_i(s)$) on a $L^2(a, b)$-dense basis $\{\phi_k(s)\}_{k=1}^{+\infty}$, and then decide to truncate the projection to a certain number of basis components, performing a reduction that will allow multidimensional standard analyses. Since the mean $\mu(s)$ of the random function $R(s)$ is related to the the frequency $f = 0$, now we don't subtract it.
In the Fourier series, a relation similar to that in equation (2.5) holds, if we take the

following orthonormal deterministic basis, called *Fourier basis*:

$$\phi_0(s) = \frac{1}{\sqrt{b-a}}, \quad \phi_{2k}(s) = \frac{\cos\left(\dfrac{2\pi k(s-a)}{b-a}\right)}{\sqrt{(b-a)/2}}, \quad \phi_{2k-1}(s) = \frac{\sin\left(\dfrac{2\pi k(s-a)}{b-a}\right)}{\sqrt{(b-a)/2}},$$

$s \in (a,b), 1 \leq k \leq \infty$.
Then we call *Fourier coefficients* the random coefficients $C_k = \langle R(s), \phi_k(s)\rangle$; note that $C_0\ \phi_0(s) = \mu(s)$.
At implementation level, passing from the Fourier series to the DFT requires some technicalities, such as the usage of the Eulero's ralation for complex numbers, truncating the expansion to the maximum number of available samples for each function (p), and introducing a discrete scalar product to define the *discrete Fourier coefficients*. This allows writing each of the p components of the sample realizations $\mathbf{r_i}$, $1 \leq i \leq n$, in the following way:

$$r_i(s) = \sum_{k=1}^{p} \hat{C}_i(k) e^{j\frac{2\pi}{p}(k-1)(s-1)}, \qquad 1 \leq s \leq p, \tag{2.7}$$

where

$$\hat{C}_i(k) = \frac{1}{p}\sum_{s=1}^{p} r_i(s) e^{-j\frac{2\pi}{p}(k-1)(s-1)}, \qquad 1 \leq k \leq p \tag{2.8}$$

are the discrete Fourier coefficients. The transformation $\{\mathbf{r_i}\} \mapsto \{\hat{\mathbf{C}}_\mathbf{i}\}, 1 \leq i \leq n$, described in equation (2.8), is the *Discrete Fourier Transform (DFT)*. It provides a matrix $\hat{\mathbf{C}}$, of the same dimensions of $\mathbf{R}$, that is calculated with the help of algorithms that reduce the computational complexity (*Fast Fourier Transform*). The elements of the matrix $\hat{\mathbf{C}}$ are the modules of the k-th complex term of the summation in equation (2.7), relative to k-th frequency $2\pi k/p$. Figure 2.4 shows the modules corresponding to the first two frequencies, divided in the nine groups of materials.

A graphic of variances similar to that on the bottom-left in Figure 2.3 would show that the variances of the variables C_k are not monotonically decreasing, even if they present a global decreasing trend. We point out that a high variance in the k-th feature C_k is not directly connected with the possibility of reducing the classification error if we use it in the construction of the classifier, but this is a first approach to feature selection problem. Moreover we will show that in our case-study this heuristic approach perform quite well. In Section 2.4 we see that Fourier-modules related to the first (low) frequencies and a group of modules related to middle-high frequencies are significant, while some central frequencies and the highest one should not be taken into account.

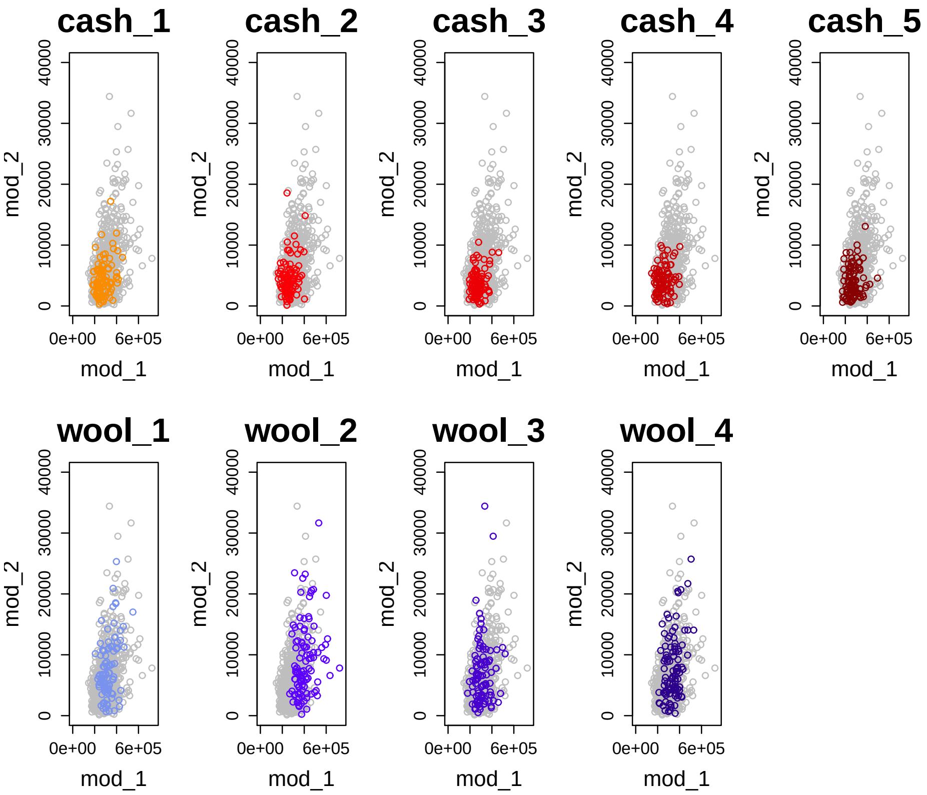


Figure 2.4: Modules $\hat{C}_k, k = 1, 2$, corresponding to the first two frequencies in the 9 groups.

2.3 Hierarchical Clustering

In this Section we support the decision of considering only two groups of materials (wool and cashmere) in the discriminant analysis performed in Section 2.15. The choice of the number of groups deeply influences the entity of the classification actual error rate (AER), i.e. the probability for a new case to be misclassified. For example the trivial classifier, the one that assign to all the samples the class of the group that has the higher prior density (corresponding to higher size if we estimate the priors with the sample information), would approximately get AER $= 4/9 \simeq 0.44$ if we consider only two groups, and AER $= 8/9 \simeq 0.89$ if we consider nine groups.

In order to perform the next analyses, a multivariate normality assumption on the distributions of the features variables must be verified. In fact we make use of densities estimated on data and follow the frequent approach of using the Gaussian model. We denote with g the number of groups (initially $g = 9$), to which correspond a label parameter $L \in \{1, \ldots, g\}$, and with $\{C_{hk}\}_{k=0}^{+\infty}$, $1 \leq h \leq g$, the sequence of Fourier coefficients relative to the g-th group. By means of multivariate Shapiro tests on the first 4 features, reported in Table 2.1 we are induced to transform the sequences $\{C_{hk}\}_k$ into the new sequences $\{X_{hk}\}_k$, where $X_{hk} = \log(C_{hk})$, since sets of m variables drawn

Table 2.1: P-values of multivariate (4 features) Shapiro-tests based on 2500 Monte Carlo iterations, relative to the 5 Cashmere groups and to the 4 Wool groups. The high sample size cause the tests to be extremely powerful; the transformed variables $\{X_{hk}\}_k$ better meet the Gaussian hypothesis.

	C_1	C_2	C_3	C_4	C_5	W_1	W_2	W_3	W_4
$\{C_{hk}\}_k$	0	0	8×10^{-4}	0	76×10^{-4}	0.01	0	0	56×10^{-4}
$\{X_{hk}\}_k$	0.66	0	0	16×10^{-4}	0	0	0.01	0.39	76×10^{-4}

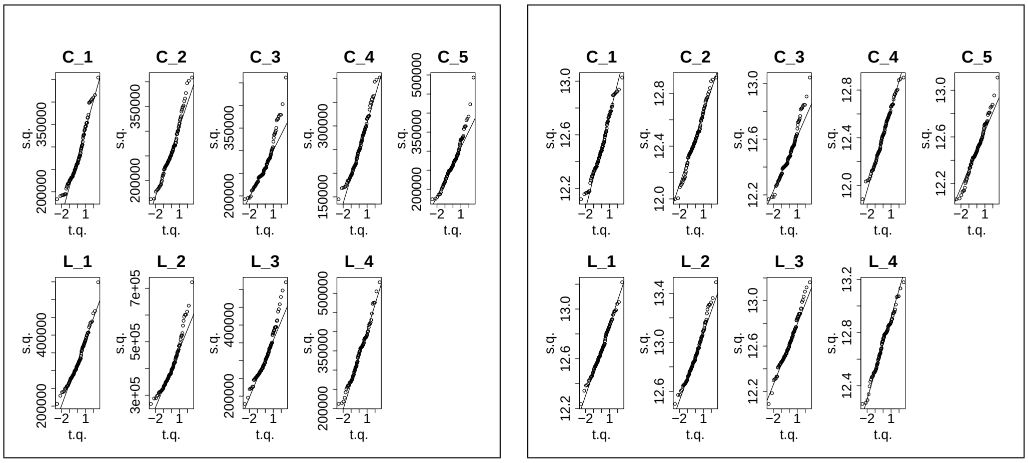


Figure 2.5: QQplot of Fourier-modules C_{h1} (left) and of their logarithms $X_{h1} = \log(C_{h1})$ (right)

from $\{X_{hk}\}_k$ better meet the normality hypothesis. This is confirmed by qqplot of the first variable in each group, shown in Figure 2.5.

Thus we perform a hierarchical clustering (HC) using as data $f_h(\mathbf{x}|\mu_\mathbf{h}, \mathbf{\Sigma_h}) = f_h(\mathbf{x}|L = h)$, the Guassian probability density function of the random vector $\mathbf{X_h} = (X_{h1}, \ldots, X_{hm})$, whose components are the logarithms of the m Fourier coefficients that we find useful for our analyses. In the computer implementation we use $m = 1$ in order to avoid multidimensional integration, being aware that an increase in the number of features taken into account should cause an increase in the number of stable cluster found by the algorithm.
The second input necessary to HC is a dissimilarity matrix between the densities:

$$\mathcal{D} = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1g} \\ d_{21} & 0 & \cdots & d_{2g} \\ \vdots & \vdots & \ddots & \vdots \\ d_{g1} & d_{g2} & \cdots & 0 \end{pmatrix},$$

$$d_{ij} = d(f(\mathbf{x}|L = i), f(\mathbf{x}|L = j)), \quad 1 \leq i, j \leq g.$$

We recall that, if $\mathcal{G}$ is the set of absolutely continuous distributions on $\mathbb{R}^m$, $f(\mathbf{x}), g(\mathbf{x}) \in \mathcal{G}$, $\mathbf{x} \in \mathbb{R}^m$, then the *dissimilarity between* $f(\mathbf{x})$ *and* $g(\mathbf{x})$ is a positive function $d(f, g) : \mathcal{G} \times \mathcal{G} \longrightarrow [0, +\infty)$ such that $d(f, f) = 0$ and $d(f, g) = d(g, f)$. In the present work we used a symmetrized version (adding the second term in equation (2.9)) of the frequently used Kullback-Leibler divergence (see Dykstra 2005) between two distributions f and g:

$$d^{KB}(f,g) = \int_{\mathbb{R}^m} f(\mathbf{x}) \log\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) d\mathbf{x} + \int_{\mathbb{R}^m} g(\mathbf{x}) \log\left(\frac{g(\mathbf{x})}{f(\mathbf{x})}\right) d\mathbf{x}. \tag{2.9}$$

Finally we need to select a *dissimilarity between clusters*. Given the sets $U, V \subseteq \mathcal{X}$, and denoting with $\#U$ the cardinality of set U, some typical choices are:

1. *single linkage* $d_{SL}(U,V) = \min\{d(f,g),\ f \in U,\ g \in V\}$;
2. *complete linkage* $d_{CL}(U,V) = \max\{d(f,g),\ f \in U,\ g \in V\}$;
3. *average linkage* $d_{AL}(U,V) = \dfrac{1}{\#U\#V} \sum_{f\in U} \sum_{g \in V} d(f,g)$.

Computations are performed for all of the three methods and then they are compared.

Algorithm 1: Hierarchical Clustering

```
Data: $\mathcal{D}$, dissimilarity between sets
Result: dendrogram grouping data from $g = 9$ to $g = 1$ clusters, in function of
        the dissimilarity level
initialization: $\mathcal{D}_0 \leftarrow \mathcal{D}$, $g \leftarrow 9$ clusters;
while $g \neq 1$ do
    the two less dissimilar clusters;
    calculate the new matrix $\mathcal{D}_i$;
    $g \leftarrow g - 1$;
end
```

HC algorithm is summarized in pseudo-code (1), and the dendrograms obtained by the three methods are proposed in Figure 2.6. The procedure leave us the task of deciding at what level to cut the tree: a long vertical line means that the for a wide range of dissimilarities the current grouping is stable. A blue box is plotted in order to identify which densities are grouped together if we decide to form two clusters, while a red box is plotted in case we decide to form three clusters. Single and average linkage methods give a similar solution: they point out that the group W_2, associated to label $L = 7$, is a wool very different both from other wool and cashmere groups. In the first case this is due to the fact that SL tends to aggregate nearby elements, despite they belong to different groups.

The classical way to evaluate the goodness of the linkage method is to compare the original dissimilarity-matrix with the matrices containing the dissimilarity level at which that method aggregates elements $f(\mathbf{x}|L = i)$ and $f(\mathbf{x}|L = j)$; the more they

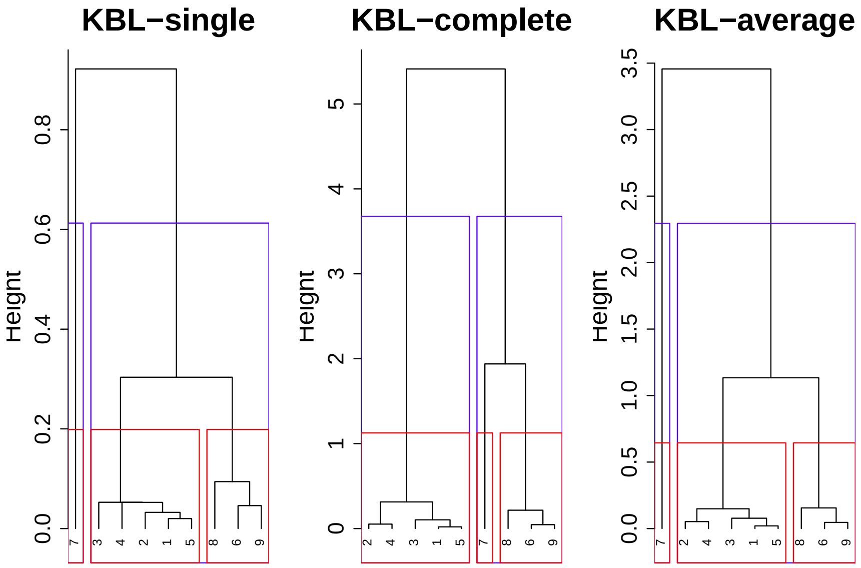


Figure 2.6: Histogram obtained according to the three between-clusters dissimilarities methods.

resemble, the more reliable is the result. Figure 2.7 shows a matrix-plot: dark colors suggest low dissimilarities, while light colors suggest high ones. The visual impression that matrices produced by SL and CL resemble more the original one is quantified by the cophenetic correlation coefficient $\rho_{coph} = Corr(\mathcal{D}, \mathcal{D}_m), m \in \{SL, CL, AL\}$. The synthetic notation $Corr(\mathcal{D}, \mathcal{D}_m)$ indicates the calculus of the correlation coefficient after vectorizing the matrices following the same order. Table 2.2(a) gives the three values.

In conclusion we would be induced to consider the existence of three group in order to be able to separate all the wool from the cashmere groups. But not to throw away the information on the fact that W_7 is a wool, even if more distinguishable from cashmere, we treat the whole wool group as a Gaussian Mixture Model (GMM) of the subgroups $W_A = W_{-7}$ and $W_B = W_7$ (W_{-7} are the not-W_7 groups), with weights given by prior probabilities ($p_A \simeq 3/4$, $p_B \simeq 1/4$). Thus the wool likelihood is:

$$f(\mathbf{x} \mid \mu_W, \Sigma_W) = p_A f(\mathbf{x} \mid \mu_A, \Sigma_A) + p_B f(\mathbf{x} \mid \mu_B, \Sigma_B) \qquad (2.10)$$

where $f(\mathbf{x} \mid \mu_A, \Sigma_A)$ and $f(\mathbf{x} \mid \mu_B, \Sigma_B)$ are Gaussian multidimensional probability density functions (pdf).

P-values from Shapiro 4-dimensional tests are reported in Table 2.2(b): low values are again caused by the high number of samples.

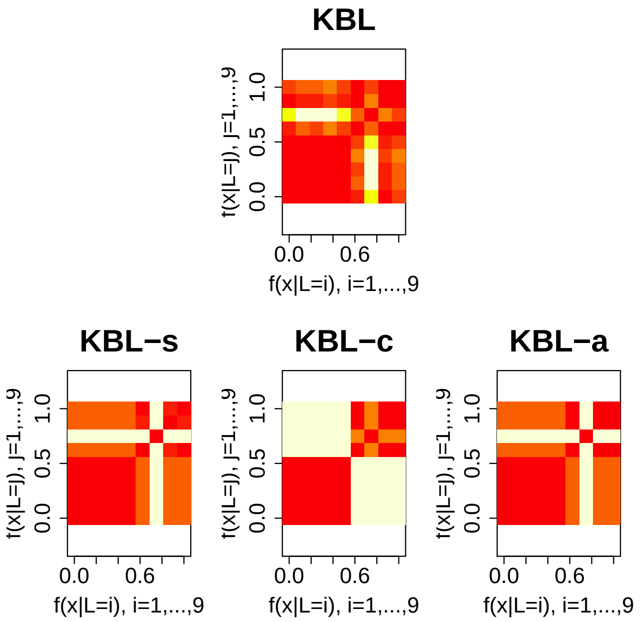


Figure 2.7: Matrix-Plot of the dissimilarity matrix $\mathcal{D}$ (top) and of matrices reprenting the dissimilarity level at which $f(\mathbf{x}|L=i)$ and $f(\mathbf{x}|L=j)$ are joined in a cluster by the method (bottom).

Table 2.2: (a)Cophenetic correlation coefficients for the single, complete and average linkage methods; (b) p-values from Shapiro 1-dimensional and 4-dimensional tests, based on 2500 Monte Carlo iterations, for the three groups Cashmere (C), Wool$_A$ (W$_A$), Wool$_B$ (W$_B$).

(a) Cophenetic coefficients

	S.L.	C.L.	A.L
ρ_{coph}	0.82	0.58	0.82

(b) Shapiro p-values

	C	W$_A$	W$_B$
1-dim p-value	0.032	0.546	0.059
4-dim p-value	0	4×10^{-4}	96×10^{-4}

2.4 Discriminant Analysis

In this Section we perform a discriminant analysis and do some comparisons with respect to the choice of the number of features, the number of groups and the variables used in the construction of the classifier.

Discriminant Analysis (DA) aims at building a classifier $\delta^*(\mathbf{x})$, i.e. a partition of the features space $\mathbb{R}^m$, that minimize the Expected Cost of Misclassification (ECM):

$$ECM(\delta) = \sum_{h=1}^{g} \int_{T_h} \sum_{i \neq h} p_i \ f_i(\mathbf{x}|L=i) \ c[h|i] \ d\mathbf{x}. \tag{2.11}$$

In expression (2.11), $g=2$ is the number of groups ($L=1$ being related with cashmere, $L=2$ with wool), p_i is the prior probability of the i-th group, $f_i(\mathbf{x}|\mu_\mathbf{i}, \mathbf{\Sigma_i}) = f_i(\mathbf{x}|L = \ i)$

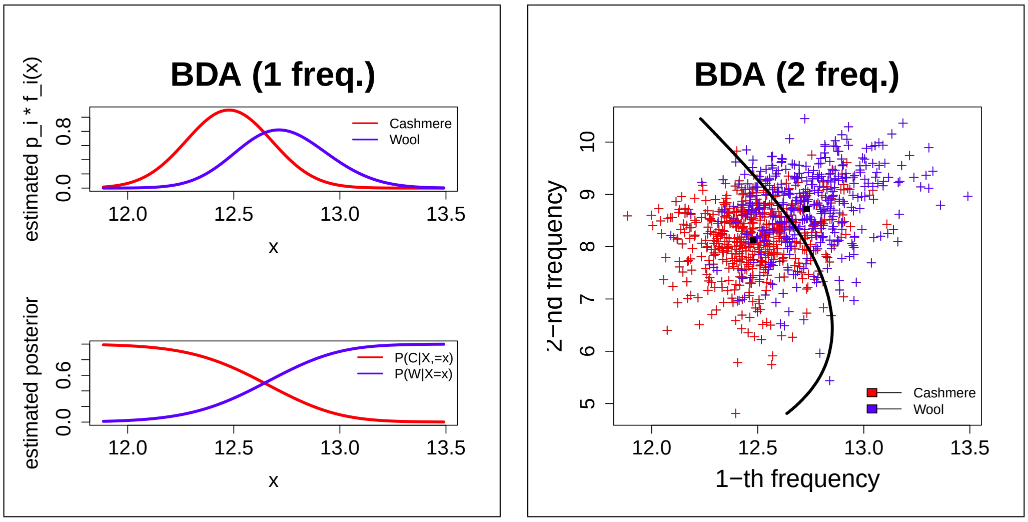


Figure 2.8: Estimates of 1-dim (left) and 2-dim (right) regions T_i^*, $i \in \{1, 2\}$, created by the Bayesian classifier. On the left is presented the equality in equation (2.12); black points on the right are the sample means of the groups.

is the pdf of the random vector $\mathbf{X} = (X_1, \ldots, X_m)$, whose components are the logarithms of modules of the m Fourier coefficients: a Gaussian model for $L = 1$ and a Gaussian mixture model for $L = 2$, as described in Section 2.3; T_h is the m-dimensional region of euclidean space predicted as belonging to the h-th group by the classifier, coefficients $c[h|i]$ represent the cost of classifying as belonging to the h-th group a fiber belonging to the i-th group. In the present work we imposed all the extra-diagonal values of the costs matrix $\mathbf{C}$ equal to one, supposing the misclassification equally significant in both directions, but this choice could be easily modified, getting other estimates of ECM.

This cause the optimal classifier δ^* to become a Bayesian classifier: the posterior probability density function of belonging to all groups is valuated for each fiber, that is labeled with the group which has the maximal posterior, as shown in equation (2.12)

$$\begin{aligned} T_i^* &= \{\mathbf{x} \in \mathbb{R}^m : c[j|i]\, p_i\, f_i(\mathbf{x}|L = i) \geq c[i|j]\, p_j\, f_j(\mathbf{x}|L = j)\} \\ &= \{\mathbf{x} \in \mathbb{R}^m : p(L = i|\mathbf{X} = \mathbf{x}) \geq p(L = j|\mathbf{X} = \mathbf{x})\}, \quad i, j \in \{1, 2\},\, i \neq j \quad (2.12) \end{aligned}$$

We recall the notation on the posterior probability functions:

$$p(L = i|\mathbf{X} = \mathbf{x}) = \frac{p_i\, f_i(\mathbf{x}|L = i)}{\sum_{j=1}^{g} p_j\, f_j(\mathbf{x}|L = j)}. \tag{2.13}$$

If, besides $c[i|j] = c$, $\forall\, 1 \leq i,\, j \leq g$, $i \neq j$, c constant, we assume the Gaussian model to hold for probability density functions $(\mathbf{X}|L = i \sim \mathcal{N}_m(\mu_{\mathbf{i}}, \mathbf{\Sigma_i}))$, optimal regions T_i^* are separated by quadratic functions (2.14) and the whole method results in Quadratic Discriminant Analysis (2.15):

$$d_i^Q(\mathbf{x}) = -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{x}-\mu_i)'\Sigma_i^{-1}(\mathbf{x}-\mu_{\mathbf{i}}) + log(p_i) \tag{2.14}$$

$$T_i^* = \{\mathbf{x} \in \mathbb{R}^m : d_i^Q(\mathbf{x}) \geqslant d_h^Q(\mathbf{x}), \quad \forall\, 1 \leq h \leq g\}\}. \tag{2.15}$$

Figure 2.8 shows estimates of the 1-dimensional and 2-dimensional optimal regions T_i^*, $1 \leq i \leq 2$ created by the Quadratic classifier. Apparent Error Rate (AER) is the ECM computed using sample parameters in the densities and it is in turn estimated typically in two ways: Apparent Error Rate (APER) and *cross-validation* estimate (AER$_{CV}$). In both cases the integrals are estimated creating the *confusion matrix* and dividing the number of misclassified fibers with the total number of fibers in the group; in the first case the matrix is computed directly on the sample, while in the second case it is computed extracting one fiber at time, building the Quadratic classifier with the remaining fibers and classifying the removed one (*leave-one-out algorithm*). Referring to Figure 2.8, the 1-dimensional classifier (the one using only the first feature) has APER = 26.84%, AER$_{CV}$ = 26.95%, while the 2-dimensional classifier has APER = 23.82%, AER$_{CV}$ = 24.16%. APER estimate is globally decreasing towards zero if we increase the number of features used in the construction of the classifier, while AER$_{CV}$ estimate initially decreases and then start increasing again, due to the *overfitting* phenomenon; this trend is shown in Figure 2.9, where we plot the errors in function of the modules related to the first 50 frequencies. We characterize the features to be used as those that contribute to decreasing AER$_{CV}$. We have chosen the log-modules $X_{1\div4}$ (low frequencies), X$_{13\div15}$ and X$_{26}$ (middle frequencies); the final classifier has APER = 17.89% and AER$_{CV}$ = 19.91%.

We now compare it with the followings: first we keep log-modules as features and change the number of groups into $g = 3$ or $g = 9$, then we do the same analyses changing the features considered (modules of DFT and scores instead of log-modules).

2.4.1 Comparisons changing number of groups and features

In the present Subsection we first show the result obtained varying the number of groups, fixing the DFT log-modules as features variables. Definition of ECM in equation (2.11) confirms the intuitable concept that increasing the number of groups causes a growth of the error, since each of the summations has more than one non-negative term. Figures 2.10(b) and 2.11(b) show that both with three and nine groups even sample means of some groups are misclassified, unlike the case in which we considered wool being the mixture of two groups. The size of the error with one feature (the starting point in Figures 2.10(c) and 2.11(c)) is comparable with that of the trivial

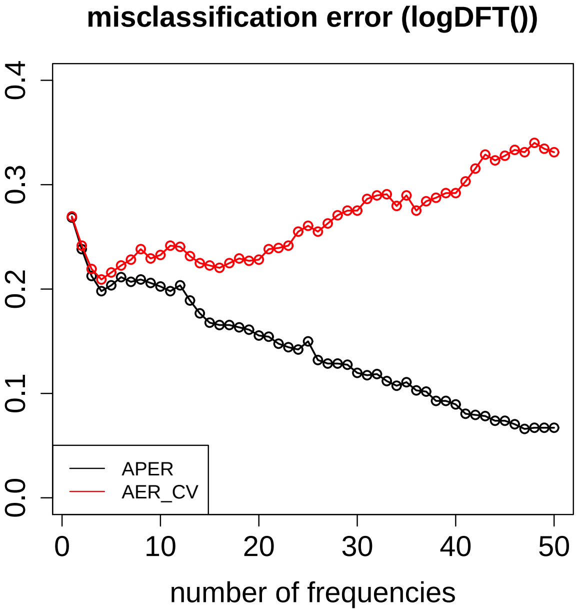
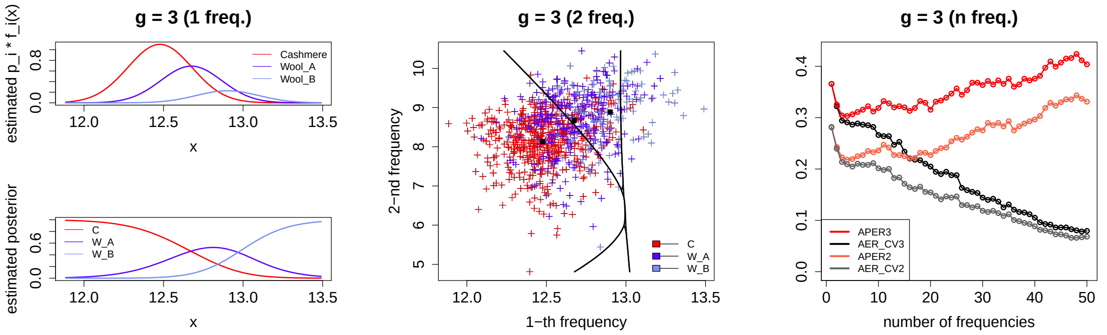


Figure 2.9: Estimates of AEPER and AER$_{CV}$. APER is globally decreasing in features (logarithm of modules considered), while AER$_{CV}$ presents a minimum.

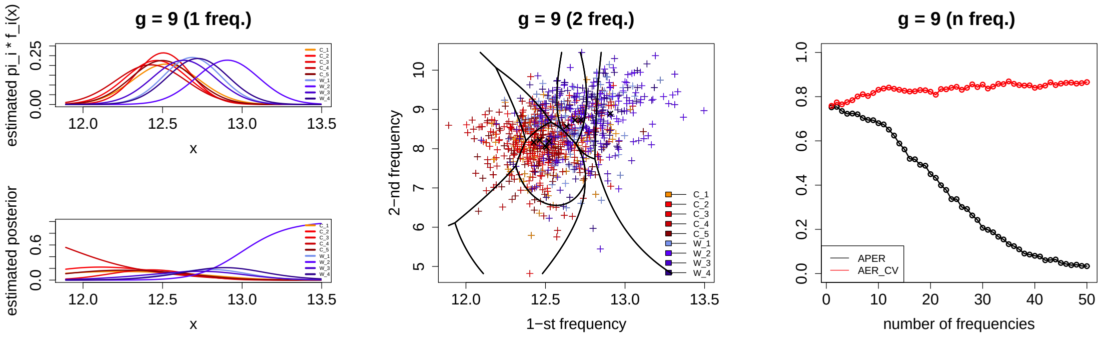
classifier (approximately AER $= 4/9 \simeq 0.44$ if $g = 3$ and AER $= 8/9 \simeq 0.98$ if $g = 9$). We have also tried to see what happens if we take into account three groups in the construction of the Quadratic classifier (so that three regions T_i^* are detected) and to estimate AER considering *a posteriori* the costs $c[W_A|W_B]$ and $c[W_B|W_A]$ null. It is not properly a correct procedure, but we can justify it with the fact the two subgroups belong to the same macro-group. Moreover, even if the error is under that estimated in the correct way with three groups (Figure 2.10(c)), it does not improve with respect to the error obtained with the Gaussian Mixture Model, as shown in the first line of Table 2.3. The reason can be understood looking for example at Figure 2.10(a): pdf of Gaussian densities mixture of the groups W$_A$ and W$_B$ (blue and lightblue densities) is higher than the single W$_A$-density in the *borderline* cashmere-wool region. This means that in the not-GMM there are some borderline wool fibers classified as cashmere, that are correctly classified in the GMM.

We have then made similar analyses changing the features from DFT log-modules into DFT modules and into scores; results, related to the Gaussian mixture model GMM ($g = 2$) and to the case of $g = 3$ groups used in building the classifier and $g = 2$ groups used in evaluating the errors, are reported in Table 2.3. If we look only at AER$_{CV}$, we notice the case with $g = 3$ has worse performances than the case $g = 2$. APER estimate is always too optimistic and less stable; moreover APER values differ slightly between the left and the right part of the table. Secondly, among the three transformations taken into account, the scores present the lower AER$_{CV}$, even if



(a) (b) (c)

Figure 2.10: 1-dim and 2-dim classifier and estimates of AER in the case $g = 3$.



(a) (b) (c)

Figure 2.11: 1-dim and 2-dim classifier and estimates of AER in the case $g = 9$.

the multivariate normality hypothesis is not as well met as by the log-modules. This indicate that, even if capturing the higher quantity of residual variance is not directly connected with the possibility of reducing classification error, in the present case-study this happen. Errors related to scores and log-modules are very similar, but maybe a classifier built with a more appropriate pdf for the scores would make the difference more marked. Finally we ascribe the high estimates of the errors of the DFT-modules classifier to the sharp asymmetry of they distribution, in which case the Gaussian assumption clearly fails.

Figure 2.12 gives a synthetic insight of the errors relative to different variables we considered in Table 2.3. With respect to it, we have added some more cases. The meaning of indices on the abscissa (features used in building the Quadrtic classifier) is presented in Table 2.4. Compared to the three groups of variables already analyzed, we have taken into consideration also the longitudinal mean ($RL_i = \frac{1}{p}\sum_{s=1}^{p} r_i(s)$) and standard deviation ($SL_i = \sqrt{\frac{1}{p}\sum_{s=1}^{p}(r_i(s) - RL_i})$ of the radii, and the first four scores for the i-th fiber, $1 \leq i \leq n$. The first ones are the most simple features usable

Table 2.3: Comparisons of APER and AER$_{CV}$ when changing the number of groups (left vs. right part of the table) and the features (different lines of the table).

	g=2, GMM			**g=3, Error as if g=2**		
	features	*APER*	*AER$_{CV}$*	*features*	*APER*	*AER$_{CV}$*
log(DFT)	1÷4, 13÷15, 26	17.89%	19.91%	1÷3, 10, 13÷15 26	21.14%	22.15%
DFT	1÷3, 12÷17, 24	19.35%	20.69%	1÷3, 11÷15, 24÷25, 28	16.67%	16.89%
Scores	1÷3, 5÷6, 28÷32	16.89%	18.23%	1÷5, 26÷33	15.21%	17.56%

in multivariate analysis, while the second are a number of sequential scores, without selection of those responsible of AER$_{CV}$ decreases.

The horizontal line represents the errors of the trivial classifiers ($4/9 \simeq 0.44$), which attributes to all the fibers the label of the group with the highest sample size. All the classifiers, with any combination of features, perform better than the trivial one. APER and AER$_{CV}$ are plotted also for a quadratic classifier built considering only two groups of fibers (without mixture). Optimal features are

- $1 \div 3$, $13 \div 15$, for the DFT log-modules;
- $1 \div 3$, $12 \div 15$, for the DFT modules;
- $1 \div 6$, $28 \div 29$, for the scores.

Since building the classifier with three groups and computing the error only with two is not properly a correct procedure, we have plotted its errors, but we should not focus on them. This is confirmed by the fact that the error in this case have an anomalous trend, decreasing with DFT modules as fetures, instead of increasing with respect to DFT log-modules, like errors of other classifiers do.

As we expected, AER$_{CV}$ in the case with two groups is higher than in the mixture model, since we lose in precision ignoring the fact that the group W_B can easily be identified. The only exception is represented by the DFT-modules features, but they do not meet at all the Normality assumption, and, in any case, the error is comparable with the mixture model.

Looking at the figure, we conclude that a Gaussian mixture model, with an optimal choice of scores, and, if possible, a better estimation of their multivariate probability density function, should be the final proposal from a mathematical point of view, in order to perform discriminant analysis of FIBER dataset, if we consider only the radii information. They are derived from the statistical procedure of computing functional

Table 2.4: Meaning of indices on the abscissa in Figure 2.12.

1	mean radius and standard deviation
2	scores 1 ÷ 4
3	optimal scores
4	optimal DFT log-modules
5	optimal DFT modules

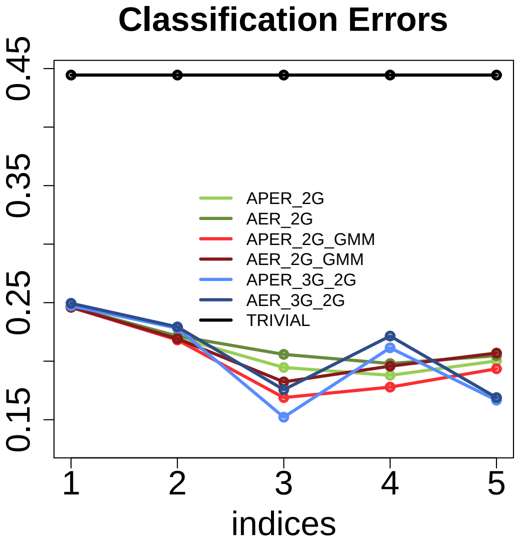


Figure 2.12: APER (light lines) and AER$_{CV}$ (dark lines) committed by the classifier build with 2 groups (green lines), with the Gaussian mixture model (red lines) and with 3 groups, computing the error as if they were 2 (blue lines). The error the trivial classifier, the same for the three cases, is plotted in black.

principal components.
An alternative proposal, maybe more common and easy to accept in an industrial context, could be using DFT log-modules and a classifier that takes into account the existence of two groups or use the Gaussian mixture model for wool fibers; its has namely AER$_{CV}$ = 19.91%, while the longitudinal mean radius and standard deviation provide AER$_{CV} \simeq$ 24.6%.

In this case-study functional analysis has let us improve the error committed with respect to that committed if we use only the synthetic information contained in RL_i and SL_i for the i-th fiber. We leave to future work to study if there are improvements analyzing in a functional framework features more sophisticated than radii and proposed in technical literature, such as the height of the scales (see Robson 1997).

Chapter 3

Sparse Sobolev Metrics

3.1 Introduction

In this chapter we deal with the problem of the choice of the metric space, which functional data belong to, as anticipated in the Introduction.
When we treat functional data, we immediately face an important statistical problem. The space E in which the variables are taking their values, as in Definition 1, is an infinite dimensional space. We are concerned with the selection of a norm in that space (and the metric induced by it) both for convergence needs (not all the norms are equivalent, unlike in the multivariate framework) and for representation needs: some metric better highlight differences between functions than others.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{T}$ be an open subset of $\mathbb{R}$. We denote by $\mathcal{H}^k(\mathcal{T}) = \{f : \mathcal{T} \to \mathbb{R},\, s.t.\, D^j f \in L^2, \forall j \leq k, j \in \mathbb{N}\}$ the Sobolev space, subset of $L^2(\mathcal{T})$, consisting of the equivalence classes of functions with weak derivatives $D^j f$ in L^2, for each derivation order $j \leq k$. The definition of weak derivative is given in the Appendix B; we recall that, when existing, the classic derivative of each order coincides with the weak one. It can be demonstrated that $\mathcal{H}^k(\mathcal{T})$ is a Hilbert space endowed with the inner product:

$$\langle f, g \rangle_{\mathcal{H}^k} = \sum_{j \leq k} \langle D^j f, D^j g \rangle_{L^2} = \sum_{j=0}^{k} \int_{\mathcal{T}} D^j f(s), D^j g(s) ds. \tag{3.1}$$

It is also a Banach space with respect to the induced norm:

$$\|f\|_{\mathcal{H}^k} = \left\{ \sum_{j \leq k} \|D^j f\|_{L^2}^2 \right\}^{\frac{1}{2}}. \tag{3.2}$$

Depending on whether or not we consider the belonging to a group, we make one of the following assumptions. We make this distinction only for simplicity of notation.

Assumption 1. We suppose to have g *functional random variables* $X_h(\omega, s) : \Omega \to \mathcal{H}^k(\mathcal{T}_h)$, $1 \leq h \leq g$, i.e. g measurable functions whose realizations are our *functional data* $x_{h,i}(s) : \mathcal{T}_h \to \mathbb{R}$, $1 \leq i \leq n_h$, belonging to $\mathcal{H}^k(\mathcal{T}_h)$; n_h is the number of observation for each of the random functions X_h. We suppose that $\cap_h \mathcal{T}_h = \mathcal{T} \neq \emptyset$.

Assumption 2. We suppose to have n *functional random variables* $X_i(\omega, s) : \Omega \to \mathcal{H}^k(\mathcal{T})$, with realizations the *functional data* $x_i(s) : \mathcal{T}_h \to \mathbb{R}$, $1 \leq i \leq n$, belonging to $\mathcal{H}^k(\mathcal{T})$.

When not ambiguous we will denote by $X(s)$ the random function $X(\omega, s)$, omitting to express the dependence on the case ω. Observe that we treat the scalar case ($\mathcal{T} \in \mathbb{R}$) for simplicity of computation and notation, but the results we give are extensible.

The choice of $\mathcal{H}^k(\mathcal{T})$ to be a Hilbert space is not necessary, since, as already stressed in the Introduction, recent works on functional data are based on the metric (or semi-metric) nature of the functional space which data belong to. Thus one could think to take into account more general Sobolev spaces, subsets of $L^p(\mathcal{T})$. Our choice is motivated by the first part of the work, in which we performed FPCA on random functions in L^2: the whole procedure of maximization is based on Theorem A.2 and on the concept of orthogonality induced by the inner product, with which the Hilbert spaces are endowed.

The aim of this Section is to provide an algorithm for computing a new Sobolev metric (or semi-metric), in which each term of the metric in equation (3.2) (expressed as $\|f\|_{\mathcal{H}^k} = d_{\mathcal{H}^k}(f, 0)$) is multiplied with coefficients that give weights to those derivatives that allow an *optimal* (in a sense that will be specified) unsupervised classification. We remember that a semi-metric is defined in the same way of a metric except for the axiom of *identity of indiscernible elements*, as follows:

Definition 2. A *semi-metric* on a given set H is a function $d_H : H \times H \to \mathbb{R}$, which satisfies the following properties for all the elements f, g, $h \in H$:

1. $d_H(f, g) \geq 0$;
2. $d_H(f, f) = 0$, but possibly $d_H(f, g) = 0$ for some $f \neq g$;
3. $d_H(f, g) = d_H(g, f)$;
4. $d_H(f, h) \leq d_H(f, g) + d_H(g, h)$.

Our aim is then to build the following semi-metric for functional data:

Definition 3. Let f, $g : \mathcal{T} \to \mathcal{H}^k(\mathcal{T})$ be two functions, and let $\mathbf{w} \in \mathbb{R}^k$ be a vector s.t. $w_j \geq 0$, $\forall j \leq k$, $j \in \mathbb{N}$; then a *Sobolev weighted semi-metric* (or *semi-norm*) is the function $d^{\mathbf{w}}_{\mathcal{H}^k} : \mathcal{H}^k(\mathcal{T}) \times \mathcal{H}^k(\mathcal{T}) \to \mathbb{R}$, s.t.

$$d^{\mathbf{w}}_{\mathcal{H}^k}(f, g) = \left\{ \sum_{j \leq k} w_j \|D^j f - D^j g\|^2_{L^2} \right\}^{\frac{1}{2}}.$$

It is not hard to see that $d^{\mathbf{w}}_{\mathcal{H}^k}$ satisfies the requirements in definition (3), and is thus a semi-metric, for every admissible value of $\mathbf{w}$, while it is a metric only if $w_0 \neq 0$. We aim at finding an optimal vector of weights $\mathbf{w}^*$ that, beyond the positivity constraint necessary in order to allow $d^{\mathbf{w}}_{\mathcal{H}^k}$ to be a semi-metric, meets the following two conditions:

1. to maximize (or at least accentuate with respect to the not weighted case) the dissimilarity between functional data that are not belonging to the same group (i.e. that are realizations of the random functional variables X_{hi} and X_{kj}, $h \neq k$) if used to build the dissimilarity matrix in a problem of not supervised classification;

2. to provide sparsity in the choices of derivatives, assigning exactly zero-weights to those terms that less accentuate the dissimilarity between functions of different groups.

We are able to formalize these requirements in the form of an optimization problem and to write an iterative algorithm for its solution taking inspiration from the work of Tibshirani and Witten (2010). They propose two methods for the variable selection problem in a framework of unsupervised classification of high-dimensional data: the first method is a form of *sparse k- means clustering*, while the second one is a *sparse hierarchical clustering*. Both approaches are a variation of a general *sparse clustering framework*, that we recall in the next Section 3.2; we focus then on showing how their second approach for variable selection can be made into an approach for metric selection in Section 3.3.

3.2 Sparse Feature Selection in Clustering

In the present Section we assume to deal with a multivariate high-dimensional dataset, in order to explain the construction performed by Tibshirani and Witten (2010). Let $\mathbf{X}$ denote a $n \times p$ matrix, with $p \gg n$, with n observations and p variables and let $\mathbf{X}_j \in \mathbb{R}^n$ denote feature j and $\mathbf{x_i}$ the observation i. Call $d_{i,i',j}$ a *dissimilarity* between observations i and i' along the feature j (e.g. d could be the squared Euclidean distance $d_{i,i',j} = (X_{ij} - X_{i'j})^2$); d is assumed to be additive in the features, i.e. $d(\mathbf{x_i}, \mathbf{x_{i'}}) = \sum_{j=1}^{p} d_{i,i',j}$. The authors show that under these hypotheses many clustering methods can be expressed as an optimization problem of the form

$$\underset{\Theta \in D}{\text{maximize}} \quad \{\sum_{j=1}^{p} f_j(\mathbf{X_j}, \mathbf{\Theta})\} \tag{3.3}$$

where $f_j(\mathbf{X_j}, \mathbf{\Theta})$ is a function that involves only the jth features, and $\mathbf{\Theta}$ is a parameter restricted to lie in a set D. We observe that a high number of variable is not a necessary condition for the methods proposed to work, but is instead the context that motivates the whole analysis; thus the fact that the hypothesis is not met by the construction we make in Section 3.3 is not problematic.

Example 1. An intuitive example of a clustering method turned in the form (3.3) is that of K-means, where $f_j(\mathbf{X_j}, \boldsymbol{\Theta}) = \sum_{j=1}^{p}(\frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{i,i',j} - \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i'\in C_k} d_{i,i',j})$ is the between cluster sum of squares (BCSS) for feature j, and $\boldsymbol{\Theta}$ is a partition of the observations into K disjoint sets of cardinality n_k, i.e. it is a vector $\boldsymbol{\Theta} \in \mathbb{R}^n$ with components $1 \leq \Theta_j \leq K$. We will explain the hierarchical clustering case in Section 3.3, since some more passages are necessary.

Then Tibshirani and Witten (2010) define *sparse clustering* as the solution to the problem

$$\begin{aligned} \underset{\mathbf{w};\Theta\in D}{\text{maximize}} \quad & \{\sum_{j=1}^{p} w_j\, f_j(\mathbf{X_j}, \boldsymbol{\Theta})\} \\ \text{subject to} \quad & \|\mathbf{w}\|_2^2 \leq 1, \qquad \|\mathbf{w}\|_1 \leq c, \\ & w_j \geq 0 \qquad \forall j, \end{aligned} \tag{3.4}$$

where w_j is a weight corresponding to feature j and c is a tuning parameter. It turns out that, for both the constraint to be active, it must be $1 \leq s \leq \sqrt{p}$. This is shown in the Appendix C, where we give proofs and considerations about the solving algorithm for a quite less general problem, that turns out to be the one necessary for the construction of the sparse Sobolev metric. The construction (3.4) for sparse clustering is motivated by the authors with the following observations:

1. If $w_1 = \ldots = w_p$ in (3.4), then the criterion reduces to (3.3).

2. The L_1 penalty on $\mathbf{w}$ results in sparsity for small values of the tuning parameter c, causing some of the w_j's to be zero.

3. The L_2 penalty is necessary since, without it, the solution would not be bounded: weights w_j would not have limited components.

4. The value of w_j can be interpreted as the contribution of feature j to the resulting sparse clustering: a large value of w_j indicates a feature that contributes greatly, while $w_j = 0$ means that feature j is not involved in the clustering.

It is immediate to prove that, for the solution not to be trivially $\mathbf{w} = \mathbf{0}$, it is necessary that $f_j(\mathbf{X_j}, \boldsymbol{\Theta}) > 0$ for some or all j, while if $f_j(\mathbf{X_j}, \boldsymbol{\Theta}) > 0$, then the nonnegativity constraint on w_j has no effect.
An iterative solving algorithm (Algorithm 2, where $a_j = f_j(\mathbf{X_j}, \boldsymbol{\Theta})$) is proposed in this general framework: it does not allow to reach a global optimum of the problem (3.4),

but guarantees that the objective function is incremented at each iteration, which is composed by two steps. In the first step, holding $\mathbf{w}$ fixed, (3.4) is optimized with respect to $\mathbf{\Theta}$; this step is typically an application of a standard clustering procedure to a weighted version of the data. In the second one, holding $\mathbf{\Theta}$ fixed, we optimize with respect to $\mathbf{w}$. This second reduced problem is convex (linear objective function and convex constraints) and it is solved thanks to optimization techniques, as stated by the following theorem.

Theorem 1. *Let x^+ denotes the positive part of x ($x^+ = x$ if $x > 0$, $x^+ = 0$ if $x \leq 0$) and define the soft-thresholding operator $S(x, c) = \text{sign}(x)(|x| - c)^+$. Then the solution to the convex sub-problem*

$$\begin{aligned} &\underset{\mathbf{w}}{\textit{maximize}} \quad \{\sum_{j=1}^{p} w_j\, a_j\} \\ &\textit{subject to} \quad \|\mathbf{w}\|_2^2 \leq 1, \qquad \|\mathbf{w}\|_1 \leq c, \\ &\qquad\qquad\quad w_j \geq 0 \qquad \forall j \end{aligned} \tag{3.5}$$

is given by $\mathbf{w} = \dfrac{S(\mathbf{a}, \Delta)}{\|S(\mathbf{a}, \Delta)\|_2}$, where $\Delta = 0$ if that results in $\|\mathbf{w}\|_1 \leq c$; otherwise, $\Delta > 0$ is chosen to yield $\|\mathbf{w}\|_1 = c$.

The proof of the theorem is given in the Appendix C, in theorem (C.5; it follows from the Karush-Kuhn-Tucker first order conditions (see e.g. Nocedal and Wright (2006), Chapter 5). Note that the stopping criterion of the iterative algorithm involves the value $\mathbf{w}^r$, which is the set of weights obtained at iteration r, and ε, a sufficiently small value for the relative error (e.g. 10^{-4}).We will make some more considerations on the stopping criterion in Section 4.1; from now on we will indicate it with the general notation of *convergence*

Algorithm 2: Sparse Clustering

```
Data: $a_j = f_j(\mathbf{X_j}, \mathbf{\Theta})$;
Result: $\mathbf{w}$, $\mathbf{\Theta}$;
Initialization: $w_1 = \ldots = w_p = \frac{1}{\sqrt{p}}$;
while $\frac{\sum_{j=1}^{p} |w_j^r - w_j^{r-1}|}{\sum_{j=1}^{p} |w_j^{r-1}|} > \varepsilon$ do
    1. $\underset{\Theta \in D}{\text{maximize}} \quad \{\sum_{j=1}^{p} w_j\, a_j\}$ by clustering reweighted data;
    2. $\mathbf{w} \leftarrow \frac{S(\mathbf{a}, \Delta)}{\|S(\mathbf{a}, \Delta)\|_2}$;
end
```

The authors then cast problem (3.4) in a version of sparse K-means and sparse hierarchical clustering, providing a method for an optimal choice of the tuning parameter c and the R-package `sparcl` for simulations and analysis (see Tibshirani and Witten (2013)). For our purposes we are interested in the second method: we provide some details on it in the next Section 3.3, where we analyze directly the problem of metric selection.

3.3 Sparse Sobolev Metrics Selection in Clustering

We aim at building a Sobolev weighted semi-metric, as in definition (3), i.e. a vector of weights $\mathbf{w}^*$ satisfying the optimality and sparsity requirements explained at the beginning of this Section. We follow what Tibshirani and Witten (2010) call *sparse hierarchical clustering*, which is indeed a technique that can be applied to any method that takes a dissimilarity matrix as its input. The main novelty that we introduce is in the type of data between which the dissimilarity is computed, and not in the steps of the optimal algorithm used to construct $\mathbf{w}^*$. We suppose to be in the conditions expressed in Assumption 2, dealing with random functions $X_i(s) : \mathcal{T} \to \mathbb{R}$, with realizations $x_i(s)$, $1 \leq i \leq n$ belonging to the Sobolev space $\mathcal{H}^k(\mathcal{T})$. Our novel idea is then to attribute to the levels of derivation the role assumed by the features in the multivariate framework. So we give a definition of dissimilarity between two observed functions according to the following associative scheme between the multivariate and the functional case:

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix} \longrightarrow \tilde{\mathcal{X}} = \begin{pmatrix} X_1(s) & D^1X_1(s) & \cdots & D^kX_1(s) \\ X_2(s) & D^2X_2(s) & \cdots & D^kX_2(s) \\ \vdots & \vdots & \ddots & \vdots \\ X_n(s) & D^1X_n(s) & \cdots & D^kX_n(s) \end{pmatrix}.$$

Definition 4. Let $x_i(s), x_{i'}(s) : \mathcal{T} \to \mathbb{R}$ be two functional data, belonging to the Sobolev space $\mathcal{H}^k(\mathcal{T})$. We define dissimilarity along the feature j as the square of the L^2-norm of the difference between their j-th derivative (denoting with zero-derivative the function itself): $d_{i,i',j} = \|D^j x_i(s) - D^j x_{i'(s)}\|^2_{L^2}$, $0 \leq j \leq k$, $1 \leq i, i' \leq n$.

Note that the dissimilarities along features are the single terms of the square of the Sobolev norm of the difference between the two functions $\|x_i(s) - x_{i'}(s)\|^2_{\mathcal{H}^k}$. Note also that d is additive in features (and we call *overall dissimilarity* between functions the quantity $U_{i,i'} = \sum d_{i,i',j}$, that is just equivalent to $\|x_i(s) - x_{i'}(s)\|^2_{\mathcal{H}^k}$), as required in the general framework for sparse feature selection in Section 3.2.

We now present the procedure both in the not sparse and in the sparse version, getting optimization problems of the form (3.3) and (3.4). We anticipate that, writing their sparse version of hierarchical clustering in the multivariate context, Tibshirani and Witten (2010) recognize the possibility of seeing it in the light of the *penalized matrix decomposition* (PMD) method that they had proposed in a previous work (Hastie, Tibshirani, Witten, 2009). This leads to the possibility of constructing clusterings that are complementary in a statistical sense (nearly orthogonal, in a mathematical sense). Be warned that this results has not in output a mathematical object with the properties of a metric, in particular it lacks the property of positivity. It provides instead a quantity that indicates the dissimilarity between data/functions in a way that is is compatible with the characteristics of insensitivity to translations of some classifiers such as hierarchical clustering.

In both the not sparse and sparse construction, we have in input the between function dissimilarities $d_{i,i',j}$ along feature j, and in output $\mathbf{U}^*$ a synthetic dissimilarity symmetric $n \times n$ matrix that can be used for unsupervised classification. In the first case $\mathbf{U}^*$ is the overall dissimilarity already introduced, whose elements are $U_{i,i'} = \sum_j d_{i,i',j}$; in the second case we have an additional output, the vector $\mathbf{w}^* \in \mathbb{R}^{(k+1)}$: then it results that the elements of $\mathbf{U}^*$ are $U_{i,i'} = \sum_j w_j^* d_{i,i',j}$ and they have to be interpreted as the square of the Sobolev weighted semi-norms $\left(d_{\mathcal{H}^k}^{\mathbf{w}^*}(x_i(s), x_{i'}(s))\right)^2 = \sum_{j \leq k} w_j \|D^j x_i(s) - D^j x_{i'}(s)\|_{L^2}^2$.
Since scaling the dissimilarity matrix by a factor does not affect the output of a unsupervised classification methods, such as hierarchical clustering, we ignore proportionality normalization constants in the following discussion. From now on, let $p = k + 1$. Then we can state the following theorem.

Theorem 2. *Let* $\mathbf{U}^*$ *optimize the following criterion*

$$\begin{aligned} &\underset{\mathbf{U}}{maximize} \quad \{\sum_{j=1}^{p} \sum_{i,i'}^{n} d_{i,i',j} U_{i,i'}\} \\ &subject\ to \quad \sum_{i,i'}^{n} U_{i,i'}^2 \leq 1. \end{aligned} \tag{3.6}$$

Then $U_{i,i'}^* \propto \sum_j d_{i,i',j}$

The proof of theorem (2) is a simple application of theorem (C.5) in the Appendix, if we recast the data structure. In fact we are in the same conditions of that theorem if we call $\mathbf{D} \in \mathbb{R}^{n^2 \times p}$ the matrix in which column j consist of the elements $d_{i,i',j}$ strung out into a vector and $\mathbf{u} \in \mathbb{R}^{n^2}$, the vector obtained by stringing out $\mathbf{U}$. Notice that the vector $\mathbf{a}$ of the theorem (C.5) is now equivalent to $\mathbf{U1}$, where $\mathbf{1} \in \mathbb{R}^p$ is a vector with unitary components.
As a consequence, if we can think of the not sparse construction of $\mathbf{U}^*$ as resulting from the criterion (3.6), then to obtain sparsity in features (i.e. in the choices of the

derivatives used in building the Sobolev metric) we modify (3.6) by multiplying each element of the summation over j, by a weight w_j, subject to constraints on the weights, as stated in the following theorem.

Theorem 3. *Let* $\mathbf{U}^{**}$ *optimize the following criterion*

$$\begin{aligned}
&\underset{\mathbf{w},\mathbf{U}}{\textit{maximize}} && \{\sum_{j=1}^{p} w_j \sum_{i,i'}^{n} d_{i,i',j} U_{i,i'}\} \\
&\textit{subject to} && \sum_{i,i'}^{n} U_{i,i'}^2 \leq 1, \\
& && \|\mathbf{w}\|_2^2 \leq 1, \qquad \|\mathbf{w}\|_1 \leq c, \qquad w_j \geq 0 \quad \forall j.
\end{aligned} \tag{3.7}$$

Then $U_{i,i'}^{**} \propto \sum_j d_{i,i',j} w_j$.

Observe that (3.6) takes the form (3.3) with $\boldsymbol{\Theta} = \mathbf{U}$, $f_j(\mathbf{X}_j, \Theta) = \sum_{j=1}^{p} \sum_{i,i'}^{n} d_{i,i',j} U_{i,i'}$. It follows directly that (3.7) takes the form (3.4), and so sparse construction of the dissimilarity matrix fits into the framework of Section 3.2. The proof of theorem (3) is the same of theorem (C.6) in the Appendix, since we can rewrite criterion (3.7) in the following way:

$$\begin{aligned}
&\underset{\mathbf{w},\mathbf{u}}{\text{maximize}} && \mathbf{u}'\mathbf{D}\mathbf{w} \\
&\text{subject to} && \|\mathbf{u}\|_2^2 \leq 1, \\
& && \|\mathbf{w}\|_2^2 \leq 1, \qquad \|\mathbf{w}\|_1 \leq c, \qquad w_j \geq 0 \quad \forall j,
\end{aligned} \tag{3.8}$$

and the nonnegativity constraint on $\mathbf{w}$ can be dropped if $d_{i,i',j} \geq 0$. Then, since the optimization problem (3.9) is bi-convex in $\mathbf{u}$ and $\mathbf{w}$ (with $\mathbf{w}$ fixed it is convex in $\mathbf{u}$, and vice versa), we can write a simple solving optimization algorithm, analogous to Algorithm 6 in the Appendix relative to the rank-1 penalized matrix decomposition PMD ($\cdot$, L_1) approach presented by Hastie, Tibshirani, Witten (2009); notice also that it is a particular case of Algorithm 2, relative to the general problem of feature selection in clustering.

Algorithm 3: Sparse Sobolev Metric Construction

```
Data: D;
Result: w, u;
Initialization: initialize w as $w_1 = \ldots = w_p = \frac{1}{\sqrt{p}}$;
while not convergence do
    1. $\mathbf{u} \leftarrow \frac{\mathbf{Dw}}{\|\mathbf{Dw}\|_2}$;
    2. $\mathbf{w} \leftarrow \frac{S(\mathbf{a},\Delta)}{\|S(\mathbf{a},\Delta)\|_2}$, where $\mathbf{a} = \mathbf{D}'\mathbf{u}$, and $\Delta = 0$ if that results in $\|\mathbf{w}\|_1 \leq c$,
       otherwise $\Delta > 0$ is chosen to be a positive constant such that $\|\mathbf{w}\|_1 = c$;
end
rewrite u as a $n \times n$ matrix U.
```

Our purpose was to build a Sobolev metric that emphasize the dissimilarity between functions not belonging to the same class, with respect to the usual not-weighted Sobolev metric, and that is sparse in features (derivatives used).
As regards the first objective, we will show how experimental results evidence the necessity of some corrections, that we explain in Chapter 4.
Our second objective is reached imposing the L_1-constraint on $\mathbf{u}$, varying the tuning parameter. As in Section 3.2, c is restricted to lie between 1 and $\sqrt{p}$. It is important to stress that, in order to gain sparsity, it's necessary that $p > 2$, as explained in the Appendix C.2: in our case this is equivalent to require that we consider at least the functions and their first derivative. Nevertheless we will not perform computations with a too high order of derivation, due to the numerical problems that arise, as we proceed increasing it. Tibshirani and Witten (2010) propose a method for choosing the value of c, that is closely related to the gap statistic of Hastie, Tibshirani and Walther (2001) for selecting the number of clusters K in the standard K-means clustering. We report it in the next Subsection 3.3.1, even if we don't use this criterion in our simulations; in fact the `sparcl` R-package provide a function for computing it for multivariate data, that should be reimplemented for functional data in an efficient way, so we simply perform simulations for a range of values of c, choosing the one that appears to perform better. More considerations on c are made in Chapter 4.

Finally we end the present Section presenting the possibility of constructing complementary sparse Sobolev metrics and we explore their meaning.

3.3.1 Selection of Tuning Parameter for Sparse Sobolev Metric Selection

We note that the parameter c cannot simply be chosen to maximize the objective function in problem (3.7), since as s is increased, the objective function will increase as well. A possible way of choosing the tuning parameter is the permutation approach consisting of the following steps:

1. obtain permuted datasets $\tilde{\mathcal{X}}_1, \ldots, \tilde{\mathcal{X}}_B$ by independently permuting the observed functions and their derivatives in the matrix $\tilde{\mathcal{X}}$;
2. chose a number of candidate tuning parameter c;
3. for each candidate:
   - (a) compute $O(c) = \sum_{j=1}^{p} w_j \sum_{i,i'}^{n} d_{i,i',j} U_{i,i'}$, the objective function obtained by performing sparse Sobolev metric selection with tuning parameter c on data $\tilde{\mathcal{X}}$;
   - (b) for $b = 1, \ldots, B$ compute $O_b(c)$, the objective function obtained by performing sparse Sobolev metric selection with tuning parameter c on data $\tilde{\mathcal{X}}_b$;
   - (c) calculate Gap(c)=log($O(c)$) $- \dfrac{1}{B} \sum_{b=1}^{B} \log(O_b(c))$;
4. choose c^* corresponding to the largest value of Gap(c); alternatively choose c^* to equal the smallest value for which Gap(c^*) is within a standard deviation of $\log(O_b(c^*))$ of the largest value of Gap(c).

The approach is based on the idea that, while there my be strong correlation between features (different order derivatives) in the original data $\tilde{\mathcal{X}}$, the features in the permuted datasets $\tilde{\mathcal{X}}_b$ are uncorrelated with each other. The gap statistic measures the strength of the clustering obtained on null data, i.e. the data that do not contain subgroups. The optimal tuning parameter value occur when this quantity is greatest.

3.3.2 Complementary Sparse Functional Data Hierarchical Clustering

Here we report a construction that Tibshirani and Witten (2010) do in order to find *complementary sparse clustering*, in the light of a previous work of Nowak and Tibshirani (2007) on *complementary clustering*. As well as standard hierarchical clustering, even sparse hierarchical clustering reveals to be often dominated by a single group of features that have high variance and are highly correlated with each other. Then they propose a method that allows for the discovery of a secondary sparse clustering after removing the signal found in the standard sparse hierarchical clustering. This objective is gained building a matrix that takes the role of the overall dissimilarity, but is not a linear combination of the feature-wise dissimilarity matrices composing $\mathbf{D}$.

In this Subsection we try to extend this method to our functional framework; even if we do not succeed in finding a complementary Sobolev metric, we are able to improve directly a classification method such as hierarchical clustering.
From a mathematical point of view, removing the nonnegativity constraint on $\mathbf{w}$, we can recognize in criterion (3.9) the rank-1 penalized matrix decomposition PMD $(\cdot, L_1)$ apprroach persented by Hastie, Tibshirani, Witten (2009) that we mention in the Appendix C.2.1. It is possible to generalize it, in order to obtain the rank-k penalized matrix decomposition, whose details useful for our purposes are briefly explained in the Appendix C.2.2. This reflects in the possibility of writing a number $r = \text{rank}(D)$ optimization problems analogous to problem (3.9), with an additional orthogonality constraint on the vectors $\mathbf{u_k}$s, $1 \leq k \leq r$. Here below we write the first subsequent optimization problem, but it could be generalized; however, according to the meaning we give to the solutions of these orthogonal problems, one should not exceed in the *orthogonality order* of the problem, since we risk capturing less and less significant phenomena. Call $\mathbf{u}_1$ and $\mathbf{w}_1$ the optimal solutions to criterion (3.9) that is, $\mathbf{U_1}$ (obtained by writing $\mathbf{u}_1$ in a matrix form) is a weighted linear combination of the feature-wise dissimilarity matrices componing $\mathbf{D}$, and $\mathbf{w}_1$ denotes the corresponding weights. Then the criterion

$$\begin{aligned} &\underset{\mathbf{w_2},\mathbf{u_2}}{\text{maximize}} && \mathbf{u_2}'\mathbf{D}\mathbf{w_2} \\ &\text{subject to} && \|\mathbf{u_2}\|_2^2 \leq 1, \qquad \mathbf{u_1}'\mathbf{u}_2 = 0 \\ & && \|\mathbf{w_2}\|_2^2 \leq 1, \qquad \|\mathbf{w_2}\|_1 \leq c, \qquad w_{2_j} \geq 0 \quad \forall j, \end{aligned} \tag{3.9}$$

results in a dissimilarity matrix obtained by writing $\mathbf{u_2}$ as a $n \times n$ matrix $\mathbf{U_2}$, and in a vector of weights $\mathbf{w_2}$. Note that in the above construction we don't impose an orthogonality constraint between vectors $\mathbf{w_1}$ and $\mathbf{w_2}$, since it is not required (nor it is clear if it is desirable) in order to obtain a rank-2 PMD. In fact the output vector $\mathbf{u_2}$ of the iterative solving Algorithm 4 is not proportional to $\mathbf{Dw_2}$, but instead $\mathbf{u_2} \propto (\mathbf{I} - \mathbf{u_1}\mathbf{u_1}')\mathbf{D}\mathbf{w_2}$, i.e. it is a linear combination of feature-wise matrices composing $\mathbf{D}$, projected in the linear space orthogonal to that generated by the columns of $\mathbf{U}$. That is, we are not gaining orthogonality by multiplying the same matrix $\mathbf{D}$ with orthogonal weights (which, in conjuction with the nonnegatity constraint on $\mathbf{w_2}$, would lead to a choice of different derivatives in building a second sparse Sobolev metric). Nevertheless we can suppose (and we have experimental evidence in our simulations) that $\mathbf{w_1}$ and $\mathbf{w_2}$ are likely to be almost orthogonal, since in the solution they are each associated with orthogonal $\mathbf{u_k}$s, $k = 1, 2$.
As we already observed, the matrix $\mathbf{U_2}$ is not a dissimilarity matrix, since its elements can be negative, due to the constraint that $\mathbf{u_1}'\mathbf{u_2} = 0$, while it has zeroes on the

diagonal. Thus, if $\mathbf{U_1}_{i,i'} = d^{\mathbf{w_1}}_{\mathcal{H}^k}(x_i(s), x_{i'}(s))$ meets the requirements to be a semi-metric (and in particular a Sobolev sparse semi-metric), $\mathbf{U_2}_{i,i'}$ does not. However the matrix $\mathbf{U_2}$ still indicates in what extent two functions are dissimilar, and in particular how do they differ if we project their dissimilarities in a orthogonal space. Thus, as suggested by Tibshirani and Witten (2010), it can be used to perform unsupervised classification techniques that are not affected by adding a constant to the off-diagonals elements and do not make use of the diagonal ones. Hierarchical clustering behaves in this way, so that we will perform directly what we call *complementary sparse functional data clustering*: without finding a complementary sparse Sobolev metric, that might be a preferable result, we classify directly our data in a somewhat orthogonal way. From now on, with an abuse of notation, we will call $\mathbf{U_1}$ *primary metric* and $\mathbf{U_2}$ *secondary metric*. In Section 4.1 we show that it is worth looking at both these metrics and at the resulting clustering.

It remains open the more abstract problem of what to do with two metrics. A first idea could be to create multidimensional labels for the data, comparing the results of both, but more work should be done in this field. Moreover we leave to our future investigations to find out properties of the sparse Sobolev metric that we could build creating the dissimilarity matrix $\mathbf{U_3} \propto \mathbf{D}\mathbf{w_2}$, where $\mathbf{w_2}$ are the almost orthogonal to $\mathbf{w_1}$ weights obtained thanks to Algorithm 4; another possibility could be to analyze $\tilde{\mathbf{U}}_\mathbf{3} \propto \mathbf{D}\tilde{\mathbf{w}}_\mathbf{2}$, where $\tilde{\mathbf{w}}_\mathbf{2}$ are exactly orthogonal to $\mathbf{w_1}$ obtained thanks to Algorithm 3 not applied to $\mathbf{D}$, but a matrix where we removed the derivatives levels used in building $\mathbf{w_1}$. In this case some attention should be paid on the fact that the tuning parameter c isn't more the same used in finding the primary metric.

Algorithm 4: Complementary Sparse Sobolev Metric Construction

```
Data: $\mathbf{D}, \mathbf{u_1}, \mathbf{w_1}$;
Result: $\mathbf{w_2}, \mathbf{u_2}$;
Initialization: initialize $\mathbf{w_2}$ as $w_{2_1} = \ldots = w_{2_p} = \frac{1}{\sqrt{p}}$;
while not convergence do
    1. $\mathbf{u_2} \leftarrow \frac{(\mathbf{I} - \mathbf{u_1}\mathbf{u_1}')\mathbf{D}\mathbf{w_2}}{\|\mathbf{U_1^\perp}\mathbf{D}\mathbf{w_2}\|_2}$, where $\mathrm{U}_1^\perp$ is a basis orthogonal to $\mathbf{u_1}$;
    2. $\mathbf{w_2} \leftarrow \frac{S(\mathbf{a}, \Delta)}{\|S(\mathbf{a}, \Delta)\|_2}$, where $\mathbf{a} = \mathbf{D}'\mathbf{u_2}$, and $\Delta = 0$ if $\|\mathbf{w_2}\|_1 \leq c$,
       otherwise $\Delta > 0$ is chosen to be a positive constant s.t. $\|\mathbf{w_2}\|_1 = c$;
end
rewrite $\mathbf{u_2}$ as a $n \times n$ matrix $\mathbf{U_2}$.
```

Chapter 4

Corrections to the Metric and Simulations

4.1 Introduction

From a theoretical point of view, we should be able to implement the procedure that we have proposed in Section 3.1, and that extend to the functional framework the multivariate one. However, there are some corrections that we think necessary before running Algorithms 3 and 4, precisely because of the different nature of our data. The corrections are motivated by different needs (measure units and variability): we will see that the first one works well, even if it is only necessary but not sufficient to make improvements, while the second is a proposal, on which more studies should be done in future. In the next Sections we expose them separately, after showing the results of the clustering without corrections, displaying what happens on two synthetic dataset, consisting of polynomial and trigonometric functions. We have close formulations for the derivatives, so that we are not concerned with numerical estimations. Here we consider the belonging to a group of functional data, adopting the notation expressed in Assumption 1, in order to be able to make considerations on the errors committed. Note that it is *a posteriori* information and that all the construction holds in a not supervised classification framework.

Dataset 1. We deal with two functional random variables $X_h(s)$, $1 \leq h \leq 2$ projected on the polynomial basis, with Gaussian coefficients, and a 100 samples balanced functional dataset. That is:

$$X_h(s) = A_{h0} + A_{h1}\left(\frac{x}{\mathcal{T}}\right) + \ldots + A_{hn}\left(\frac{x}{\mathcal{T}}\right)^n, \qquad 1 \leq h \leq 2, \tag{4.1}$$

are the two functional random variables, and

$$x_{hi}(s) = a_{h0i} + a_{h1i}x + \ldots + a_{hni}x^n, \qquad 1 \leq h \leq 2,\ 1 \leq i \leq 50 \tag{4.2}$$

are their realizations. The randomness is expressed by the coefficients A_{hk}, $0 \leq k \leq n$, that are realizations of 11-variate Gaussian variables: $\{A_{hk}\}_{k=0}^{n} = \mathbf{A}_h \sim \mathcal{N}_{n+1}(\mu_h, \mathbf{\Sigma})$. For simplicity we assume that the covariances matrices $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ are the same in the two groups, and that the difference between groups is expressed only by the means. Then it is easy to see that the j-th derivative of the random function $X_h(s)$ is

$$D^j X_h(s) = \begin{cases} \sum_{k=1}^{n} \dfrac{k!}{(k-j)!} A_{hk} x^{(k-j)}, & 0 \leq j \leq n \\ 0, & j > n. \end{cases} \tag{4.3}$$

Dataset 2. We deal with two functional random variables $Y_h(s)$, $1 \leq h \leq 2$ projected on the Fourier basis, with finite terms, with Gaussian amplitudes, and a 100 samples balanced functional dataset:

$$Y_h(s) = B_{h0}\phi_0 + B_{h1}\phi_1 + \ldots + B_{hn}\phi_n, \qquad 1 \leq h \leq 2, \tag{4.4}$$

are the functional random variables, and

$$y_{hi}(s) = b_{h0i}\phi_0 + b_{h1i}\phi_1 + \ldots + b_{hni}\phi_n, \qquad 1 \leq h \leq 2,\ 1 \leq i \leq 50 \tag{4.5}$$

are their realizations, with

$$\phi_0(s) = 1,\ \ \phi_{2k-1}(s) = \sin\left(\frac{2\pi ks}{\mathcal{T}}\right),\ \ \phi_{2k}(s) = \cos\left(\frac{2\pi ks}{\mathcal{T}}\right), \qquad 1 \leq k \leq n/2.$$

Again $\mathbf{B}_h \sim \mathcal{N}_{k+1}(\mu_h, \mathbf{\Sigma})$ and $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ is the same for the two groups. Also in this case there is a close formula for the j-th derivative; in particular, denoting with $\mod(r,s)$ the rest of the integer division between r and s, we have:

$$D^j Y_h(s) = (2\pi k)^j \begin{cases} \sum_{k=1}^{n/2} B_{h,2k-1}\phi_{2k-1} + B_{h,2k}\phi_{2k}, & \mod(k,4) = 0 \\ \sum_{k=1}^{n/2} B_{h,2k-1}\phi_{2k} - B_{h,2k}\phi_{2k-1}, & \mod(k,4) = 1 \\ \sum_{k=1}^{n/2} -B_{h,2k-1}\phi_{2k-1} - B_{h,2k}\phi_{2k}, & \mod(k,4) = 2 \\ \sum_{k=1}^{n/2} -B_{h,2k-1}\phi_{2k} + B_{h,2k}\phi_{2(k-1)}, & \mod(k,4) = 3 \end{cases} \quad j \geq 1 \tag{4.6}$$

Observe that we are normalizing the independent variable s, so that changes in the measure unit (and consequently in the length of $\mathcal{T}$) don't vary the information brought by a function, leaving the same shape it has for $\mathcal{T} = (0,1)$.
In our analysis we vary the following quantities: mean (μ_h) and variance (σ^2) of the coefficients, measure unit of the independent variable ($[s]$), value of the tuning parameter (c).

4.2 Simulations Without Corrections

Here we report the results of the primary and the secondary weighted Sobolev semimetrics obtained according to methodologies described in Section 3.1. We analyze the

Table 4.1: Instances of Dataset 1, varying the mean and the variance of the coefficients.

instance	mean vectors	variances	$\frac{\sigma_{\mathbf{min}}}{\lvert\boldsymbol{\Delta}\mu_{\mathbf{max}}\rvert}$
$\mathbf{POL_1}$	$\mu_1 = (2, 1, 0, 3, 2, 1.8, -0.5, 2, 1.5, 1, 1)$ $\mu_2 = (2.1, 0.9, -0.1, 3.1, 2.1, 1.7, -0.6, 1.9, 1.6, 0.9, 1.1)$	10^{-3}, 10^{-2}	30%
$\mathbf{POL_2}$	$\mu_1 = \mathbf{1}'_{11}$ $\mu_2 = 3 * \mathbf{1}'_{11}$	0.2, 1, 1.5	22%
$\mathbf{POL_3}$	$\mu_1 = (2, 1, 0, 3, 2, 1.8, 0, 0, 0, 0, 0)$ $\mu_2 = (2.1, 0.9, -0.1, 3.1, 2.1, 1.7, 0, 0, 0, 0, 0)$	10^{-12}, 10^{-11} 10^{-10}, 10^{-1}	10^{-5}
$\mathbf{POL_4}$	$\mu_1 = (\mathbf{1}'_6, \mathbf{0}'_5)$ $\mu_2 = (3 * \mathbf{1}'_6, \mathbf{0}'_5)$	10^{-10}, 10^{-9} 10^{-8}, 10^{-1}	$5 * 10^{-6}$

goodness of the results through quantitative and qualitative comparisons with the not sparse Sobolev and the L^2 square norms. In particular we build the *confusion matrix* for the output of the hierarchical clustering method implemented on the matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ provided by the optimization algorithms and we count the percentage of misclassified functions; image-plot of the matrices and the function `ColorDendrogram` available in the `R`-package `sparcl` (see Tibshirani and Witten (2013)), allows us a quick insight of the result.

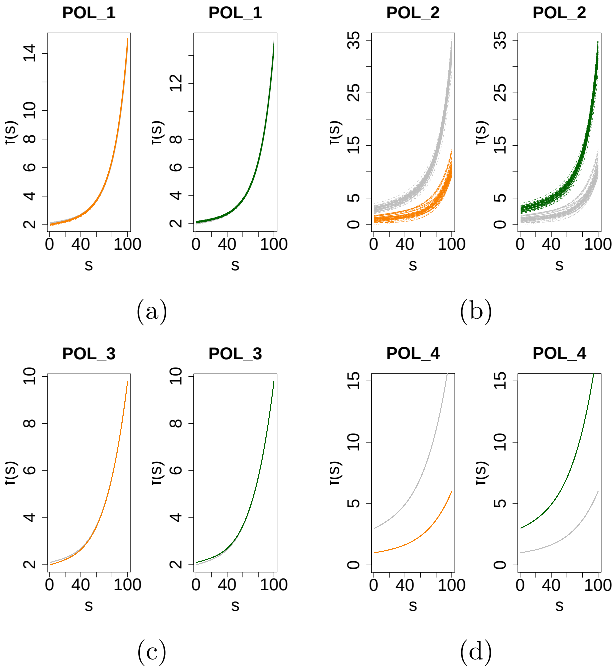
4.2.1 Polynomial Dataset

We are going to analyze four instances of Dataset 1, as in Table 4.1. They are effectively 10-degree polynomial functions, since we impose variability also on null coefficients. Remember that we compare the (square) norms of the differences of the functions, so the various cases correspond to the following situations. An index which explains how robust is our method is the ratio $CV = \sigma_{min}/|\Delta\mu_{max}|$, where σ_{min} is the the first lower standard deviation that we considered acceptable in terms of the error committed by a hierarchical clustering and dissimilarities build with the standard Sobolev norm, and $|\Delta\mu_{max}|$ is the larger difference between the mean coefficients.

$\mathbf{POL_1}$: the differences between the functions represent noise with mean's components in $\{-0.1, 0.1\}$. The method behave well also with quite high CV;

$\mathbf{POL_2}$: there is a shift between functions, that are well separated 10- degree polynomials. The method is quite robust;

$\mathbf{POL_3}$ **and** $\mathbf{POL_4}$: in in our intents these last two instances correspond to 5-degree polynomial functions with noise (the polynomials are mixed in the first case and well separated in the second). We are aware that there are better models



(a)

(b)

(c)

(d)

Figure 4.1: 10-degree overlapped and separated polynomials; realizations of $X_1(s)$ are in orange, while those of $X_2(s)$ are in green (a), (b); 5-degree overlapped and separated polynomials, with 6 to 10-degree polynomial noise; realizations of $X_1(s)$ are in orange, while those of $X_2(s)$ are in green (c), (d).

representing this situation, such as B- splines, and that no one would really use it, but we were interested in seeing how much robust our procedure is. We anticipate that the last two situations are those that perform poorly. The decision of putting differences on the first terms is due to the fact that we know what we should expect from a good method: high weights w_j, $0 \leq j \leq n$, on the lower derivatives and low weights on the higher ones. In a symmetrical problem with differences on the high-degree terms, we don't expect a symmetric behavior, since also the first derivatives are generally relevant.

In our simulations we estimate numerically the integrals implicit in the norm symbol withe the rectangle method and a stable grid. Figure 4.1 shows the four cases for $\mathcal{T} = (0, 1)$, a 100-points grid, and the lowest variances taken into consideration in each case: increasing the variances the functions get closer (and do overlap in the instances POL$_1$ and POL$_3$).

The results of the simulations relative to $\mathcal{T} = (0, 1)$ are summarized in Table 4.2. We use as features all the 11 nonzero derivatives, even if in the applications one rarely goes beyond the $4-$th, always with the aim of testing robustness. Observe that this cause that, when considering the sparse Sobolev metric, the admissible range for the tuning parameter is $1 \leq c \leq \sqrt{11} = 3.31$: we do computations for the values

$\mathbf{c} = (1.01,\, 1.5,\, 2,\, 2.5,\, \sqrt{11})$, and observe which perform better. We leave to future work the implementation of the optimal tuning parameter selection, as explained in Subsection 3.3.1. The errors represent the percentage of fibers misclassified by hierarchical clustering algorithm performed using Ward's between cluster dissimilarity, which usually behaves better than those explained in the first part of this work (for details see Fionn and Legendre (2011)). Errors are relative to the following metrics:

- Sobolev: the dissimilarities between functions are computed weighting in the same way all the submatrices contained in $\mathbf{D}$, i.e. all the derivatives;

- L_2: the dissimilarities are computed with $w_0 = 1$ and $w_j = 0$, $1 \leq j \leq k$.;

- Sparse Sobolev (primary and secondary), varying the tuning parameter we obtain more or less sparsity: in Table 4.3 are represented the nonzero weights w_j, $0 \leq j \leq k$, for $c \to 1$, and the significantly nonzero weights for $c \to \sqrt{11}$; we compare them with the features that are visually significant, detected trough inspection of image plots shown in the Appendix D. With *all* we mean that the image plots reveal differences in all the derivatives (and in particular in those indicated), while *null* means that no derivative allows clear detection of clusters.

The first thing to observe is that there is not an ordering on the performance of the three methods. As one might expect, increasing variability makes all the methods work worse. The quite surprising thing is that varying the tuning parameter has not so a grate effect on the sparse Sobolev metric, at least in these simulations in which no correction has been made. According to us, the reason is the same for which, even if more information is available in the Sobolev and in the sparse Sobolev metrics with respect to the L_2 metric, due to the derivative terms, the latter almost always has better performance. We can summarize our observations as follows.

1. Let's start comparing Sobolev with L_2 metrics. The lower error produced using L^2 metric in the instances POL$_2$ and POL$_4$ is consistent with the fact that functions do not overlap, but probably their derivatives (surely the high order ones) do. Even the image plot reported in figures D.2 and D.4 in the Appendix D confirm that giving a nonzero weights to terms different from $D^0, \ldots, D^6$ increases the error, since they add noise to the dissimilarities on which the clustering bases. On the contrary when the natural variability between functions could be observed on high derivatives, as in POL$_1$, the Sobolev metric perform better than L_2, at least for not too high variance, when noise dominates. The instance POL$_3$ is an almost degenerate input for every method, since we valuate the differences between functions that has similar means and variances in all the polynomial degrees. Only with very low variances the methods present acceptable errors, even if decreasing continuously variances makes the errors to decrease in a discontinue

Table 4.2: HC errors on instances of Dataset 1, varying the metric, σ^2 and c, without corrections, for $\mathcal{T} = (0, 1)$.

		$\mathbf{POL_1}$(%)				$\mathbf{POL_2}$(%)			
	$\sigma^2 =$	$\mathbf{10^{-3}}$	$\mathbf{10^{-2}}$			**0.2**	**1**	**1.5**	
Sobolev		7	65			0	16	17	
$\mathbf{L_2}$		37	40			0	1	2	
Sp.Sob.1	$c = 1.01$	7	29			7	16	23	
	$c \neq 1.01$	7	64			7	16	23	
Sp.Sob.2	$\forall c$	52	44			17	24	24	
		$\mathbf{POL_3}$(%)				$\mathbf{POL_4}$(%)			
	$\sigma^2 =$	$\mathbf{10^{-12}}$	$\mathbf{10^{-11}}$	$\mathbf{10^{-10}}$	**0.1**	$\mathbf{10^{-10}}$	$\mathbf{10^{-9}}$	$\mathbf{10^{-8}}$	**0.1**
Sobolev		0	45	48	48	0	0	56	48
$\mathbf{L_2}$		0	0	0	39	0	0	0	0
Sp.Sob.1	$c = 1.01$	0	45	45	45	0	45	45	45
	$c \neq 1.01$	0	48	47	45	0	56	47	47
Sp.Sob.2	$\forall c$	50	51	51	51	49	0	51	51

way. This instability of the Sobolev metric could be observed also in instance POL$_4$, and a reason could be the too high polynomial degree, that make the error dominate above a certain variance threshold and be not significant under it.

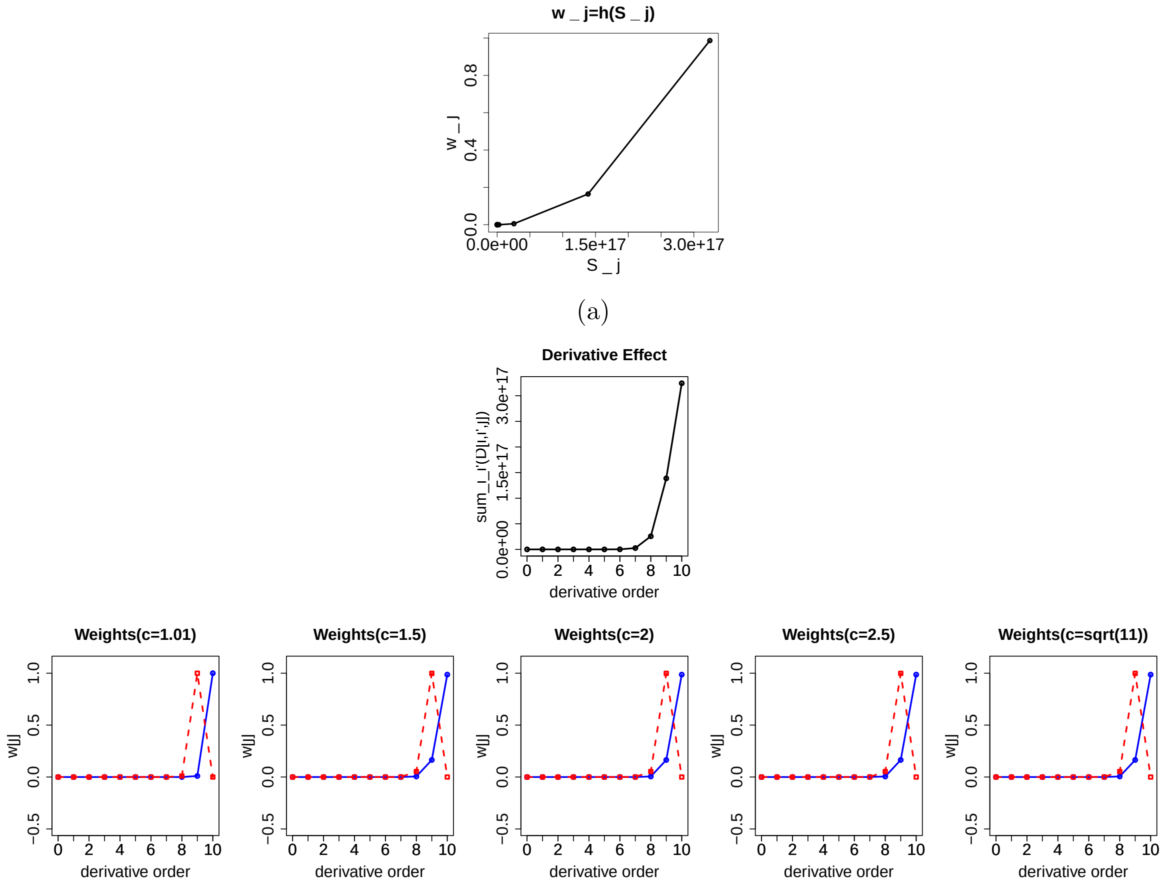
2. When in Section 3.1 we defined *optimality* for the sparse Sobolev metric, we intended to build a metric able to behave in an adaptive way, being, for example, more like L_2 in instances POL$_2$ and POL$_4$ and more like Sobolev in instance POL$_1$. However the simulations show that, fixed a scenario, the metric seems to give always the same weights to the same terms. We think that it is due to what we call *derivative effect*: the not homogeneity in magnitude between different derivatives that we are going to explain, and that justifies the two corrections we propose. As confirmed by the second step of the iterative Algorithm 3, we can empirically observe that, at least when $c \to \sqrt{p}$, each weight w_j is in a strictly relationship with the quantity $S_J = \sum_{i,i'} d_{i,i',j}$, $0 \leq j \leq k$, that is the sum of each respective sub-matrix, composing $\mathbf{D}$, relatively to each derivative order. The main open issue we leave for future studies is to try to quantify this relationship, since we think that the second proposal of correction we do doesn't behave so well due to the approximation we have made on it. In particular, observing the

Table 4.3: Indexes of matrices composing **D**, where differentiation is observed through image plots; indexes of nonzero weights w_j for the primary and secondary sparse Sobolev metrics

		POL$_1$				**POL$_2$**			
	$\sigma^2 =$	$\mathbf{10^{-3}}$	$\mathbf{10^{-2}}$			**0.2**	**1**	**1.5**	
IP		7÷10	null			all, 0÷3	all, 0÷5	all, 0÷3	
Sp.Sob.1	$c = 1.01$	10	10			10	10	10	
	$c \neq 1.01$	9÷10	9÷10			9÷10	9÷10	9÷10	
Sp.Sob.2	$\forall c$	9	9			9	9	9	
		POL$_3$				**POL$_4$**			
	$\sigma^2 =$	$\mathbf{10^{-12}}$	$\mathbf{10^{-11}}$	$\mathbf{10^{-10}}$	**0.1**	$\mathbf{10^{-10}}$	$\mathbf{10^{-9}}$	$\mathbf{10^{-8}}$	**0.1**
IP		0÷5	0÷5	null	null	0÷5	0÷ 5	0÷ 5	0÷1
Sp.Sob.1	$c = 1.01$	6, 10	10	9÷10	9÷10	5	10	10	9÷10
	$c \neq 1.01$	6, 10	10	9÷10	9÷10	4÷5	5,10	9÷10	9÷10
Sp.Sob.2	$c = 1.01$	9÷10	9	9	9	10	4÷5	9	9
	$c \neq 1.01$	9÷10	9	9	9	9÷10	4÷5	9	9

function $w_j = h(S_j)$, we can note that h has a quadratic-like shape. Thus from now on we consider that $w_j = \alpha S_j^2$. The experimental facts on which we base our hypotheses are shown in Figure 4.2(a), where the relationship $w_j = h(S_j)$ is plotted for $c = \sqrt{11}$, and in Figure 4.2(b), where S_j and the weights of the primary and secondary sparse Sobolev metrics are presented. Both figures are relative to the POL$_2$ instance, with $\sigma^2 = 0.2$. Changing the variance σ^2 in the same instance, and changing instance, similar graphics are obtained, but with different slope (different α under our hypothesis).

Then we call *derivative effect* this relationship in conjunction to the fact that S_j is not always a faithful representation of the real variability between data along the features (derivative orders), as happens in instance POL$_2$, to which Figure 4.2 refers. Comparing the black line that represents S_j, we see high values on $S_{8:10}$, while image plot presented in Figure D.2 in the Appendix tells us that the derivative orders for which the greater between-groups differences can be observed are the $1th$ to the $5th$. The discrepancy is attributable to the not homogeneity between derivatives: first of all they have not the same measure unit, as we explain in details in Section 4.3, and secondly they have intrinsic different magnitudes. To fix the ideas, in this simple polynomial case, increasing the derivative order j, the multiplicative factor $k!/(k-j)!$ in expression (4.3) increases, while the power



(a)

(b)

Figure 4.2: POL$_2$, $\sigma^2 = 0.2$: (a) quadratic-like relationship between w_j and S_j, for $c = \sqrt{11}$; (b) *derivative effect*: black curve represent S_j, the blue sold lines are the weights of the primary metric and the red dashed those of the secondary metric, varying the tuning parameter.

term x^{k-j} decreases, and they vary with different rates. Image plots compare the relative values of dissimilarities $d_{i,i',j}$ for a fixed j, while the plot of S_j gives us a synthetic information about their absolute values. It is thus possible, as in this case, that some derivative order have high weight w_j even if they are not really responsible for differentiation.

3. Observe that if we do computations with a low variance (it is sufficient $\sigma^2 = 0.1$) in POL$_2$, Sobolev and sparse Sobolev metric do allow good classification, with null error; what changes are not the weights w_j, nor the shape of S_j, but the fact that in image plots we are able to identify block structures also for high derivative orders, as in Figure D.2, with $\sigma^2 = 0.2$
In other situations (such as in instances POL$_3$ and POL$_4$) there is instead a change in the shape of w_j and of S_j as shown in Figure 4.3(a), relative to $\sigma^2 = 10^{-12}$ and (b), relative to $\sigma^2 = 10^-10$. Compatibly the image plots in Figure D.3 of

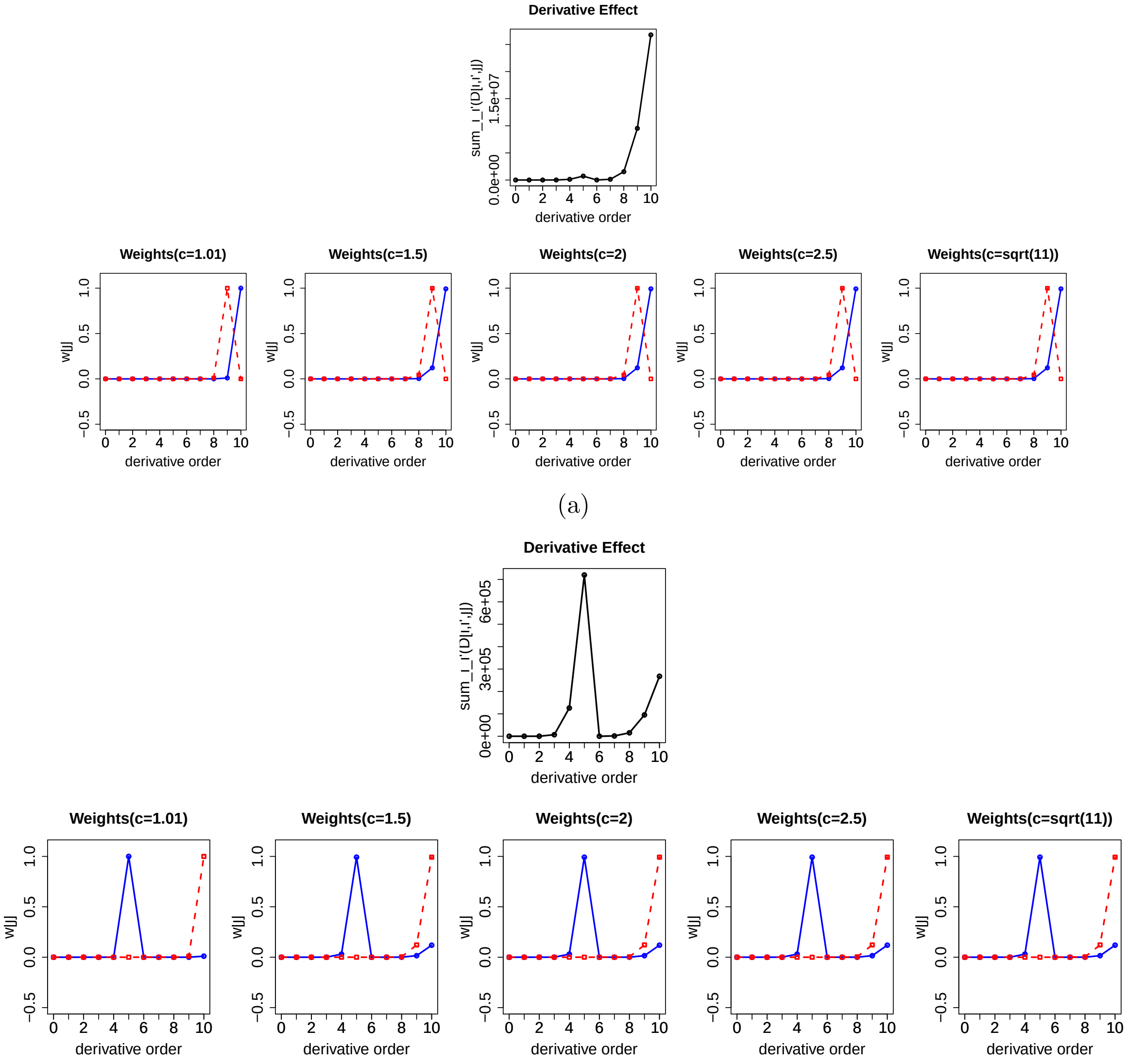
the two cases are similar, meaning that, in both of them, low derivatives counts toward classification. With high variance the method fails, while low variance let the method to recognize that the real variability lies in low derivative terms. As already stressed, the fact that this happens at so low variances and with discontinuity should be attribute to the high polynomial degree. At this point we can add to considerations regarding the Sobolev metric, that if σ^2 is so low that there is very few distinction between functions of the same group, also the Sobolev metric perform well (null error); even if it takes into consideration all the derivatives, the magnitude of the higher is null if compared to that of the lower ones, as we can see in the graphic of S_j, in figure 4.3.

4. The previous observations allow to guess that the stronger and more important of the corrections we are going to propose is the one aimed at override the intrinsic derivative effect, and not the one that avoid considering the unit measures, looking at adimensional phenomena. In fact even if making variance to decrease does not makes comparable the derivatives measure unit, both the Sobolev and the sparse Sobolev primary metrics allow good classification. This indicate that removing the dimensions is necessary for a good formulation of the problem, but not sufficient.

5. Note that another important aspect on which the corrections we are going to propose have impact is the convergence criterion used in the optimization algorithms. They stop iterating when at iteration r holds the following relation on the relative error:
$$\frac{\sum_{j=1}^{p} |w_j^r - w_j^{r-1}|}{\sum_{j=1}^{p} |w_j^{r-1}|} > \varepsilon$$
where ε is a sufficiently grate number. If weights w_j^r are affected by different measure units and intrinsic distortion, also this criterion is. It arises again the necessity of corrections.

6. We close this Section with considerations on the secondary metric, represented in red in the figures. As already stressed in Subsection 3.3.2, it is not a metric, if used on matrices orthogonal to $\mathbf{D}$, that is what we have done in computing the errors reported in Table 4.2. The number of functions misclassified is bigger with respect to the primary metric, but the intent of Tibshirani and Witten (2010) in the multivariate context was not to build a feature selection method with low errors, but capable of capturing *orthogonal* high-variance structures. This happens also in our functional framework, as we can see looking to typical dendrograms in output from the hierarchical clustering (see Figure 4.4, relative to instances POL_1, $\sigma^2 = 0.001$, POL_2, $\sigma^2 = 0.2$ and POL_4, $\sigma^2 = 10^{-10}$,).
We can see that the functions are well separated if we use the primary metric,

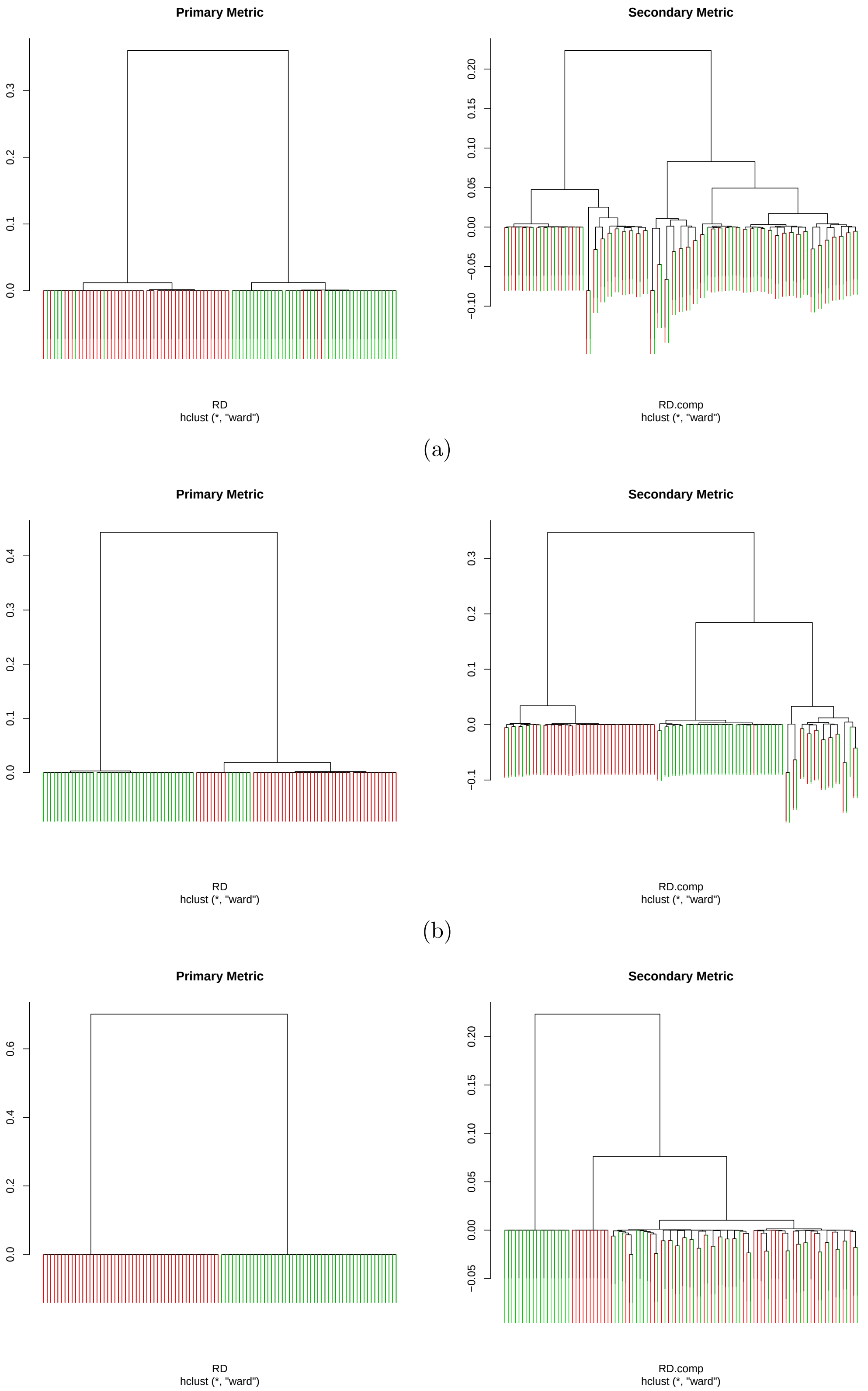
while they are mixed if we use the secondary. As we have anticipated, more studies should be done on the statistical meaning of the secondary metric, either if we use it on a matrix orthogonal to $\mathbf{D}$, or if we use it directly on $\mathbf{D}$. The only exception is POL_3, for $\sigma^2 = 10^{-9}$ when the secondary metric assumes the role that for lower variances will be covered by the primary metric, showing the presence of a double attractor for the solution of the iterative Algorithm 3.



(a)

(b)

Figure 4.3: Changes in the metric caused by changes in variability; (a) POL$_3$ instance with $\sigma^2 = 10^{-10}$; (b) POL$_3$ instance with $\sigma^2 = 10^{-12}$. Black curve represent S_j, the blue sold lines are the weights of the primary metric and the red dashed those of the secondary metric, varying the tuning parameter.



(a)

(b)

(c)

Figure 4.4: Dendrograms in output from hierarchical clustering, in instances POL_1, $\sigma^2 = 0.001$ (a), POL_2, $\sigma^2 = 0.2$ (b), and POL_4, $\sigma^2 = 10^{-10}$ (c).

4.2.2 Fourier Dataset

We present two instances of the Fourier Dataset 2. In this case we found more stability (if a stable grid is used for computations) then in the polynomial instances: low derivatives allow generally to well discriminate functions, even if there is some variability along high derivatives. Details of the instances are presented in Table 4.4. They are Fourier expansions truncated to the 10-th harmonic, with the same variance on terms with null and not null mean. Remember that in order to avoid *aliasing*, according to Shannon's theorem (see Shannon (1949)), it in necessary to work with a number of samples corresponding to a frequency that is at least twice the maximal frequency of the phenomenon. In our case $f_0 = 5$, $f_{max} = 10 \times 5 = 50$ since we have 10 harmonics, $2f_{max} = 100$ (we work under the hypothesis $|\mathcal{T}| = (0, 1)$, which mean seeing 5 periods). We have experimented however that the grid must be more dense (500 samples), if we want to achieve stability of the integrals approximating the norms (further increases in the number of samples do not change the results, fixed the remaining parameters).

FOU$_1$: the differences between functions of different groups correspond to the case of a Fourier expansion with not null mean on the low-frequency (the first 3) harmonics, and (statistically) null mean otherwise;

FOU$_2$: the differences are functions with both high and low (the first 3 and the last 3) frequency components.

Graphics of instance FOU$_1$, $\sigma^2 = 0.05$ and FOU$_1$, $\sigma^2 = 0.2$ are shown in Figure 4.5, while Table 4.5 contains the errors committed if we run hierarchical clustering algorithm using Sobolev, L_2, or sparse Sobolev (semi-)metrics. In these simulations we analyze differences between functions and their derivatives until the 4-th: $k = 4$, so admissible values for the tuning parameter are $1 \leq c \leq \sqrt{5} \simeq 2.23$. As for polynomial instances, we try a number of values: $c = (1.01, 1.4, 1.8, \sqrt{5})$; nevertheless one can see that without making corrections to the procedure, the choice of the tuning parameter is not so influential, due to the *implicit derivative effect.*
In the same table we indicate in brackets the positions for which we have significantly

Table 4.4: Instances of Dataset 2, varying the mean and the variance of the coefficients. Indices of vectors μ_h start from 0: $\mu_{h0} = 0$, $h = 1, 2$.

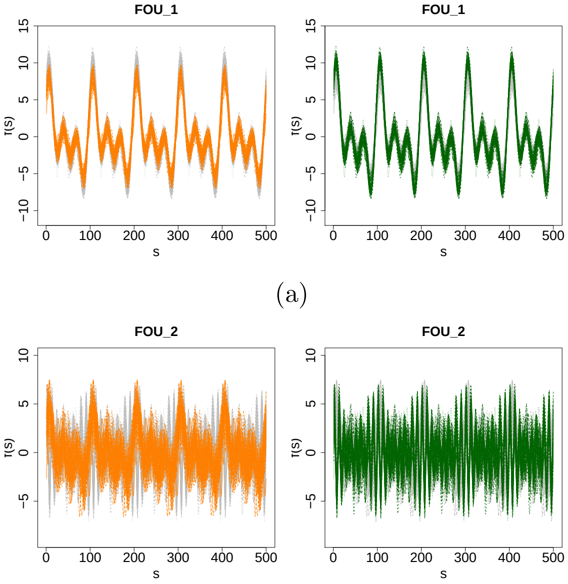
instance	nonzero elements of μ	variances
FOU$_1$	$\mu_{1,1\div6} = 2$ $\mu_{2,1\div6} = 2.5$	0.05, 0.1, 0.2, 0.5
FOU$_2$	$\mu_{1,1\div6} = 1$ $\mu_{2,16\div21} = 1$	0.2, 0.5, 1

nonzero weights for sparse Sobolev primary and secondary metrics. Increasing variance makes the methods commit higher errors and there is not a strict sorting. Observe that primary sparse Sobolev metric perform well both in instance FOU$_1$, where implicit derivative effect agrees with the natural differentiation along low derivatives, and in FOU$_2$, where it is able to capture more variability than L_2 (giving small weights also to derived functions) and than Sobolev (since all the derivatives allow differentiation, but the low one more).

Similar considerations to those done for polynomial instances holds about the quadratic relationship between w_j and S_j (Figure 4.6(a)) and about the role of the secondary metric (Figure 4.6(b) and 4.6(c)); both figures are relative to instance FOU$_2$, $\sigma^2 = 0.5$.

Table 4.5: HC errors on instances of Dataset 2, varying the metric, σ^2 and c, without corrections, for $\mathcal{T} = (0, 1)$. In brackets we indicate significantly nonzero weights for sparse Sobolev metrics, and those visually detected in image plots (IP).

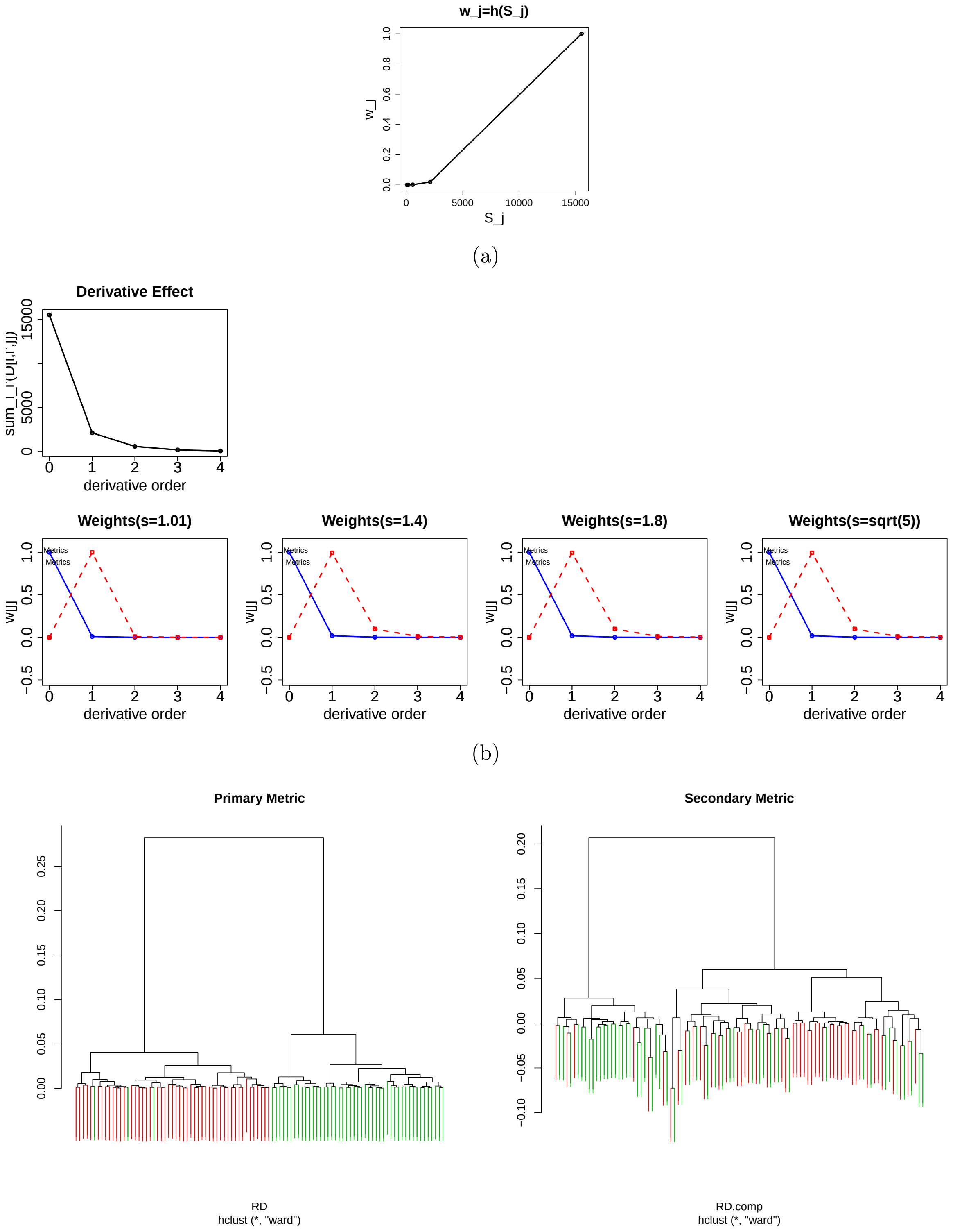
		FOU$_1$(%)				**FOU$_2$**(%)		
	$\sigma^2 =$	**0.05**	**0.1**	**0.2**	**0.5**	**0.2**	**0.5**	**1**
IP		(0)	(0)	(null)	(null)	(1:4)	(1:4)	(null)
Sobolev		1	14	34	56	0	9	23
L$_2$		0	8	40	43	0	4	31
Sp.Sob.1	$\forall c$	0 (0)	5 (0)	43 (0)	43 (0)	0 (0)	3 (0)	15 (0)
Sp.Sob.2	$c = 1.01$	52 (1)	53 (1)	53 (1)	45 (1)	70 (1)	31 (1)	30 (1)
	$c \neq 1.01$	52 (1+2)	53 (1+2)	53 (1+2)	45 (1+2)	70 (1+2)	31 (1+2)	30 (1+2)



(a)

(b)

Figure 4.5: 10-degree overlapped and separated fourier expansions: (a) low-frequency differences, (b) low and high frequency differences; realizations of $Y_1(s)$ are in orange, while those of $Y_2(s)$ are in green.



(a)

(b)

(c)

Figure 4.6: FOU$_2$, $\sigma^2 = 0.5$, no correction: (a) quadratic relationship between w_j and S_j; (b) sparse Sobolev primary and secondary metrics; (c) dendrograms obtained with them.

4.3 Π-Correction

The first correction we propose is concerned with a practical matter: units of measures. In the multivariate application proposed by Tibshirani and Witten (2010) such a need is not felt, since the features they treat are homogeneous (micro-array gene data). However in general applications, the quantities we are dealing with have a natural unit of measure, related to their physical meaning. The functional framework is very sensitive in this sense. To fix the ideas one can think that the independent variable s represents time (for example measured in seconds) and that the dependent functional variable $x(s)$ represents space (expressed for example in meters). Then we have that the unit of the first derivative is in $[D^1x(s)] = m^1s^{-1}$, and generally that of the j-th derivative is $[D^jx(s)] = m^1s^{-j}$, $0 \leq j \leq k$. This causes that every term of the Sobolev square norm $\|x(s)\|^2_{\mathcal{H}^k}$ has a different unit, that we can express as $\left[\|D^jx(s)\|^2_{L^2}\right] = m^2s^{1-2j}$, taking into consideration also the dimensionality introduced by integration. We are in general able to state the following proposition.

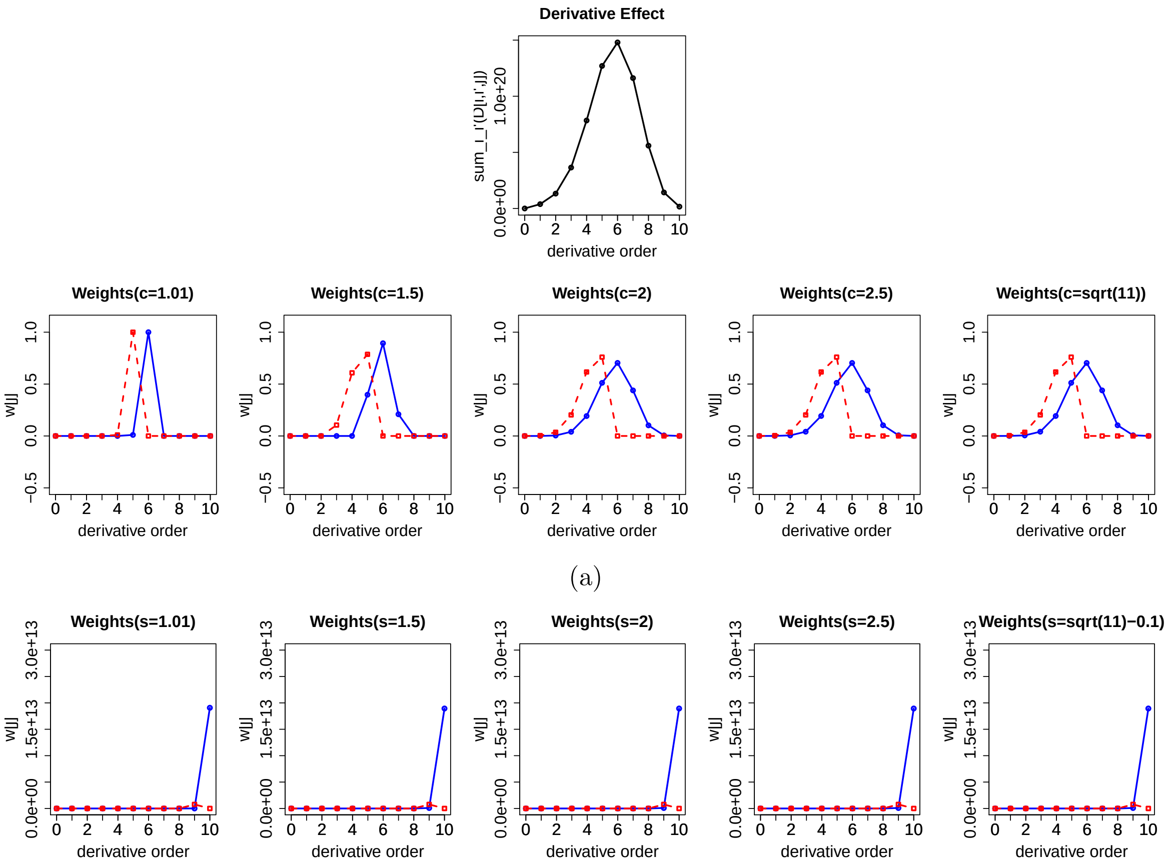
Proposition 1. *Let $[x(s)]$ denote the measure unit of the amplitude of functional data $x(s)$, and let $[s]$ denote the measure unit of the independent variable $s \in \mathcal{T}$. Then*

$$\left[\|D^jx_i(s) - D^jx_{i'}(s)\|^2_{L^2}\right] = [x(s)]^2[s]^{1-2j} \qquad 0 \leq j \leq k. \tag{4.7}$$

This means that if we want to perform a weighted summation of such terms through coefficients w_j in order to obtain the Sobolev weighted semi-norm $d^{\mathbf{w}}_{\mathcal{H}^k}(x_i(s), x_{i'}(s))$ (Definition 3), the weights w_j must have different units. In particular if the squared distance between functions $(d^{\mathbf{w}}_{\mathcal{H}^k})^2$ has unit $[x(s)]^2$, then each weight must have unit $[s]^{1-2j}, 0 \leq j \leq k$. This fact has two consequences: we are not able to compare directly the weights, since their values are conditioned by their units and, above all, the method is not invariant to change of measure unit. For example in Table 4.6 we show what happens to the L_2, Sobolev and the sparse Sobolev (with $c = 1.5$) metrics if we impose $\mathcal{T} = (0, 5)$, in the polynomial instances introduced in Subsection 4.2.1. The values have to be compared with those in Table 4.2, relative to $\mathcal{T} = (0, 1)$. This change in length of domain corresponds to varying the measure unit of the independent variable, without varying the information contained in the data, since we have normalized the functions dividing the independent variable with $\mathcal{T}$.

In some cases there is an improvement, in other a worsening of the error. What is important is that there is a change not justified, since the available information is not changed. Observe that also the weights are differently shaped from those when $\mathcal{T} = (0, 1)$, as shown for example in Figure 4.7(a), to be compared with Figure 4.2(b), that is relative to $\mathcal{T} = (0, 1)$, POL$_2$, $\sigma^2 = 0.2$.

We aim at correcting this dependence from measure unit with a procedure inspired to the Buckingham's theorem (also known as π-theorem) used especially in fluid dynamics theory (see Buckingham (1914)). Let us indicate with $\mathbf{w^a}$ the original vector of



(a)

(b)

Figure 4.7: Shape of the weights w_j^a (a) and shape of the weights w_j^b (b), for $|\mathcal{T}| = (0, 5)$, POL$_2$, $\sigma^2 = 0.2$. Black curve represent S_j, the blue sold lines are the weights of the primary metric and the red dashed those of the secondary metric, varying the tuning parameter.

weights obtained performing the optimization Algorithm 3 on the original data matrix $\mathbf{D}$. These weights have different measure unit. We pass to analyzing the phenomenon in a demensionless framework by transforming data contained in $\mathbf{D}$, obtaining a new input matrix $\tilde{\mathbf{D}}$, on which we run the algorithm. Call $\mathbf{D}_j$, $\forall 0 \leq j \leq k$ each $n \times n$ sub-matrix composing $\mathbf{D}_j$ and containing between functions dissimilarities relative to $j - th$ order derivatives. Call $\tilde{\mathbf{D}}_j$ the correspondent transformed matrices. It follows from Proposition 1 that the right transformation to perform is

$$\tilde{\mathbf{D}}_j = \frac{\mathbf{D}_j}{|\mathcal{T}|^{1-2j}} \quad \forall j = 0, \ldots, k..$$

As output we have a new vector of weights $\tilde{\mathbf{w}}$, that are no more sensitive to changes of unit measure. In particular they reveal to be numerically the same to $\mathbf{w^a}$ that we have if $\mathcal{T} = (0, 1)$. In general the relationship $\tilde{w_j} = w_j|\mathcal{T}|^{1-2j}$ holds. The vector $\mathbf{w}$ is generally different from $\mathbf{w^a}$, since the optimization process is not linear. Call this new

Table 4.6: HC errors on instances of Dataset 1, varying the metric, σ^2, for $c = 1.5$, without corrections, for $\mathcal{T} = (0, 5)$.

$\mathcal{T} = (\mathbf{0}, \mathbf{5})$		$\mathbf{POL_1}(\%)$				$\mathbf{POL_2}(\%)$			
	$\sigma^{\mathbf{2}} =$	$\mathbf{10^{-3}}$	$\mathbf{10^{-2}}$			**0.2**	**1**	**1.5**	
Sobolev		20	32			0	17	18	
$\mathbf{L_2}$		37	40			0	1	2	
Sp.Sob.1	$c = 1.5$	9	32			0	15	17	
$\mathcal{T} = (\mathbf{0}, \mathbf{5})$		$\mathbf{POL_3}(\%)$				$\mathbf{POL_4}(\%)$			
	$\sigma^{\mathbf{2}} =$	$\mathbf{10^{-12}}$	$\mathbf{10^{-11}}$	$\mathbf{10^{-10}}$	**0.1**	$\mathbf{10^{-10}}$	$\mathbf{10^{-9}}$	$\mathbf{10^{-8}}$	**0.1**
Sobolev		0	54	47	45	0	0	45	45
$\mathbf{L_2}$		0	0	0	39	0	0	0	0
Sp.Sob.1	$c = 1.5$	45	48	45	48	0	53	45	48

vector $\mathbf{w^b}$: if we want to go back to the dimensional world, we have to transform the weights $\tilde{\mathbf{w}}$, in order to obtain new weights $\mathbf{w^b}$. These last are such that the weighted semi-metric constructable with the pair $(\mathbf{w^b}, \mathbf{D})$ is the same constructable with $(\tilde{\mathbf{w}}, \tilde{\mathbf{D}})$. In particular:

$$w_j^b = \frac{\tilde{w_j}}{|\mathcal{T}|^{1-2j}}, \quad \forall j = 0, \ldots, k.$$

The following scheme shows synthetically the procedure we have explained.

Scheme 1. $\forall\ 1 \leq i, i' \leq n :$

$$\begin{aligned}
\left(d_{\mathcal{H}^k}^{\mathbf{w^a}}(x_i(s), x_{i'}(s))\right)^2 &= \sum_j \underbrace{w_j^a}_{[s]^{1-2j}} \underbrace{\left\|D^j x_i(s) - D^j x_{i'}(s)\right\|_{L^2}^2}_{[x(s)]^2[s]^{-(1-2j)}} \\
&\downarrow \\
\left(d_{\mathcal{H}^k}^{\tilde{\mathbf{w}}}(x_i(s), x_{i'}(s))\right)^2 &= \sum_j \underbrace{\tilde{w_j}}_{[1]} \underbrace{\frac{\left\|D^j x_i(s) - D^j x_{i'}(s)\right\|_{L^2}^2}{|\mathcal{T}|^{1-2j}}}_{[x(s)]^2} \\
&= \sum_j \frac{\tilde{w_j}}{|\mathcal{T}|^{1-2j}} \left\|D^j x_i(s) - D^j x_{i'}(s)\right\|_{L^2}^2 \\
&= \sum_j \underbrace{w_j^b}_{[s]^{-(1-2j)}} \left\|D^j x_i(s) - D^j x_{i'}(s)\right\|_{L^2}^2 \\
&= \left(d_{\mathcal{H}^k}^{\mathbf{w^b}}(x_i(s), x_{i'}(s))\right)^2.
\end{aligned}$$

Some relevant observations have to be made:

1. weights $\mathbf{w^b}$ do not respect the L_1 constraint ($\|\mathbf{w^b}\|_1 \nleq c$), since the optimization procedure isn't performed on them. We can see it in Figure 4.7(b), where it is given insight of what looks like vector $\mathbf{w^b}$ in instance POL$_2$, $\sigma^2 = 0.2$. Remember that $\tilde{\mathbf{w}}$ is the same as $\mathbf{w}^a$ when $\mathcal{T} = (0,1)$ and for instance POL$_2$ they are presented in Figure 4.2;

2. weights $\tilde{\mathbf{w}}$ are adimensional in order to let the squared weighted Sobolev semi-metric built with them to have dimension $[x(s)]^2$;

3. a priori (but rarely from the experimental point of view) we lose sparsity in weights $\mathbf{w^b}$, but we gain independence from measure unit. Both needs (sparsity and independence from measure unit) follows our concept of a *data-driven* Sobolev semi-metric, but the second is maybe a more necessary condition;

4. the Sobolev weighted semi-metric $\tilde{\mathbf{w}}$ is the correct one from a dimensional point of view, even if it is still dependent from the intrinsic derivative effect, as we have seen in Subsection 4.2.1; hence this first proposal is only necessary, but not sufficient in order to capture the real variability between functions.

4.4 Normalization Correction

In the present Section we propose a correction aimed at overriding what in Section 4.2 we called *implicit derivative effect*. Remember that with this term we indicate the fact that the real differentiation between functions can lie in derivative orders that have an absolute *magnitude* smaller than other derivative orders.

In particular, with *magnitude* we mean the absolute value of the summation of between functions dissimilarities along feature (derivative) j, $0 \leq j \leq k$: $S_j = \sum_{i,i'} d_{i,i',j}$. A qualitative insight of real between functions differentiation is obtainable looking at image plots of matrices $\mathbf{D}_j$, presented in the Appendix D for polynomial and Fourier instances. Remember that dark colors indicate low dissimilarities and light colors high ones: a block-like structure means groups differentiation. Than we have observed, especially in the polynomial dataset, that S_j has not always maximum and high values in correspondence of the j-th level for which the block-like structure is sharper in image plots. The arising problem, requiring a correction to the method proposed until now, is that weights w_j strictly follows the values of S_j, at least in the not sparse case $c \to \sqrt{k+1}$ (but decreasing c makes going to zero w_j values close to zero when c is big, without moving the position of the maximum). Note that this behavior is a peculiarity of datasets in which features are not homogeneous not only from a measure unit point of view, but especially from a magnitude point of view. Again, as for the first

correction, the work of Tibshirani and Witten (2010) is not affected by this problem as they treat homogeneous quantities, but the construction we are going to propose is valid also in the multivariate framework.
As anticipated, we make the following assumption on the relationship between w_j and S_j, for which an empirical justification was given in Section 4.2.

Assumption 3. Let $\alpha \in \mathbb{R}^+$, let $\mathbf{w} \in \mathbb{R}^{k+1}$ be the optimal vector of weights in output from optimization Algorithm 3, for $c \to \sqrt{k+1}$, and let $S_j = \sum_{i,i'} d_{i,i',j}$, $\forall\, 0 \leq j \leq k$. Then we assume that the following relation holds:

$$w_j = \alpha S_j^2. \tag{4.8}$$

We want to stress that the relationship is only approximate, and that the true one is the outcome of the iterative Algorithm 3, in which we use the threshold operator S. Remember that at each step we impose:

$$\mathbf{w} = \frac{S(\mathbf{a}, \Delta)}{\|S(\mathbf{a}, \Delta)\|_2}, \tag{4.9}$$

with $\mathbf{a} = \mathbf{D}'\mathbf{u}$, and $\Delta = 0$ if that results in $\|\mathbf{w}\|_1 \leq c$, otherwise $\Delta > 0$ is chosen to be a positive constant such that $\|\mathbf{w}\|_1 = c$.
According to us this is one of the reasons for which the correction we are going to propose doesn't always gain optimality (i.e. a performance of the weighted Sobolev semi-norm that is always equal or better than that of both the Sobolev and the L_2 one). Nevertheless an indication that we are on track is the fact that, as we show both in synthetic and in CLASH datasets, it never provide the bigger error. We leave to future work the task of exploring if improvements in understanding the relationship can be made. As we are going to show, there is also a computational failure of the algorithm, that should be corrected, or, at least, studied more deeply.

Our idea is to make the changing in sparsity parameter let to pass from a situation in which components of $\mathbf{w}$ are equal to each other, if $c \to \sqrt{k+1}$, to a situation in which only some of the components are nonzero, if $c \to 1$. The first case correspond to an equivalence between sparse Sobolev semi-metric and Sobolev metric, while the second case should let emerge the derivative order(s) along which the functions really differentiate (e.g. making sparse Sobolev equivalent to L_2 metric if the difference is in not derived functions).
We show two attempts in doing this, the first non successful, due to a lack in the algorithm or in its implementation, the second works, but not optimally, due to a failure of the algorithm and to the approximation in Assumption 3. Both of them are based on the following Scheme 2. As in the π- correction, we pass from weights $\mathbf{w^a}$ relative to the not transformed input $\mathbf{D}$, to weights $\tilde{\mathbf{w}}$ relative to a transformed input $\tilde{\mathbf{D}}$ for Algorithm 3. In particular, if $\mathbf{D}_j$ are the $n \times n$ matrices composing $\mathbf{D}$, we

normalize each as

$$\tilde{\mathbf{D}}_j = \frac{\mathbf{D}_j}{S_j^m}, \qquad m \in \mathbb{R}, \, \forall j = 0, \dots k.$$

We conclude going back to new weights $w_j^b = \tilde{w}_j / S_j^m$, relative to input $\mathbf{D}$, that are such that the pair $(\tilde{\mathbf{w}}, \tilde{\mathbf{D}})$ furnishes the same weighted Sobolev semi-metric as $(\mathbf{w^b}, \mathbf{D})$.

Scheme 2. Let $c \to \sqrt{k+1}$. $\forall\, 1 \leq i, i' \leq n$, $\forall\, m \in \mathbb{R}$, for $S_j = \sum_{i,i'} d_{i,i',j}$, $\forall\, 0 \leq j \leq k$:

$$\begin{aligned}
\left(d_{\mathcal{H}^k}^{\mathbf{w^a}}(x_i(s), x_{i'}(s))\right)^2 &= \sum_j w_j^a \|D^j x_i(s) - D^j x_{i'}(s)\|_{L^2}^2 \qquad \overset{Ass.\,3}{\Longrightarrow} \quad w_j = \alpha S_j^2 \\
&\downarrow \\
\left(d_{\mathcal{H}^k}^{\tilde{\mathbf{w}}}(x_i(s), x_{i'}(s))\right)^2 &= \sum_j \tilde{w}_j \frac{\|D^j x_i(s) - D^j x_{i'}(s)\|_{L^2}^2}{S_j^m} \qquad \overset{Ass.\,3}{\Longrightarrow} \quad \tilde{w}_j = \alpha \tilde{S}_j^2 = \alpha \frac{S_j^2}{S_j^{2m}} \\
&= \sum_j \frac{\tilde{w}_j}{S_j^m} \|D^j x_i(s) - D^j x_{i'}(s)\|_{L^2}^2 \\
&= \sum_j w_j^b \|D^j x_i(s) - D^j x_{i'}(s)\|_{L^2}^2 \\
&= \left(d_{\mathcal{H}^k}^{\mathbf{w^b}}(x_i(s), x_{i'}(s))\right)^2.
\end{aligned}$$

It follows that $w_j^b = \dfrac{\tilde{w}_j}{S_j^m} = \alpha S_j^{2-3m}$.

One can think to realize our proposal of having weights equal to each other for $c \to \sqrt{k+1}$ in two ways. Or we impose $\tilde{w}_0 = \ldots = \tilde{w}_k = \alpha$, which means enforcing $m = 1$, and deriving the resulting w_j^b, or we impose $w_0^b = \ldots = w_k^b = \alpha$, which means enforcing $m = 2/3$.

Observe that the first correction ($m = 1$) has a further statistical meaning. Since, given a functional random variable X

$$E\left[\|X\|_{L^2}^2\right] = \int_\Omega \int_{\mathcal{T}} |X(\omega, s)|^2 ds \, \mathbb{P}(d\omega),$$

to less than an additive constant, $S_j^1 = \sum_{i,i'} d_{i,i',j} = \sum_{i,i'} \int_{\mathcal{T}} (D^j x_i(s) - D^j x_{i'}(s))^2 ds$ is something similar to an estimate of the variance of a random function representing the j-th derivative. Thus, dividing $\mathbf{D}_j$ with S_j^1 is how to operate a standardization. However, as already said, this operation cause the algorithm (or its implementation) to fail: we have an output that does not respect the L_1 constraint $\|\mathbf{w}\|_1 \leq c$ for $c \to 1$ (or in general, for low values of c). Remember that, as in the π-correction, differently from weights $\tilde{\mathbf{w}}$, weights $\mathbf{w^b}$ have not to respect the constraints. We suppose that this behavior can be due to the presence of unstable equilibria that arise just because, if m=1, then $\tilde{S}_j = \alpha$ and it is not set a direction toward which the solution $\mathbf{w}$ has to

Table 4.7: Statistical correction: weights $\tilde{\mathbf{w}}$ for $m = 1$, $c = 1.01$, POL$_2$, $\sigma^2 = 0.2$.

	$\tilde{w}_0$	$\tilde{w}_1$	$\tilde{w}_2$	$\tilde{w}_3$	$\tilde{w}_4$	$\tilde{w}_5$	$\tilde{w}_6$	$\tilde{w}_7$	$\tilde{w}_8$	$\tilde{w}_9$	$\tilde{w}_{10}$
POL$_1$	0.26	0.27	0.28	0.29	0.29	0.30	0.30	0.31	0.32	0.32	0.33

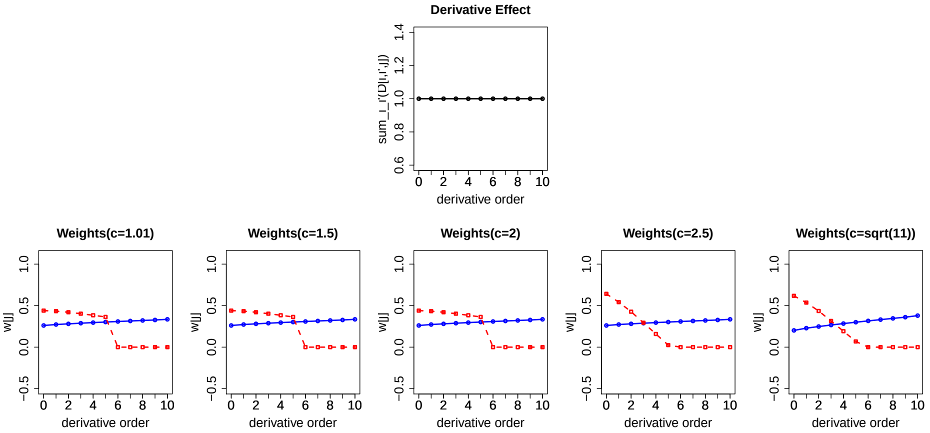


Figure 4.8: Normalization correction: weights $\tilde{\mathbf{w}}$ for $m = 1$, instance POL$_2$, $\sigma^2 = 0.2$. Black curve represent $\tilde{S}_j$, the blue sold lines are the weights of the primary metric and the red dashed those of the secondary metric, varying the tuning parameter.

move. We show in Figure 4.8 what happens in the polynomial instance POL$_2$, $\sigma^2 = 0.2$, but the same effect is present also in other instances, and in the multivariate framework if we standardize variables. Values of $\tilde{\mathbf{w}}$ for $c = 1.01$ are reported in Table 4.7. From now on we analyze the case $m = 2/3$, in the polynomial and in the Fourier instances.

4.4.1 Polynomial Dataset

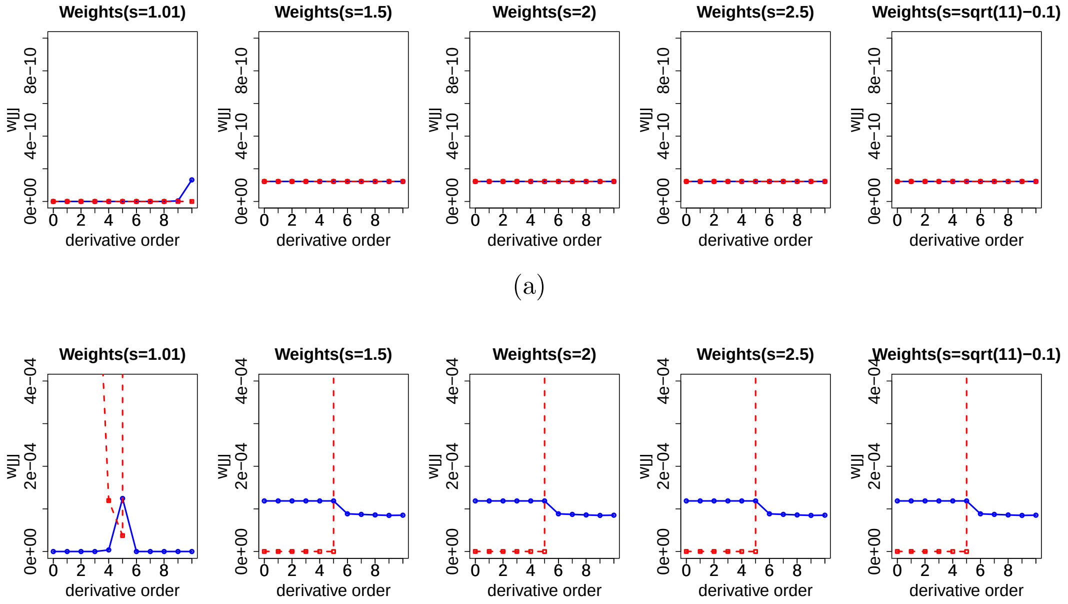
We analyze what happens to the four polynomial instances presented in Subsection 4.2.1 if we apply the normalization correction for $m = 2/3$ to sparse Sobolev semimetric. In Table 4.8 we report the errors gained varying the variances, in the case $\mathcal{T} = (0, 1)$. In order to have an idea of how do weights $\mathbf{w^b}$ are arranged (we would like to see if they respect the real variability responsible derivative levels presented in the image plots better than weights without correction $\mathbf{w^a}$), we plot them in Figures 4.10 and 4.11, in correspondence of the lower variance taken into consideration. Note the following facts.

1. As for π-correction, weights $\mathbf{w}^b$ have not to respect the L_1 constraint, even if in this case they do so, since the transformation we operate makes them very low. This reflects in the fact that the vector has not to be sparse for decreasing values of c. Therefore it would have been preferable to succeed in implementing the normalization correction with $m = 1$, if it had not caused the algorithm to fail. However we experimentally observe sparsity, for $m = 2/3$ and $c \to 1$, so we report the results obtained in this case, leaving space for future researches.

2. The arrangement of the weights $\mathbf{w^b}$ is globally similar to that obtained without correction ($\mathbf{w^a}$), but, unlike it, it doesn't follow strictly the shape of S_j for every c. The construction makes $\mathbf{w^b}$ to have constant values for $c \to \sqrt{k+1}$: this is respected quite well at least for small variances, as we can see in Figure 4.9, where we plot the instance POL$_2$ (analogous to POL$_1$) and POL$_3$ (analogous to POL$_4$) for $\sigma^2 = 10^-20$. Observe that in instance POL$_3$ the curve is not perfectly horizontal and we justify it with the fact that Assumption 3 is only an approximation.
   This let this corrected metric behave like the Sobolev one for high values of c and like the sparse Sobolev for low ones. It was our intent make it highlight the derivative orders really responsible of variability for $c \to 1$, but we didn't succeed in it, since the optimization process is performed on $\tilde{\mathbf{w}}$ and not on $\mathbf{w^b}$.
   However we have allowed more degrees of freedom, with respect to the not corrected metric, which makes the error almost always intermediate between the three metrics considered in Subsection 4.2.1. Our proposal of correction give worse results only in instance POL$_3$ for $\sigma^2 = 10^{-12}$: in our opinion it is due to the fact that it is a transitional value of variance for which even small differences in weights can make the difference.

Table 4.8: HC errors on instances of Dataset 1, with normalization correction ($m = 2/3$), for weighted Sobolev semi-metric, varying σ^2 and c, for $\mathcal{T} = (0, 1)$. * for $c = 2$, err = 29%; ** for $c = 2$, err = 49%.

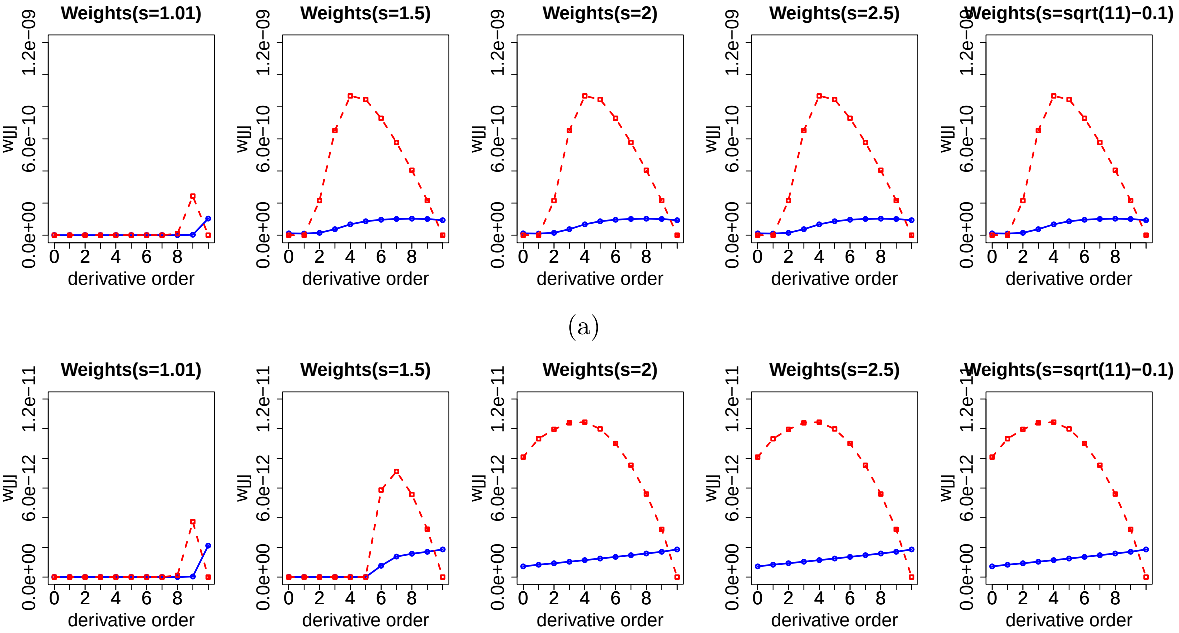
		$\mathbf{POL_1}$(%)				$\mathbf{POL_2}$(%)			
	$\sigma^2 =$	$\mathbf{10^{-3}}$	$\mathbf{10^{-2}}$			**0.2**	**1**	**1.5**	
Sp.Sob.1	$c = 1.01$	7	64			7	16	23	
	$c = 1.5$	7	28			0	16	17	
	$c \neq \{1.01, 1.5\}$	7	28^{*}			0	16	17	
Sp.Sob.2	$c = 1.01$	52	48			17	24	24	
	$c = 1.5$	52	48			17	24	24	
	$c \neq \{1.01, 1.5\}$	52	48^{**}			17	24	24	
		$\mathbf{POL_3}$(%)				$\mathbf{POL_4}$(%)			
	$\sigma^2 =$	$\mathbf{10^{-12}}$	$\mathbf{10^{-11}}$	$\mathbf{10^{-10}}$	**0.1**	$\mathbf{10^{-10}}$	$\mathbf{10^{-9}}$	$\mathbf{10^{-8}}$	**0.1**
Sp.Sob.1	$c = 1.01$	47	47	47	47	0	47	47	47
	$c = 1.5$	45	45	48	45	0	48	48	45
	$c \neq \{1.01, 1.5\}$	28	48	48	45	0	45	48	45
Sp.Sob.2	$c = 1.01$	0	51	51	51	49	0	51	51
	$c = 1.5$	0	0	50	50	50	0	55	50
	$c \neq \{1.01, 1.5\}$	0	0	50	50	50	0	48	50



(a)

(b)

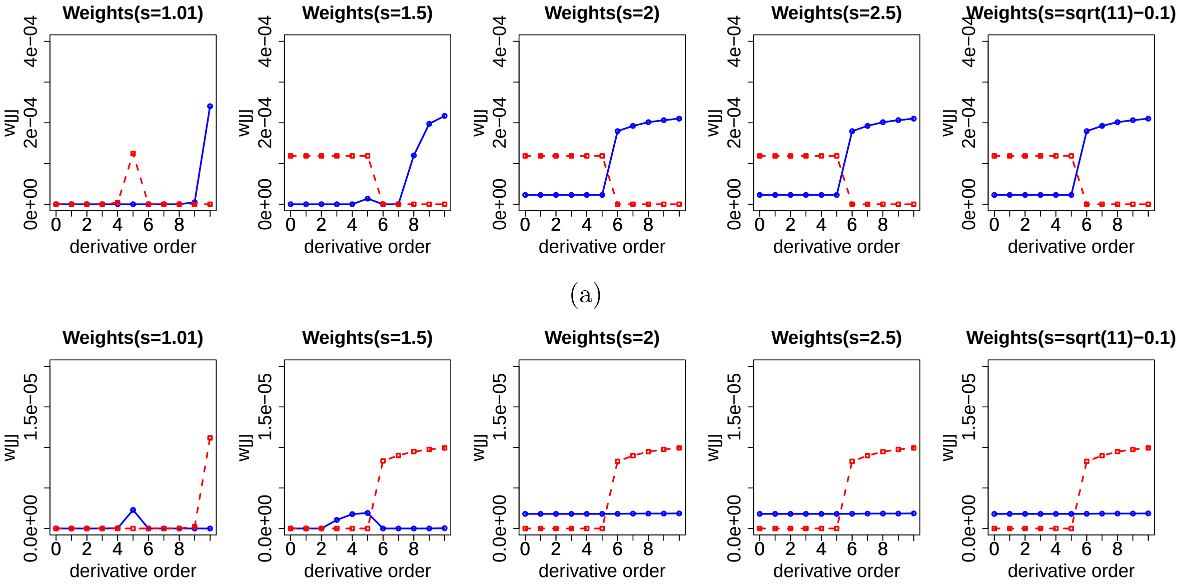
Figure 4.9: Statistical correction: shape of the weights w_j^b (a) POL$_2$,, (b) POL$_3$, $\sigma^2 = 10^{-20}$, for $|\mathcal{T}| = (0,1)$. The blue sold lines represents the weights of the primary metric and the red dashed those of the secondary metric, varying the tuning parameter.



(a)

(b)

Figure 4.10: Normalization correction: shape of the weights w_j^b (a) POL$_1$, $\sigma^2 = 0.001$, (b) POL$_2$, $\sigma^2 = 0.2$, for $|\mathcal{T}| = (0,1)$. The blue sold lines represents the weights of the primary metric and the red dashed those of the secondary metric, varying the tuning parameter.



(a)

(b)

Figure 4.11: Statistical correction: shape of the weights w_j^b (a) POL$_3$, $\sigma^2 = 10^{-12}$, (b) POL$_4$, $\sigma^2 = 10^{-10}$, for $|\mathcal{T}| = (0, 1)$. The blue sold lines represents the weights of the primary metric and the red dashed those of the secondary metric, varying the tuning parameter.

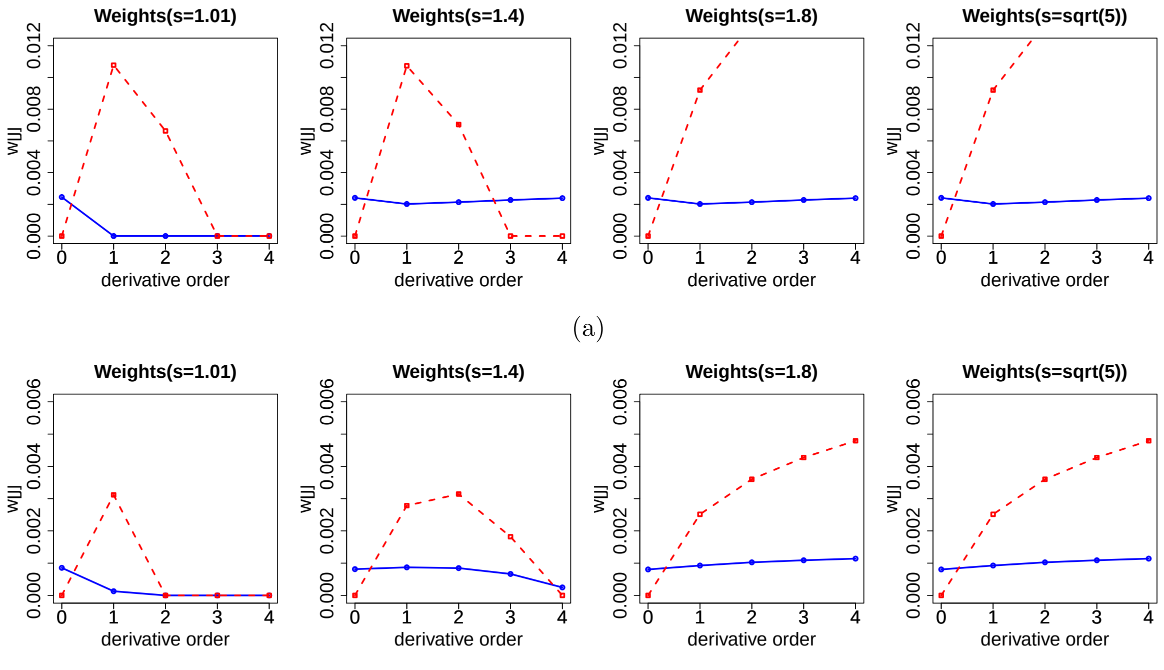
4.4.2 Fourier Dataset

Experimental observations feasible in this second synthetic dataset do confirm those obtained in the polynomial dataset. Note first of all that corrected sparse Sobolev semi-metric isn't forced to respect L_1 constraint ant to gain sparsity for $c \to 1$, but it does. And we can see that letting $c \to \sqrt{5}$ makes the corrected metric behave like the Sobolev, with some little variations from weights exactly constant, due to approximation implicit in Assumption 3. Since the order of magnitude of weights is low, the method isn't stable to these small changing (compare for example instance FOU_2, with $\sigma^2 = 1$, $c \geq 1.4$: the corrected metric generate an error of 15%, against the 23%error generated by the Sobolev metric).
Vice versa, if we let $c \to 1$, the corrected metric is like the sparse Sobolev one: the correction has generated an improvement, since we can choose to which of the two move. However, as for the polynomial case, we have not exactly reached our objective of making emerge the natural level of variability (for example high-order derivatives in some FOU_1 instances), along which functions differ the most. We leave this task to future works.

Observe finally that even here, increasing the variance of the coefficients makes things worse, which, jointly to sensibility to small values changes, can cause worse performances, compared to the two extreme metrics (see FOU_1, with $\sigma^2 = 0.5$, $c \geq 1.4$ with error 62%, versus Sobolev metric with error 56%).

Table 4.9: HC errors on instances of Dataset 2, with normalization correction ($m = 2/3$), for weighted Sovolev semi-metric, varying σ^2 and c, for $\mathcal{T} = (0, 1)$.

		$\mathbf{FOU_1}$(%)				**$\mathbf{FOU_2}$**(%)		
	$\sigma^2 =$	**0.05**	**0.1**	**0.2**	**0.5**	**0.2**	**0.5**	**1**
IP		(0)	(0)	(null)	(null)	(1:4)	(1:4)	(null)
Sp.Sob.1	$c = 1.01$	0	5	45	43	0	3	12
	$c \neq 1.01$	0	5	38	62	0	3	15
Sp.Sob.2	$c = 1.01$	50	49	53	45	70	31	30
	$c = 1.4$	50	53	47	49	34	33	32
	other c	48	48	52	49	69	65	62



(a)

(b)

Figure 4.12: Normalization correction: shape of the weights w_j^b (a) FOU$_1$, $\sigma^2 = 0.05$, (b) FOU$_2$, $\sigma^2 = 0.2$, for $|\mathcal{T}| = (0, 1)$. The blue sold lines represents the weights of the primary metric and the red dashed those of the secondary metric, varying the tuning parameter.

4.4.3 FIBER Dataset

In this Subection we analyze dataset FIBER explained in Chapter 2, where we have performed the functional principal component analysis. We build the sparse Sobolev metric and its second correction, based on Assumption 3. As in simulation examples, we make comparisons with results obtained using Sobolev and L^2 metrics, getting a dissimilarity matrix on which we run hierarchical clustering algorithm.

Remember that each fiber is represented by a function $r_i(s)$, $1 \leq i \leq n$, $n = 894$, that indicates the evolution of the radius along the curvilinear abscissa s; the number of materials groups taken into consideration can vary from 1 to 9: we fix two groups (wool and cashmere). This is equivalent to computing the error by cutting the dendrogram at a height at which two clusters could be seen, building a 2×2 confusion matrix and dividing the number of extra diagonal elements with n.

We valuate derivatives up to the third, thus $k + 1 = p = 4$ is the number of features, causing the tuning parameter c have limits $1 \leq c \leq 2$ (we have tested $c \in \{1, 1.3, 1.61.82\}$). In this example we have computed derivatives using the DFT representation of each fiber, described in equation (2.7) in Chapter 2:

$$r_i(s) = \sum_{k=1}^{p} \hat{C}_i(k) e^{j\frac{2\pi}{p}(k-1)(s-1)}, \qquad 1 \leq s \leq p.$$

Table 4.10: HC errors for FIBER dataset, varying the metric.

		2 groups(%)
Sobolev		40.15
$\mathbf{L^2}$		24.81
Sp.Sobolev.1	$\forall c$	40.26
Sp.Sobolev.2	$\forall c$	46.42
Sp.Sobolev.1.c	$\forall c$	40.26
Sp.Sobolev.2.c	$\forall c$	41.94

It can be proved that in this case the L^2 square norm of differences between h-th derivatives of functions $r_i(s)$ and $r_{i'}(s)$, $1 \le i,\ i' \le n$, $0 \le h \le k$, can be written in the following way:

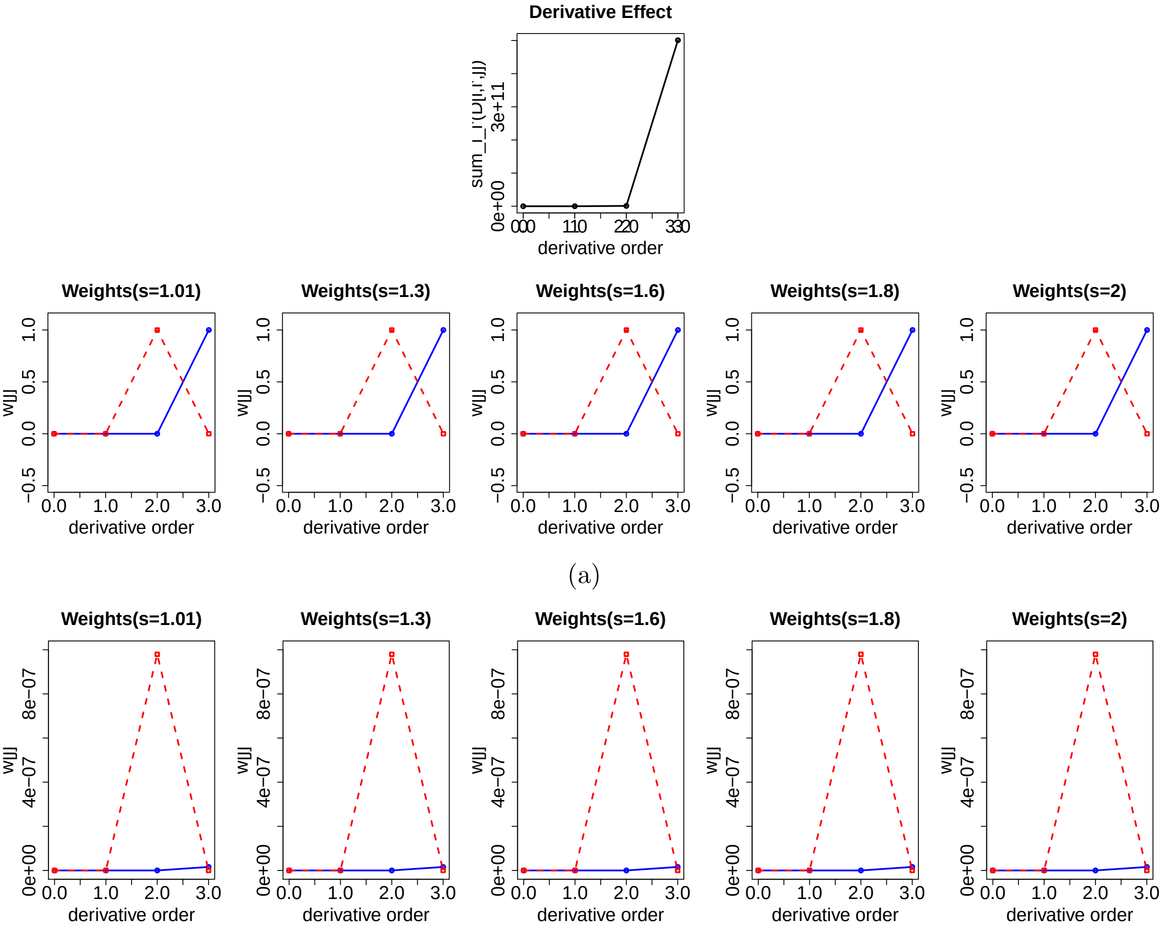
$$\|D^h r_i(s) - D^h r_i{}'(s)\|_2^2 = \left(\frac{2\pi}{p}\right)^{2h} \sum_{k=1}^{p} \hat{C}_i(k) e^{j\frac{2\pi}{p}(k-1)(s-1)}, \qquad 1 \le s \le p. \tag{4.10}$$

Numerical results are reported in Table 4.10, while Figure 4.13 gives an insight of the sparse Sobolev weights without correction (a) and with correction (b).

Three main observations can be made: first of all, errors committed using the primary metrics (without and with correction) are slightly lower than the one committed by the trivial classifier ($4/9 \simeq 44.4\%$), and remarkably lower only if we use L^2 metric. Qualitative matrices plot are not possible due to the high number of samples, but this fact means that more differences between functions can be seen along low derivatives than along higher derivatives orders.

Secondly sparse Sobolev metrics do not depend from the tuning parameter. From the optimization point of view this corresponds to the fact that the L^1-constraint is not active, because it is already satisfied for $c = 1$. As for the metrics without corrections this can be due to the high absolute value of higher derivatives; this reflects in the corrected metrics, that have the same qualitative trend and not an intermediate behavior between the Sobolev metric and the sparse not corrected metric, as we observed in simulation examples.

Thus, as third observation, we can infer that the correction proposed in Assumption 3 is not valid for these data, and more studies should be done. We leave it to future work, having provided in the meanwhile a quadratic classifier, derived from functional data analysis techniques, that performs the error obtained with multivariate analyses, for this industrial problem.



(a)

(b)

Figure 4.13: Sparse Sobolev weights for FIBER dataset, without correction (a) and with correction (b).

Chapter 5

Conclusions and Future Developments

In the first part of this work we have faced the problem of natural fibers classification trough data coming from bidimensional electronic microscope images, with the aim of implementing automatic morphological analysis for fraud situations detection. As far as we know only multivariate analysis have been performed until now and this is the first attempt in a functional framework.

Functional principal component analysis carried on radii of FIBER dataset, seen as functions of the curvilinear abscissa, let us obtain scores functions, defining the most important mode of variation of radii, subject to each mode being orthogonal to all modes defined on previous steps. It does not follow directly that a classifier built with scores (or an optimal selection of their valuations) have good performances, directions of maximal residual variability have not to coincide with those along which we see between-groups differences.

Nevertheless we have concluded that this happens in the FIBER dataset, by computing AER_{CV} estimates of quadratic discriminant analysis by varying both the number of groups and the features used in building the classifier. Since inspection of eigenfunctions of covariance function has indicated that scores meaning is similar to coefficients of radii DFT, we have compared scores, DFT modules, DFT log-modules (that better meet Normality assumption than the second), longitudinal mean radius and standard variation (the simpler indices used in multivariate analysis). All of them have an error below that of the trivial classifier, either we consider the existence of two or three groups (as suggested by hierarchical clustering on densities) and either the two groups are pure or the wool group is seen as a mixture of two sub-groups

The best result we have reached is obtained by combining a selection of *optimal* scores and a gaussian mixture model for the groups ($\text{AER}_{CV} = 18.23\%$). However, in the eventuality that in the industrial context it appears more immediate, even the application of DFT log-modules let us gain a good classifier, both in terms of Normality assumption and error (AER$_{CV}$ = 19.91%, while longitudinal mean radius and standar deviation gives ARE$_{CV}$ $\simeq$ 24.6%, for each choice on the number of groups). Given these positive results, more future work is encouraged in order to obtain even better perfomance:

- optimal scores have been selected through visual inspection, by looking at what of them caused a decrease in AER$_{CV}$; better features selection techniques could improve the classifier;
- Normality assumption is only a first attempt on the variables distribution; a different hypothesis would cause a use of more generic, but hopely more precise, variants of discriminant analysis;
- variables different from radius, such as scales height, perimeter and area, that in multivariate framework bring substantial improvement, could be studied in functional context, in order to see if they retain this trend;
- FPCA performed on radii derivatives, instead of functions themselves, might bring more information (consider e.g. that roughness of materials is connected to 2-nd derivative).

In the second part of the work we have dealt with a more theoretical issue: the importance of the choice of the metric in functional data analysis. A direction in which this modern branch of statistics could be addressed is the integration of this passage in the analytic process; our work has been a first approach to Sobolev-like semi-metrics.

We have attempted to extend the *hierarchical clustering* optimization algorithm proposed by Tibshirani and Witten (2010), relative to future selection in multivariate framework when $n >> p$. The innovation of our proposal has been changing the data on which the algorithm is run: we have replaced feature variables whit functions derivatives, without altering the structure of the optimization process based on KKT conditions.

The method gives in output a weight vector, whose components are referred each to a derivative order, and the between-functions dissimilarity $n \times n$ matrix built weighting by means of it, that allows running hierarchical clustering algorithm. Constraints on weights provide sparsity, so that decreasing the tuning parameter makes only some components non-zero: adaptivity of the method is reached if non-zero weights correspond to derivatives along which functions of different groups differentiate.

In this point we have experimented the influence of different nature of data with respect to the multivariate case, or better to homogeneity of the variables the for which the algorithm was originally thought. Studying empirically the relationship between weights and S_j (the summation of single Sobolev norm terms for differences between functions, relative to $j-$th derivative) we have noticed that they have a similar pattern, at least in the not-sparse case. This let us understand that j-th weight is defined according to the absolute value of S_j, which, in functional case, is likely not to significantly indicate if along that differentiation order exists between-functions variability. This phenomenon (*implicit derivative effect*) isn't present in micro-array gene data studied by Tibshirni and Witten.

We have analyzed two synthetic datasets (polynomial and a trigonometric) and FIBER data, noting sensibility to the fact that measure unit of derivatives changes when derivative order is changed, and to implicit derivative effect, both linked to not homogeneity in data. As a consequence we have proposed two corrections: the first provides independence from changes of measure unit of the independent variable; the second one is a normalization guess on which more work should be done in future. In particular:

- the authors of the iterative solving algorithm don't propose a close solution for weights: some theoretical studies should be made in order to understand if any property could be inferred by induction, in order to do more appropriate corrections;

- if, in our second proposal of correction, one chooses to standardize S_j, the algorithm (or its implementation) fails, not respecting the L_1 constraint: a deeper understanding of such a failure should be gained and an eventual correction to the algorithm should be made. It would be interesting (and necessary) verifying if the problem presents even in multivariate framework, when variables are not homogeneous and one decides to standardize;

- a unique proposal of correction should be made, bringing together both needs: measure units and implicit derivative effect.

Another area that allows future investigations is relative to the so called *secondary metric*, that isn't indeed a metric, but a index quantifying how much two functions differ, made naturally available by the method, and allowing hierarchical clustering. It provides orthogonality of final global dissimilarity matrices and something similar to orthogonality on weights with respect to the primary metric. Orthogonality of matrices is obtained weighting an input matrix different from that used for the primary metric. If we use secondary metric weights and the primary metric input matrix then in output we properly have dissimilarities in output; almost orthogonality between primary and secondary weights and nonnegativity constraint suggest that these new dissimilarities should select derivatives that hasn't been selected by the primary metric: studies of their properties and simulations appear necessary to us.

Since without correction the tuning parameter has little effect and the function for its selection by maximizing the gap statistic, provided in `sparcl` package isn't suitable for our data, we have made simulations for a number of values and chosen the best parameter *a posteriori*. We leave to future work the implementation of an automatic way for the parameter selection, according to this or any other criterion of optimality.

Finally it would be interesting trying to extend our work to other metrics than Sobolev-like metrics. Theoretically it sounds feasible, but we expect new not homogeneity of data to arise, requiring new corrections, in order to make this approach for metric selection able to really identify between functions variability.

Appendix A

FPCA Proofs

Here we formalize some statements made in the Chapter relative to functional principal component analysis.

Theorem A.1. *Given the random function $R(s)$ and the deterministic function $\phi(s) \in L^2(a,b)$, the inner product $\langle \phi, R-\mu \rangle = \int_a^b \phi(s)(R(s)-\mu(s))ds$ is a random function, representing the projection of $(R(s)-\mu(s))$ on $\phi(s)$, with mean function $E[\langle \phi, R-\mu \rangle] = \langle \phi, \mu-\mu \rangle = 0$ and covariance function $Var[\langle \phi, R-\mu \rangle] = \langle \phi, V_\Sigma \phi \rangle$*

Proof. As regards mean function:

$$\begin{aligned}
E\left[\langle \phi, R-\mu \rangle\right] &= E\left[\int_a^b \phi(s)(R(s)-\mu(s))ds\right] \\
&= \int_a^b E\left[\phi(s)(R(s)-\mu(s))ds\right] \\
&= 0.
\end{aligned}$$

As for the covariance function:

$$\begin{aligned}
Var[\langle \phi, R-\mu \rangle] &= E\left[(\langle \phi(R-\mu) \rangle)^2\right] \\
&= E\left[\int_a^b \phi(s)(R(s)-\mu(s))ds \int_a^b \phi(t)(R(t)-\mu(t))dt\right] \\
&= \int_a^b \int_a^b \phi(s)\Sigma(s,t)\phi(t)ds\,dt \\
&= \langle \phi, V_\Sigma \phi \rangle,
\end{aligned}$$

where V_Σ is defined as the *covariance operator*

$$\begin{aligned}
V_\Sigma : L^2(a,b) &\longrightarrow L^2(a,b) \\
\phi &\longmapsto V_\Sigma(t) = \int_a^b \Sigma(t,s)\phi(s)ds.
\end{aligned}$$

□

Theorem A.2. *Let* $R(s) = \mu + \sum_{k=1}^{+\infty} C_k\phi_k(s)$ *be the representation of a random function through its projections on* $\phi_k(s)$*, the eigenfunctions of the covarince function* $\Sigma(s,t)$*,* $a \leq s,\ t \leq b$*. Let* $\tilde{R}(s) = \mu + \sum_{k=1}^{m} C_k\phi_k(s)$ *be a random function expressing the truncation of such a representation to* m *terms. The following relation on the ratio between the variance of* $R(s)$ *that of* $\tilde{R}(s)$ *holds:*

$$\frac{E[\|\tilde{R}-\mu\|^2]}{E[\|R-\mu\|^2]} = \frac{\sum_{k=1}^{m}\lambda_k}{\sum_{k=1}^{+\infty}\lambda_k}. \tag{A.1}$$

Proof. Consider the variance of $R(s)$. By means of the spectral decomposition $\Sigma(t,s) = \sum_{k=1}^{+\infty}\lambda_k\phi_k(s)\phi_k(t)$, we can write:

$$\begin{aligned}
E\left[\|R-\mu\|^2\right] &= E\left[\int_a^b (R(s)-\mu(s))^2 ds\right] \\
&= \int_a^b E\left[(R(s)-\mu(s))^2\right] ds \\
&= \int_a^b \sum_{k=1}^{+\infty} \lambda_k\phi_k(s)\phi_k(s) ds \\
&= \sum_{k=1}^{+\infty}\lambda_k \underbrace{\int_a^b \phi_k(s)\phi_k(s)ds}_{=1,\ \phi_k \text{ orthonormal}} = \sum_{k=1}^{+\infty}\lambda_k.
\end{aligned}$$

Analogous considerations can be made on $E[\|\tilde{R}-\mu\|^2]$. □

Appendix B

Weak Derivative

Let $\mathcal{T}$ be an open set of $\mathbb{R}$ and let $f : \mathcal{T} \to \mathbb{R}$ be a Lebesgue measurable function. We denote by $C^\infty(\overline{\mathcal{T}})$ the vectorial space of the functions that are derivable in $(\overline{\mathcal{T}})$ an infinite number of times and with $C_0^\infty(\mathcal{T})$ its vector subspace consisting of functions that have a compact support contained in $\mathcal{T}$.

Definition B.1. We say that f is a *locally integrable function* in $\mathcal{T}$ (we write $f \in L^1_{loc}(\mathcal{T})$) if its Lebesgue integral is finite on all compact subsets K of $\mathcal{T}$:

$$\int_K |f| ds < \infty.$$

In particular if $f \in L^1_{loc}(\mathcal{T})$ and $\phi \in C_0^\infty(\mathcal{T})$, with $supp(\phi) = \{s \in \mathcal{T} s.t. \phi(s) \neq 0\}$ then $f\phi$ is integrable in $\mathcal{T}$ and

$$\int_{\mathcal{T}} f(s)\phi(s)ds = \int_{supp(\phi)} f(s)\phi(s)ds.$$

Definition B.2. Let $\alpha \in \mathbb{N}$ and $f \in L^1_{loc}(\mathcal{T})$. We say that a function $\nu \in L^1_{loc}(\mathcal{T})$ is the *weak derivative of* f (we write $\nu = D^\alpha f$) if the following relation holds

$$\int_{\mathcal{T}} \nu(s)\phi(s)ds = (-1)^\alpha \int_{\mathcal{T}} f(s) D^\alpha \phi(s) ds \tag{B.1}$$

for each $\phi \in C_0^\infty(\mathcal{T})$.

It can be demonstrated that the relation (B.1) univocally determines the weak derivative of a function.

Appendix C

Details on Sparse Feature Selection and Sparse Hierarchical Clustering

The considerations and proofs of theorems we set out in Section 3.1 come out from a general optimization framework, which is for example described in Hastie, Tibshirani, Witten (2009). Their construction is in turn founded on the necessary Karush-Kuhn-Tucker optimality conditions for constrained problems (see e.g. Nocedal and Wright (2006)). Thus we recall here the elements we need to motivate why the admissible value for the tuning parameter s is $1 < s < \sqrt{p}$, and to prove theorems (1), (2), (3).

C.1 Karush-Kuhn-Tucker Conditions

Let's start recalling the Karush-Kuhn-Tucker first order necessary conditions for optimality of a constrained problem. We write them in the case that all the constraints are inequality constraints and there are not set constraints $\{\mathbf{x} \in S\}$. We suppose moreover that f, $g_i \in C^1(\mathbb{R}^n)$. These hypotheses can be relaxed; we are assuming them since the problems we face are in the form

$$\begin{aligned} \underset{\mathbf{x}}{\text{maximize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0, \qquad i \in I = \{1, \dots, m\}, \\ & f,\, g_i \in C^1, \\ & \mathbf{x} \in \mathbb{R}^n. \end{aligned} \tag{C.1}$$

Suppose that the feasible region is not empty:

$$S = \{\mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) \leq 0, i \in I\} \neq \emptyset.$$

Definition C.1. Let $\overline{\mathbf{x}} \in S$ be a feasible point. We define the following sets:

1. *cone of feasible directions*

$$\mathcal{D}(\overline{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^n : \exists\, \overline{\alpha} > 0 \; s.t.\; \overline{\mathbf{x}} + \alpha \mathbf{d} \in S, \forall \alpha \in [0, \overline{\alpha}]\};$$

2. *set of indices of active constraints, i.e. of constraints valid with equality*

$$I(\overline{\mathbf{x}}) = \{i \in I : g_i(\overline{\mathbf{x}}) = 0\} \subseteq I;$$

3. *cone of directions limited by the directions of active constraints*

$$\mathrm{D}(\overline{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^n : \nabla'(g_i(\overline{\mathbf{x}}))\mathbf{d} \leq 0, \forall i \in I(\overline{\mathbf{x}})\}.$$

It is easy to show that $\mathcal{D}(\overline{\mathbf{x}})$ can be an open set and that $\overline{\mathcal{D}(\overline{\mathbf{x}})} \subseteq \mathrm{D}(\overline{\mathbf{x}})$, where $\overline{\mathcal{D}(\overline{\mathbf{x}})}$ denote the topological closure of $\mathcal{D}(\overline{\mathbf{x}})$. Then we are able to write a first theorem giving a necessary condition for optimality.

Theorem C.1. *Let $f \in C^1(S)$ and let $\mathbf{x}^* \in S$ be a local maximum of f on S. Then $\nabla'(f(\mathbf{x}^*))\mathbf{d} \geq 0$, $\forall \mathbf{d} \in \overline{\mathcal{D}(\mathbf{x}^*)}$, i.e. all feasible directions are ascendant.*

The problem is that this result is not easy to use, since it is difficult to characterize $\overline{\mathcal{D}(\overline{\mathbf{x}})}$. We give then an explicit name to the case when $\overline{\mathcal{D}(\overline{\mathbf{x}})} = \mathrm{D}(\overline{\mathbf{x}})$.

Definition C.2. Let $\overline{\mathbf{x}} \in S$ be a feasible point. Then we say that the *constraint qualification* (CQ) condition holds, if

$$\overline{\mathcal{D}(\overline{\mathbf{x}})} = \mathrm{D}(\overline{\mathbf{x}}).$$

In this framework we are able to write the Karush-Kuhn-Tucker optimality conditions.

Theorem C.2. *Let $f, g_i \in C^1(S)$ and assume the CQ holds in $\overline{\mathbf{x}} \in S$. If $\overline{\mathbf{x}}$ is a local maximum of f on S, then exists a vector $\lambda = (\lambda_1, \ldots, \lambda_m)' \geq 0$ (KKT multipliers), such that $\overline{\mathbf{x}}$ and $\mathbf{u}$ are the solution of the system*

$$\begin{cases} \nabla(f(\overline{\mathbf{x}})) - \sum_{i=1}^m \lambda_i \nabla g_i(\overline{\mathbf{x}}) = 0 \\ \lambda_i g_i(\overline{\mathbf{x}}) = 0, \; \forall i \in I \qquad\qquad (\lambda_i = 0, \; \forall i \in I \backslash I(\overline{\mathbf{x}})). \end{cases} \tag{C.2}$$

Notice that theorem (C.2) gives only necessary conditions and that we have nor established yet a way to verify the QC. The following definitions and theorems reveal helpful.

Definition C.3. Let $f : C \to \mathbb{R}$ be a real-valued function, with $C \in \mathbb{R}^n$.

- f is a *linear function* if

$$f(\delta_1 \mathbf{x}_1 + \delta_2 \mathbf{x}_2) = \delta_1 f(\mathbf{x}_1) + \delta_2 f(\mathbf{x}_2), \qquad \forall \mathbf{x}_1, \mathbf{x}_2 \in C, \quad \forall \delta_1, \delta_2 \in \mathbb{R}.$$

- f is a *convex function* if

$$f(\delta\mathbf{x}_1 + (1-\delta)\mathbf{x}_2) \leq \delta f(\mathbf{x}_1) + (1-\delta) f(\mathbf{x}_2), \qquad \forall \mathbf{x}_1, \mathbf{x}_2 \in C, \quad \forall \delta \in [0,1].$$

Theorem C.3. *The following conditions are equivalent for CQ to hold* $\forall \mathbf{x} \in S$:

- $g_i(\mathbf{x})$ *is a linear function* $\forall i \in I$;
- $g_i(\mathbf{x})$ *is a convex function and* $\exists\, \mathbf{a}$ *s.t.* $g_i(\mathbf{a}) < 0$, $\forall i \in I$.

Theorem C.4. *If (C.1) is a convex optimzation problem (convex objective function f and convex constraints g_i) and the CQ condition is respected, then $\mathbf{x}^*$ is a global optimum for f if and only if the KKT conditions (C.2) hold.*

C.2 Optimization Framework for Sparse Hierarchical Clustering

In this part we report a quite general framework for optimization proposed by Hastie, Tibshirani, Witten (2009). They present a penalized matrix decomposition (PMD) and derive a method for sparse principal components (SPC). Here we both recall the results relative to the general problem, without going into details of the PMD, and analyze its declinations that are useful for the proof of the theorems in Section 3.1.

Let $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{w} \in \mathbb{R}^p$ be two unit vectors and let $P_1\ , P_2 : \mathbb{R}^n \to \mathbb{R}$ be convex penalty functions. To fix the ideas we will often use *lasso* functions $P_j(\mathbf{u}) = \sum_{i=1}^n |u_i|,\ j = 1, 2$. Let $\mathbf{X}$ the $n \times p$ be the data matrix, as in Section 3.2. We call a PMD(P_1, P_2) problem the following:

$$\begin{aligned} \underset{\mathbf{u},\mathbf{w}}{\text{maximize}} \quad & \mathbf{u}'\mathbf{X}\mathbf{w} \\ \text{subject to} \quad & \|\mathbf{u}\|_2^2 \leq 1, \qquad P_1(\mathbf{u}) \leq c_1, \\ & \|\mathbf{w}\|_2^2 \leq 1, \qquad P_2(\mathbf{w}) \leq c_2. \end{aligned} \tag{C.3}$$

If we fix one of the variables the problem (C.3) is convex, which means that it is a biconvex problem. This suggests an iterative algorithm for optimizing it: at each iteration we maximize the objective function first with respect to $\mathbf{u}$ and then with respect to $\mathbf{w}$. For example the sub-problem relative to $\mathbf{w}$ takes the form:

$$\begin{aligned} \underset{\mathbf{w}}{\text{maximize}} \quad & \mathbf{u}'X\mathbf{w} \\ \text{subject to} \quad & \|\mathbf{w}\|_2^2 \leq 1, \qquad P_2(\mathbf{w}) \leq c_2. \end{aligned} \tag{C.4}$$

We find a global maximum imposing the KKT conditions that reveal necessary and sufficient, thanks to theorem (C.4). Then we write the algorithm for solving problem

(C.3) in the following way

```
Algorithm 5: PMD($P_1$, $P_2$)
  Data: $\mathbf{X}$;
  Result: $\mathbf{w}$, $\mathbf{u}$;
  Initialization: initialize $\mathbf{w}$ to have $L_2$-norm 1;
  while not convergence do
      1. $\mathbf{u} \leftarrow \underset{\mathbf{u}}{\text{argmax}}\, \mathbf{u}'X\mathbf{w}$; s.t. $P_1(\mathbf{u}) \leq c_1$ and $\|\mathbf{u}\|_2^2 \leq 1$;
      2. $\mathbf{w} \leftarrow \underset{\mathbf{w}}{\text{argmax}}\, \mathbf{u}'X\mathbf{w}$; s.t. $P_2(\mathbf{w}) \leq c_2$ and $\|\mathbf{w}\|_2^2 \leq 1$;
  end
```

In general Algorithm 5 does not converge to a global optimum for (C.3); however the authors refer that, according to empirical studies, it does converge to interpretable factors for appropriate choices of the penalty terms. As seen in Section 3.3 and in Section 4.1, this fact is confirmed by our analyses too, but only after making the necessary corrections. In the next subsections we write the KKT conditions for the two types of penalties necessary for our purposes, obtaining the PMD($\cdot$,L_1) problem, that is the PMD problem with no-penalty on $\mathbf{u}$ and *lasso* penalty on $\mathbf{w}$. We then show what does it mean writing a sequence of analogous problems with additional orthogonality constraints on $\mathbf{u}$.

C.2.1 PMD($\cdot$, L_1)

Consider the problem PMD($\cdot$, L_1) problem

$$\begin{aligned} \underset{\mathbf{u},\mathbf{w}}{\text{maximize}} \quad & \mathbf{u}'\mathbf{X}\mathbf{w} \\ \text{subject to} \quad & \|\mathbf{u}\|_2^2 \leq 1, \\ & \|\mathbf{w}\|_2^2 \leq 1, \qquad \|\mathbf{w}\|_1 \leq c. \end{aligned} \tag{C.5}$$

Observation C.1. We first notice that for both the constraints on $\mathbf{w}$ to be active it is necessary the condition $1 \leq c \leq \sqrt{p}$.

The reason for the observation is clear looking at Figure C.1, that shows the two-dimensional case ($p = 2$): the feasible region S is the one that is both inside the circle (L_2 constraint) and inside the square (L_1 constraint). The figure shows that in 2D, the points where both the L_1- and the L_2-constraints are active do not have either w_1 and w_2 equal to 0, except for the limit case $c = 1$. However, when $p > 2$, the dimension of $\mathbf{w}$ is at least 3; then the right panel of Figure C.1 can be thought of as

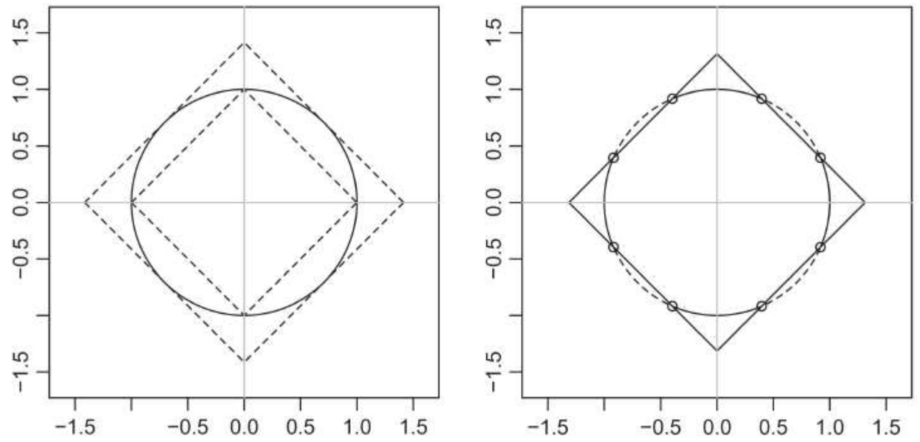


Figure C.1: A graphical representation of the L_1- and L_2-constraints on $\mathbf{u}$ in the PMD($\cdot$,L_1) criterion. Here $\mathbf{w}$ is a two-dimensional vector and the gray lines indicate the coordinates axes u_1 and u_2. Left: the L_2-constraint is the black solid circle. The constraints $\|\mathbf{w}\|_1 = 1$ and $\|\mathbf{w}\|_1 = \sqrt{p}$ are shown using dashed lines. Right: the L_1- and the L_1-constraints are shown for $c_1 = 1.2$. Small circles indicate the points where both the L_1- and the L_2-constraints are active. Along the solid arcs ate indicate the solutions that occur when $\Delta = 0$ in Algorithm 6, since the L_1-constraint is not active ($\|\mathbf{w}\|_1 < c$). But notice that there could be other feasible points in which holds $\Delta = 0$ and either $\|\mathbf{w}\|_1 = c$ (the solid lines) or $\|\mathbf{w}\|_1 < c$ (inner points).

the hyperplane $\{u_i = 0, \ \forall i > 2\}$. In this case, the small circles indicate regions where both constraints are active and the solution is sparse (since $u_i = 0$ for $i > 2$). This is the reason why, in Section 3.1, although we can't take a too high derivative order for numerical reasons, we have to consider always more than two derivatives as features.

As in the general framework in Appendix C.2, we can then split the problem C.5 into two sub-problems and solve them separately.

Theorem C.5. *Let* $\mathbf{a} = \mathbf{X}\mathbf{w}$. *The solution to the convex sub-problem*

$$\begin{aligned} &\underset{\mathbf{u}}{\textit{maximize}} && \mathbf{u}'\mathbf{a} \\ &\textit{subject to} && \|\mathbf{u}\|_2^2 \leq 1, \end{aligned} \tag{C.6}$$

is given by $\mathbf{u} = \dfrac{\mathbf{a}}{\|\mathbf{a}\|_2} = \dfrac{\mathbf{X}\,\mathbf{w}}{\|\mathbf{X}\,\mathbf{w}\|_2}$.

Proof. The KKT conditions (C.2) written for the problem (C.6) are

$$\begin{cases} \mathbf{a} - 2\lambda\mathbf{u} = 0 \\ \lambda(\|\mathbf{u}\|_2^2 - 1) = 0. \end{cases} \tag{C.7}$$

If $\lambda \neq 0$ then

$$\begin{cases} \mathbf{u} = \dfrac{\mathbf{a}}{2\lambda} \\ \|\mathbf{u}\|_2 = 1 \end{cases} \qquad \begin{cases} \mathbf{u} = \mathbf{a}/\|\mathbf{a}\|_2 \\ \lambda = \|\mathbf{a}\|_2/2. \end{cases}$$

The case $\lambda = 0$ is a feasible solution only if the problem (C.6) is trivial ($\mathbf{a} = \mathbf{0}$). □

Theorem C.6. *Let x^+ denotes the positive part of x ($x^+ = x$ if $x > 0$, $x^+ = 0$ if $x \leq 0$) and define the soft-thresholding operator $S(x, c) = \operatorname{sign}(x)(|x| - c)^+$. Let $\mathbf{b} = \mathbf{X}'\,\mathbf{u}$. Then the solution to the convex sub-problem*

$$\begin{aligned} &\underset{\mathbf{w}}{\textit{maximize}} \quad \mathbf{b}'\mathbf{w} \\ &\textit{subject to} \quad \|\mathbf{w}\|_2^2 \leq 1, \qquad \|\mathbf{w}\|_1 \leq c, \end{aligned} \tag{C.8}$$

is given by $\mathbf{w} = \dfrac{S(\mathbf{b}, \Delta)}{\|S(\mathbf{b}, \Delta)\|_2}$, where $\Delta = 0$ if that results in $\|\mathbf{w}\|_1 \leq c$; otherwise, $\Delta > 0$ is chosen to yield $\|\mathbf{w}\|_1 = c$.

Proof. Let's call λ and Δ the two KKT multipliers. The KKT conditions (C.2) written for the problem (C.8) are

$$\begin{cases} \mathbf{b} - 2\lambda\mathbf{w} - \Delta\boldsymbol{\Gamma} = \mathbf{0} \\ \lambda(\|\mathbf{w}\|_2^2 - 1) = 0 \\ \Delta(\|\mathbf{w}\|_1 - c) = 0, \end{cases} \tag{C.9}$$

where $\Gamma_i = \operatorname{sign}(w_i)$ if $w_i \neq 0$; otherwise $\Gamma_i \in [-1, 1]$.
If $\lambda \neq 0$ then we can write the first equation for each component of $\mathbf{w}$ as

$$b_i - 2\lambda w_i - \Delta \operatorname{sign}(w_i) = 0.$$

Considering that $w_i = \operatorname{sign}(w_i)|w_i|$, it follows that $\operatorname{sign}(b_i) = \operatorname{sign}(w_i)$, since the equation can be rewritten as

$$\operatorname{sign}(w_i)(2\lambda|w_i| + \Delta) = \operatorname{sign}(b_i)|b_i|,$$

where λ and Δ are nonnegative according to the KKT construction. Moreover this implies that $\lambda \neq 0$ ($\lambda > 0$) $\Leftrightarrow |b_i| - \Delta > 0$ and $\lambda = 0 \Leftrightarrow |b_i| - \Delta = 0$ Then the first equation written for each component of $\mathbf{w}$ becomes

$$\begin{cases} w_i = \dfrac{b_i - \Delta \operatorname{sign}(b_i)}{2\lambda}, & \text{if } |b_i| - \Delta > 0 \\ w_i = 0, & \text{if } |b_i| - \Delta = 0 \end{cases}$$

i.e.

$$\begin{cases} w_i = \dfrac{\operatorname{sign} b_i(|b_i| - \Delta)}{2\lambda}, & \text{if } |b_i| - \Delta > 0 \\ w_i = 0, & \text{if } |b_i| - \Delta \leq 0 \end{cases}$$

i.e. $w_i = \dfrac{S(b_i, \Delta)}{2\lambda}$. But if $\lambda \neq 0$, for the second equation in C.9 to hold, it must be $\|\mathbf{w}\|_2 = 1$, i.e. λ must be chosen so that $\mathbf{w} = \dfrac{S(\mathbf{b}, \Delta)}{\|S(\mathbf{b}, \Delta)\|_2}$.

The case $\lambda = 0$ can lead to a feasible or non feasible solution of the system (C.9) depending on the values of $\mathbf{b}$, Δ and $\mathbf{\Gamma}$, but we are unable to express an optimal value for $\mathbf{w}$; thus this case has not to be taken into consideration.
As regards Δ, either $\Delta = 0$ if this results in a feasible solution (if both $\Delta = 0$ and $\lambda = 0$, then the system is feasible only if the problem is trivial, with $\mathbf{b} = 0$), or Δ must be chosen such that $\|\mathbf{w}\|_1 = c$. So $\Delta = 0$ if this results in $\|\mathbf{w}\|_1 \leq c$; otherwise, we choose Δ such that $\|\mathbf{w}\|_1 = c$, implementing for example a Binary Search.

□

Let $\mathbf{a} = \mathbf{X}\mathbf{w}$ and $\mathbf{b} = \mathbf{X}'\,\mathbf{u}$. Consequently to theorems C.5 and C.6, we can write the following iterative algorithm for the solution of problem (C.5).

```
Algorithm 6: PMD(·, $L_1$)
  Data: $\mathbf{X}$;
  Result: $\mathbf{w}$, $\mathbf{u}$
  Initialization: initialize $\mathbf{w}$ to have $L_2$-norm 1;
  while not convergence do
      1. $\mathbf{u} \leftarrow \frac{\mathbf{a}}{\|\mathbf{a}\|_2}$;
      2. $\mathbf{w} \leftarrow \frac{S(\mathbf{b},\Delta)}{\|S(\mathbf{b},\Delta)\|_2}$, where $\Delta = 0$ if that results in $\|\mathbf{w}\|_1 \leq c$,
         otherwise $\Delta > 0$ is chosen to be a positive constant such that $\|\mathbf{w}\|_1 = c$;
  end
```

C.2.2 Complementary PMD(·, L_1)

Since the PMD problem is generally related to the approximation of the matrix $\mathbf{X}$ with a rank-k matrix (and the PMD(·, L_1) is only a particular rank-1 approximation) Hastie, Tibshirani, Witten (2009) study also the possibility of writing a number of problems similar to problem (C.5), with an additional constraint of orthogonality on $\mathbf{u}$. Notice that they do not require orthogonality on $\mathbf{w}$ but that the constraint on $\mathbf{u}$ give something similar to orthogonality also on $\mathbf{w}$.
So the complementary problems can be written in the form

$$\begin{aligned} \underset{\mathbf{u}_k,\mathbf{w}_k}{\text{maximize}} \quad & {\mathbf{u}_k}'\mathbf{X}\mathbf{w}_k \\ \text{subject to} \quad & \|\mathbf{u_k}\|_2^2 \leq 1, \qquad \mathbf{u}_k \perp \mathbf{u}_1, \ldots, \mathbf{u}_{k-1} \\ & \|\mathbf{w_k}\|_2^2 \leq 1, \qquad \|\mathbf{w}_k\|_1 \leq c. \end{aligned} \tag{C.10}$$

With $\mathbf{u}_k$ fixed, one can solve (C.10) for $\mathbf{w}_k$ easily, using theorem C.6. With $\mathbf{w}_k$ fixed, the problem is as follows

$$\begin{aligned} \underset{\mathbf{u}_k}{\text{maximize}} \quad & \mathbf{u}_k{}'\mathbf{X}\mathbf{w}_k \\ \text{subject to} \quad & \|\mathbf{u_k}\|_2^2 \le 1, \qquad \mathbf{u}_k \perp \mathbf{u}_1, \dots, \mathbf{u}_{k-1}. \end{aligned} \tag{C.11}$$

Let $\mathbf{U}_k^{\perp}$ denote an orthogonal basis that is orthogonal to $\mathbf{U}_{k-1}$, the matrix with columns $\mathbf{u}_1, \dots, \mathbf{u}_{k-1}$. It follows that $\mathbf{u}_k$ is in the column space of $\mathbf{U}_k^{\perp}$, and so can be written as a linear combination of the basis elements $\mathbf{u}_k = \mathbf{U}_{k-1}^{\perp}\theta$, with $\theta \in \mathbb{R}^{k-1}$. Note also that $\|\mathbf{u}_k\|_2 = \|\theta\|_2$. Thus problem C.11 can be written equivalently as

$$\begin{aligned} \underset{\theta}{\text{maximize}} \quad & \theta'\mathbf{U_{k-1}^{\perp}}{}'\mathbf{X}\mathbf{w}_k \\ \text{subject to} \quad & \|\theta\|_2^2 \le 1. \end{aligned} \tag{C.12}$$

and, according to theorem C.5, we find that the optimal θ is

$$\theta = \frac{\mathbf{U_{k-1}^{\perp}}{}'\mathbf{X}\mathbf{w}_k}{\|\mathbf{U_{k-1}^{\perp}}{}'\mathbf{X}\mathbf{w}_k\|_2}.$$

Therefore the value of $\mathbf{u}_k$ that solves (C.11) is

$$\mathbf{u}_k = \frac{\mathbf{U_{k-1}^{\perp}}\mathbf{U_{k-1}^{\perp}}{}'\mathbf{X}\mathbf{w}_k}{\|\mathbf{U_{k-1}^{\perp}}{}'\mathbf{X}\mathbf{w}_k\|_2} = \frac{(\mathbf{I} - \mathbf{U_{k-1}}\mathbf{U_{k-1}}{}')\mathbf{X}\mathbf{w}_k}{\|\mathbf{U_{k-1}^{\perp}}{}'\mathbf{X}\mathbf{w}_k\|_2}.$$

So we can use this update to develop an iterative algorithm to find multiple factors for the problem (C.5), the single PMD criterion, that yields orthogonal $\mathbf{u}_k$s. The algorithm can be run for $1 \le k \le r$, where it can be shown that $r \le \min(n, p)$ is the rank of the matrix $\mathbf{X}$. Attempts to running the algorithm a major number of time cause the results to repeat cyclically. It has to be underlined that, thought is not guaranteed that the $\mathbf{w}_k$s will be exactly orthogonal, they are unlikely to be very correlated, since the different $\mathbf{v}_k$s each are associated with orthogonal $\mathbf{u}_k$s.

Algorithm 7: Complementary PMD($\cdot$, L_1)

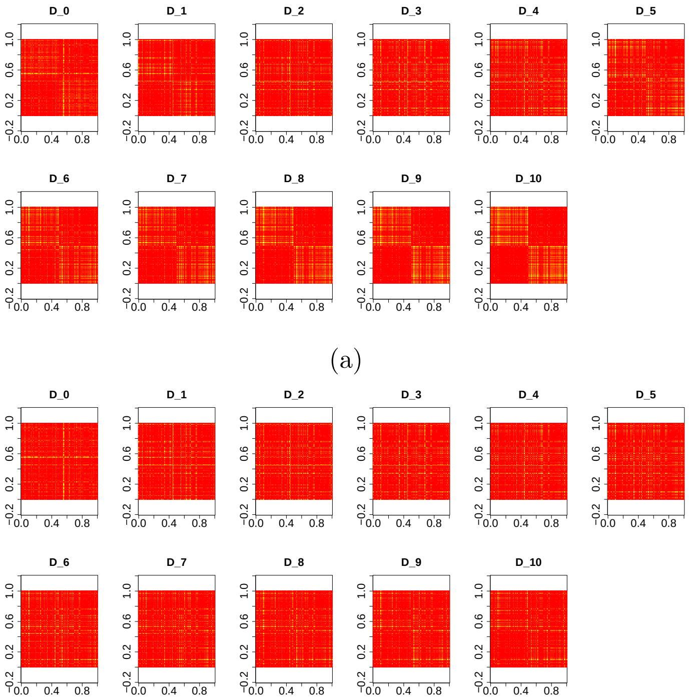
```
Data: $\mathbf{X}$, $\mathbf{U_{k-1}}$;
Result: $\mathbf{w_k}$, $\mathbf{u_k}$;
Initialization: initialize $\mathbf{w_k}$ to have $L_2$-norm 1;
while not convergence do
    1. $\mathbf{u_k} \leftarrow \frac{(\mathbf{I} - \mathbf{U_{k-1}}\mathbf{U_{k-1}}{}')\mathbf{X}\mathbf{w}_k}{\|\mathbf{U_{k-1}^{\perp}}{}'\mathbf{X}\mathbf{w}_k\|_2}$;
    2. $\mathbf{w_k} \leftarrow \frac{S(\mathbf{b_k}, \Delta)}{\|S(\mathbf{b_k}, \Delta)\|_2}$, where $\mathbf{b}_k = \mathbf{X}'\mathbf{u_k}$ and $\Delta = 0$ if $\|\mathbf{w}\|_1 \le c$,
       otherwise $\Delta > 0$ is chosen to be a positive constant such that $\|\mathbf{w}\|_1 = c$;
end
```

Appendix D

Image Plots

Here we report the main image plots of the $n \times n$ matrices composing $\mathbf{D} \in \mathbb{R}^{n^2 \times p}$, the matrix in which column j consists of the elements $d_{i,i',j}$ strung out into a vector, relative to instances of the Polynomial dataset taken into consideration in Subsection 4.2.1, and of the Fourier dataset analyzed in Subsection 4.2.2. We recall that the bottom-left corner of an image-plot corresponds to the top-left one of the matrix. Note that the matrices are symmetric, and that dark colors indicate low dissimilarities, while light colors indicate high ones, but the colors are relative to the values of the single matrix and have not an absolute meaning.

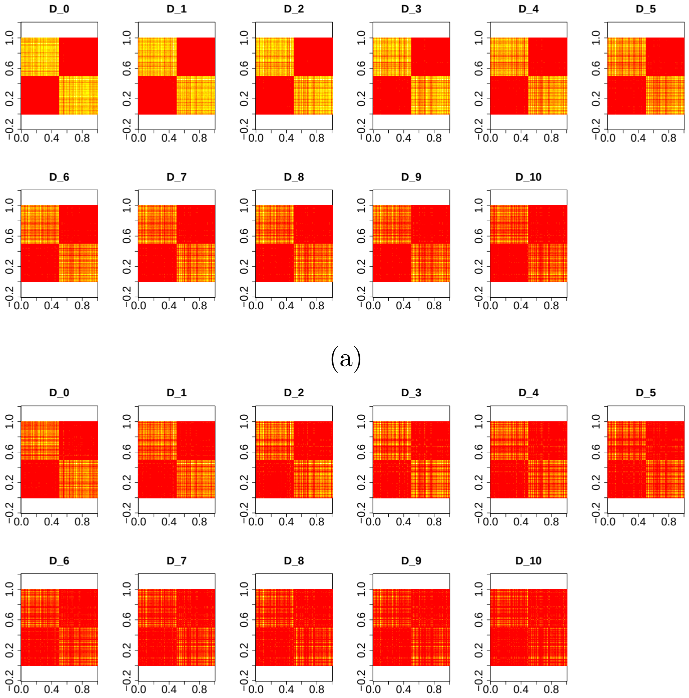
An optimal sparse adaptive Sobolev metric should be able to give nonzero weights to derivative levels in which we can see separations in blocks. The Sobolev metric equally weights all the matrices, while the L_2 metric give null weights to all the matrices except the first.



(a)

(b)

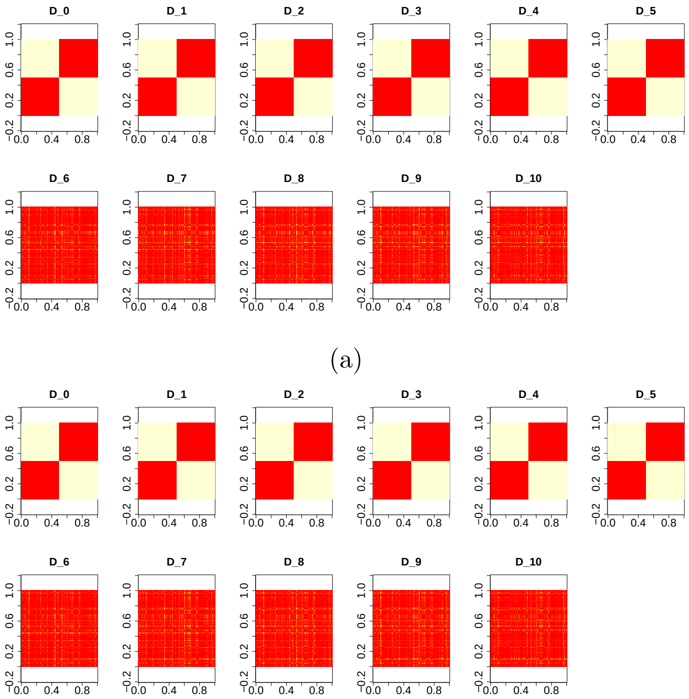
Figure D.1: POL$_1$, $\sigma^2 = 10^{-3}$(a), $\sigma^2 = 10^{-2}$(b).



(a)

(b)

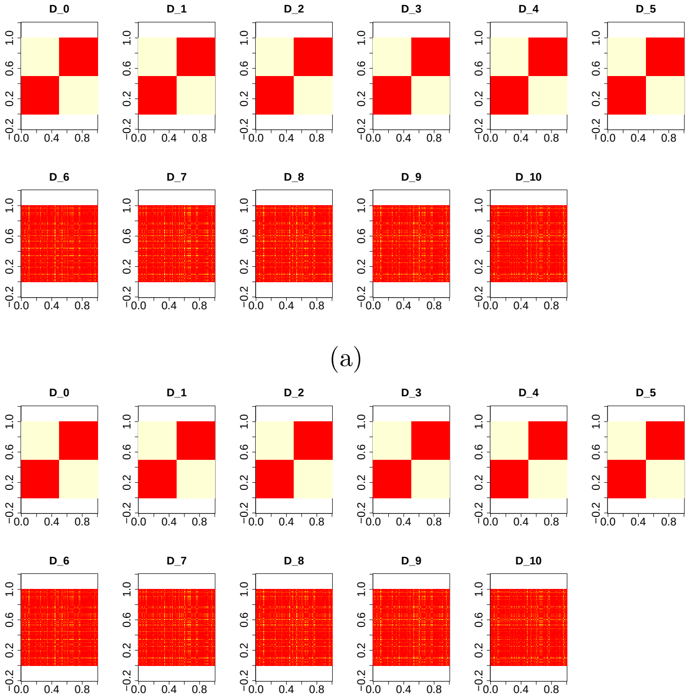
Figure D.2: POL$_2$, $\sigma^2 = 0.2$(a), $\sigma^2 = 1$(b).



(a)

(b)

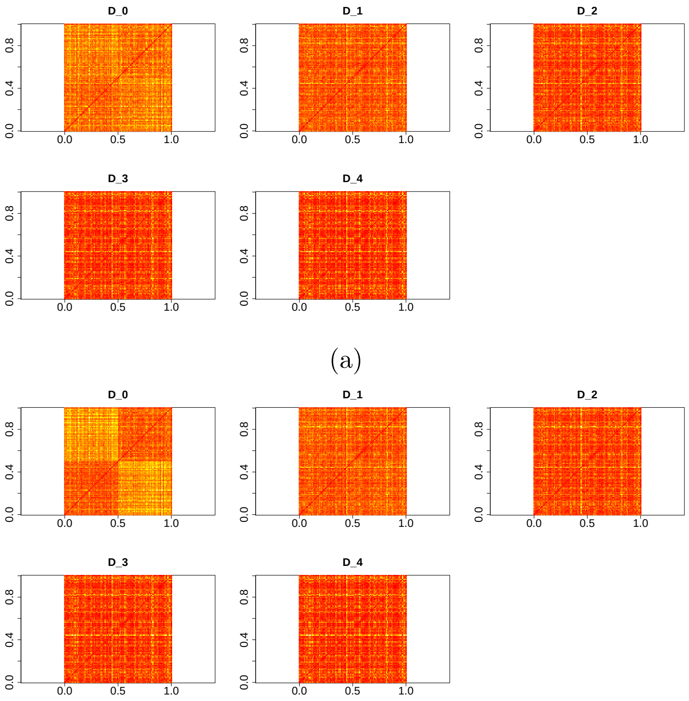
Figure D.3: POL$_3$, $\sigma^2 = 10^{-12}$(a), $\sigma^2 = 10^{-10}$(b).



(a)

(b)

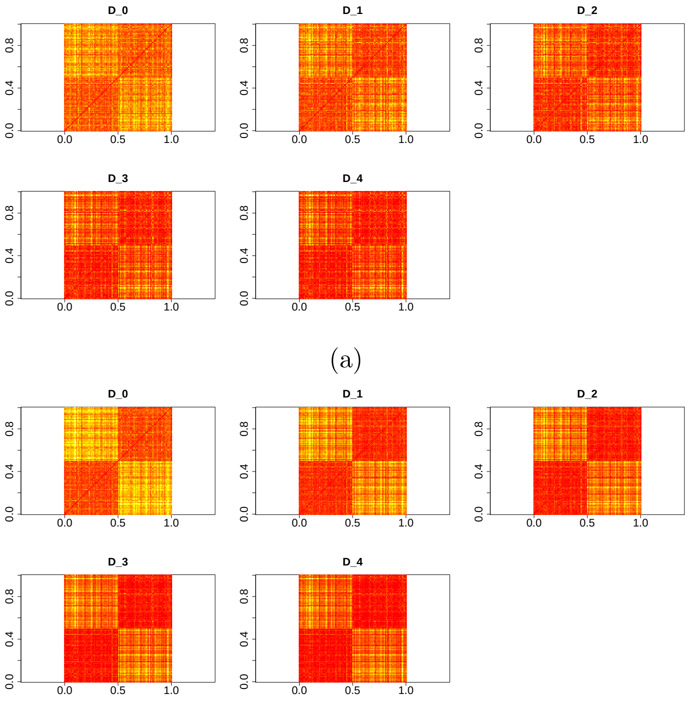
Figure D.4: POL$_4$, $\sigma^2 = 10^{-10}$(a), $\sigma^2 = 10^{-8}$(b).



(a)

(b)

Figure D.5: FOU_1, $\sigma^2 = 0.2$(a), $\sigma^2 = 0.005$(b).



(a)

(b)

Figure D.6: FOU_2, $\sigma^2 = 0.5$(a), $\sigma^2 = 0.2$(b).

Bibliography

`http://www.bi.ismac.cnr.it/kashmir/aziende.html`

ISO 17751:2007 normative, *Textiles – Quantitative analysis of animal fibers by microscopy – Cashmere, wool, specialty fibers and their blends.*

Boyd, S. and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge, U.K.: Cambridge University Press.

Buckingham, E., (1914), "On physically similar systems", *Physical Review*, 4, 354-367.

Dykstra, R., (2005), "Kullback-Leibler Information", DOI: 10.1002/0470011815.b2a15065.

Ferraty, F., Vieu, P., (2002), *The functional nonparametric modeland application to spectrometric data*, *Computational Statistics*, 17(4), 545-564.

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

Ferraty, F., Laksaci, A., Tadj, A., Vieu, P. (2010), "Rate of uniform consistency for nonparametric estimates with functional variables", *Journal of Atatisticl Planning and Inference*, 140(2), 335-352.

Fionn, M. and Legendre, P., (2011), "Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm", *ArXiv e-prints*, arXiv:1111.6285.

Hastie, T., Tibshirani, R. and Walther, G. (2001), "Estimating the Number of Clusters in a Dataset via the Gap Statistics", *Journal of the Royal Statistical Society*, 32, 2, 411-423.

Hastie, T., Tibshirani, R. and Witten, D. M. (2009), "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis", *Biostatistics*, 10, 3, 515-534.

Hastie, T., Tibshirani, R. and Zou, H. (2006), "Sparse principal component analysis", *Journal of Computational and Graphical Statistics*, 15, 265-286.

Huang, J., Z., Shen, H., (2008) "Sparse principal component analysis via regularized low rank matrix approximation", *Journal of Multivariate Analysis*, 99, 1015-1034.

Jolliffe, I., Trendafilov, N. and Uddin, M. (2003), "A modified principal component technique based on the lasso", *Journal of Computational and Graphical Statistics*, 12, 531-547.

Johnson, R. A. and Wichern, D. W. (2007), *Applied Multivariate Statistical Analysis* (6th ed.), Prentice Hall, Upper Saddle River

Nocedl, J. and Wright, S. (2006), *Numerical Optimization*, New York: Springer.

Nowak, G., and Tibshirani, R. (2007), “Complementary hierarchical clustering”, *Biostatistics*, 9, 3, 467-483.

Ramsay, J. O. and Silverman, B. W. (2002), *Applied Functional Data Analysis: Methods and Case Studies*, (1st ed.), New York: Springer.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, (2nd ed.), New York: Springer.

Robson, D., Weedall, P. J., Harwood, R. J., (1989), “Cuticular Scale Measurements Using Image Analysis Techniques”, *Textile Res. J.* 59, 713?719.

Robson, D., Weedall, P. J., (1990), “Cuticular Scale Pattern Description using Image Processing and Analysis Techniques”, *Proc. 8th Int. Wool Text. Rex. Conf.* II, 402-410.

Robson, D.,(1996), “Imaging Techniques for Animal Fiber Identification: A Critical Analysis”, *Metrology and Identification of Specialty Animal Fibers, European Fine Fiber Network* 4, 33-43.

Robson, D.,(1997), “Animal Fiber Analysis Using Imaging Techniques Part I: Scale Pattern Data”, *Textile Res. J.* 67, 747-752.

Robson, D.,(2000), “Animal Fiber Analysis Using Imaging Techniques Part II: Addition of Scale Height Data”, *Textile Res. J.* 70, 116-120.

Sangalli, L. M., Secchi, P., Vantini, S., and Veneziani, A. (2009), “A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery”, *Journal of the American Statistical Association,* 104, 37-48.

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010), “K-Mean Alignment for Curve Clustering”, *Computational Statistics & Data Analysis,* 54, 1219-1233.

Shannon, C., E., (1949), “Communication in the presence of noise”, *Proc. IRE*, 37, 10-21.

She, F. H., Kong, L. X., Nahavandi, S., Kouzani, A. Z., (2002) “Intelligent Animal Fiber Classification with Artificial Neural Networks”, *Textile Res. J.*, 72, 594-600.

Zhang, J., Palmer, S., Wang, X., (2010) “Identification of Animal Fibers withWavelet Texture Analysis”, Proceedings of the World Congress on Engineering, Vol I.

Tibshirani, R. and Witten, D. M. (2013), `http://cran.r-project.org/web/packages/sparcl/sparcl.pdf`, `R`-package documentation.

Tibshirani, R. and Witten, D. M. (2010), “A Framework for Feature Selection in Clustering”, *Journal of the American Statistical Association,* 105, 713-726.

Vantini, S. (2012), *On the definition of phase and amplitude variability in functional data analysis*, DOI: 10.1007//s11749-011-0268-9.

Wildman, A. B., (1954), *The microscopy of animal textile fibres*, WIRA, Leeds, United Kingdoms