POLITECNICO DI MILANO

Dipartimento di Matematica "F. Brioschi"

Ph.D. Course in
Mathematical Models and Methods for Engineering

XXIV Cycle

Classification of Functional Data in the presence of Spatial Dependence

Methods, Algorithms and Case Studies

PhD candidate: Valeria Vitelli,
matr. 738812

Advisorr: Prof. Piercesare Secchi
Co-advisor: Ing. Simone Vantini
Supervisor of the PhD Program: Prof. Paolo Biscari

Milano, April 26, 2012

A Gabriele

Acknowledgments

The first and most relevant thank goes to my advisor Piercesare Secchi. His wide statistical and mathematical competence, together with the passion he dedicates to all his scientific activities, have taught me the deep meaning of being a researcher.

Special thanks also to Simone Vantini, with whom I shared all the new ideas and moments of difficulty. He has been a real "benchmark" and a scientific brother to compare with in everyday work.

Thanks to Marco Gianinetto, for having efficiently coordinated all the scientific activities of the eni project which funded my PhD grant. Thanks to all eni employees involved in the project, and in particular to Mario Chiaramonte, for the stimulating interactions and for the efforts to find a common ground on which building a connection between our different cultures.

Final thanks go to all MOXians, particularly to the PhD students who shared with me these beautiful years, and to the "Tender" inhabitants: all of you gave a special contribution to this unique PhD experience.

Contents

Introduction

The general framework in which the present dissertation is placed is the analysis of high dimensional and complex data. The peculiarity of all the problems and case studies that will be described in the following is the combination of different sources of complexity: we will consider high–dimensional (functional) data, spatially dependent data, geo–referenced data on a (possibly huge) lattice. The main aim is performing data classification, i.e. reconstructing a latent field of labels which influences the distribution of the observed (functional) signal. This problem motivates the first part, and most of the second part, of the dissertation. Indeed, the general paradigm here discussed, at least in the framework of spatially dependent functional data, can be adapted to different problems arising in the applications, and to other purposes, e.g. dimensional reduction, regression.

Hence, the main focus of the dissertation is the development of methods and algorithms for the classification of functional data in the presence of spatial dependence. It is articulated in three parts: the first part concerns the classification of functional data when spatial dependence is not present, as a specific case of a more general framework in which spatial dependence is taken into account. This latter general framework is developed in the second part, where innovative classification and dimensional reduction methods, suited for the case of functional data indexed by the sites of a spatial lattice, are presented. Finally, in the third part of the dissertation, the analyses conducted within the *Microseepage eni project* are detailed: this scientific endeavour, which funded the PhD grant, has motivated the development of many of the methods described in the first two parts of the dissertation.

More specifically, the first two parts of the dissertation correspond to the "disclosable" part: the methods here described have been applied either to freely available datasets, or to case studies where results and insights can be publicly discussed, without any restriction due to confidentiality. They have been chosen due to the connection with the main topic of the dissertation, and their analysis is motivated by the necessity to test the proposed techniques in real situations, whose findings can be discussed with the scientific community. Each Chapter in the first two parts of the dissertation is hence organized as follows: in the first part of each Chapter the problem addressed, and the proposed innovative methods and algorithms are described; in the second part, a case study is presented, and the results of the application of the previously discussed methods are detailed. The final part of the dissertation, instead, concerns all the analyses we have conducted, and all the results we have attained, within the eni Microseepage project. It is thus subject to confidentiality constraints, and can not be disclosed.

The first part of the dissertation is composed by Chapter 1, 2 and 3. In Chapter 1, an introduction to many issues arising in the treatment of functional data is given. The focus is on the problem of clustering functional data, and on their possible misalignment. A case study is also discussed, concerning the morphological analysis of ECG curves.

Chapter 2 focusses on the problem of jointly clustering and aligning a set of functional data. Indeed, when treating functional data, one has to take into account a very specific source of variability, which can confound the results of the classification: phase variability. Starting from the methods discussed in the previous Chapter, an innovative joint solution to the two problems of clustering and alignment of functional data, which proved to be effective in real data analysis, is described. The proposed framework for performing joint clustering and alignment is tested on a benchmark dataset for functional data analysis: the Berkeley Growth Study dataset.

In Chapter 3 we further discuss the problem of clustering functional data, by considering also the issue of feature selection, i.e., the definition of a proper methodology able to cluster functional data and jointly to select the features which are the most relevant to the clustering scope. Many methods for sparse clustering suited for data belonging to a finite dimensional space have already been proposed. We start from the one described in Tibshirani and Witten [83], and we propose an innovative generalization of this method, which defines the problem of sparse clustering in infinite dimensional spaces. This is stated as an optimization problem whose solution exists under reasonable assumptions. The method is again tested on the Berkeley Growth Study dataset, and a comparison with the analysis described in the previous Chapter is also discussed.

The second part of the dissertation is composed by Chapter 4 and 5. In Chapter 4 the innovative methodology proposed for the analysis of functional data with spatial dependence, based on the bagging of coarse classifiers and on the treatment of spatial dependence via random systems of neighborhoods, is fully described. It has been extensively tested via a battery of simulation studies. The application to a real data set, concerning clustering of irradiance patterns for solar power applications, is also reported. Chapter 5, instead, describes a generalization of the strategy proposed in Chapter 4 to a problem of dimensional reduction: this generalization is motivated by a case study concerning functional telephonic patterns in time, geo-referenced on the metropolitan area of Milan. The analysis of these data has a possible innovative impact on a city planning analysis of Milan; for this reason, this case study has been conducted in collaboration with architects from DiAP (Dipartimento di Architettura e Pianificazione, Politecnico di Milano).

Finally, the third part of the dissertation is composed by Chapter 6 e 7. These two Chapters are not disclosable due to the confidentiality of the case studies and analyses here reported. Chapter 6 contains all the case studies considered during the Microseepage eni project 1, which lasted from October 2008 till November 2010. The innovative methodology developed thanks to the case studies here reported is the one described in Chapter 4. In Chapter 7 are instead discussed all the case studies, the methodological problems, and the innovative strategies for their solution, that have emerged during the second Microseepage eni project,

which lasted from February till December 2011. The framework here considered is more general, since not only the case of functional data is considered: the datasets we analyze are complex data with spatial dependence, spatial fields, compositional data.

All simulations and analysis of real data sets are performed in R (R Development Core Team 2006).

Part I

Classification of Independent Functional Data

Introduction to Part I

Classification problems motivate a large number of works in statistics, and great part of ongoing research on functional data analysis can be included in this framework. The guideline for this area of research is to split a large collection of objects into homogeneous groups. The classification domain can be divided into two main subcategories: *supervised* and *unsupervised* classification. The former refers to the situation when a learning sample, for which class memberships are known, is available; thus, the class structure is observed and the aim is to find the best rule which assigns each object to one of the classes. Usupervised classification (or *cluster analysis*), instead, refers to the situation when we do not observe the class membership of the considered collection of objects: this statistical problem is much more difficult because we have to jointly decide the number of classes, to define classes of objects, and to find a way to assign each object to a particular class [1] . Both supervised and unsupervised classification have been intensively studied in the multivariate case, and many references can be found in the monographs by Kaufman and Rousseeuw [36], Gordon [25] and Hand [26]; for the most recent advances see Hastie *et al.* [30].

The focus of this dissertation is on unsupervised classification problems for functional data. A recent introduction to functional data analysis has been given both in Ramsay and Silverman [61], in which different techniques to deal with functional data are described and detailed, and in Ferraty and Vieu [17], where a more theoretical approach to non parametric functional data analysis is presented. A focus on applied problems is instead given in Ramsay and Silverman [60]. In this first part we will focus on clustering independent samples of functional data.

A common problem in functional data analysis is the presence of a variability among data, which is ancillary to the classification scopes, and which is called *phase variability*. The focus of both unsupervised and supervised classification procedures, instead, is on *amplitude variability*, which induces a grouping structure among data. To decouple phase and amplitude variability is the scope of *alignment procedures*, which will be introduced in Chapter 1 together with the most common approaches to functional clustering. In Chapter 2, instead, a different light will be given on the problem of phase and amplitude variability decoupling, with the introduction of a method able to jointly handle clustering and aligment of functional data. Finally, in Chapter 3, we will deal with another common issue arising in the classification context, *sparse clustering*: the problem of selecting the subset of features which are more relevant to the classification scopes is widely studied in the multivariate framework, while in the functional framework no attempt has so far been made to handle this problem; in this dis-

[1] We will not include in the present dissertation the case of *fuzzy clustering* (also referred to as soft clustering), in which objects can belong to more than one cluster, and associated with each element is not a class membership but a set of membership levels, which indicate the strength of the association between that data element and a particular cluster. However, in Chapter 4, we will describe a clustering framework, suited for the case of spatially dependent functional data, in which final cluster memberships are frequencies of assignment to each of the clusters.

cussion we will show a possible generalization to the functional framework of the multivariate sparse clustering method discussed in Tibshirani and Witten [83].

According to the spirit of this work, methodological aspects are often presented in the light of the applications which stimulated their birth and development: the reader will find very close and explicit connections between methodological sections and more applicative ones. Moreover, connections between different chapters are also made explicit in the different approaches used to the same data anlaysis problem, as for the case of Berkeley Growth Study dataset (see Tuddenham and Snyder [85]), analyzed both in Chapter 2 and in Chapter 3.

Chapter 1

Disjointed Clustering and Alignment of Functional Data

Unsupervised classification is an important domain of statistics with many applications in various fields. The aim of this Chapter is to provide the basic tools to set a clustering problem in the functional framework; the idea is to generalize the optimal solution of iterative non–hierarchical clustering methods (e.g. k-mean method) to the case of functional data.

This dissertation on clustering methods for functional data can not disregard the misalignment problem, i.e. the possible existence of a phase variability among functional data. This aspect will also be considered in the following sections, but in this Chapter alignment procedures will be kept separate from the clustering solutions. In the final section we will give the details of a data analysis problem in which this choice is the most suited to the application at hand.

This Chapter is organized as follows. In Section 1.1 we introduce a proper framework for the analysis of functional data, which will be useful for the whole dissertation; here we mainly follow the introduction to functional data analysis given in Ferraty and Vieu [17]. In Section 1.2 we state the problem of curve clustering as a generalization of the multivariate setting, and we describe two iterative prototype methods for clustering functional data. In Section 1.3 we consider the problem of misalignment of the data, and we discuss two possible approaches to solve it. The final Section 1.4 is devoted to a case study, concerning the morphological analysis of a sample of ECG traces, with the scope of separating the physiological from the pathological ones.

1.1 Basics on functional data

There are an increasing number of situations coming from different fields of applied sciences (environmetrics, telecommunications, medicine, ..., only to cite those which will be considered in the present dissertation) in which the collected data are curves. The progress of computing tools, both in terms of memory and of computational capacities, has made available a huge amount of data in a widespread range of applications. The most common situation, often encountered in the stastistical literature, is the observation of a particular phenomenon

at several different times in the range $T := (t_{min}, t_{max})$; an observation can thus be expressed by the random family $\{X(t_j)\}_{j=1}^{J}$. However, if the collection $\{t_1, \ldots, t_J\}$ which defines the time grid can conceptually be made fine at will, data could be considered as observations from the family $\mathcal{X} = \{X(t); t \in T\}$.

In general,

Definition 1.1.1. *$\mathcal{X}$ is defined* functional random variable *(f.r.v.) if it is a random variable with respect to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which takes values in an infinite dimensional Hilbert space E, where $||\cdot||$ is the natural norm in E and we denote with $\langle \cdot, \cdot \rangle$ the inner product. An observation χ of $\mathcal{X}$ is called a* functional data.

Most of the methods for functional data considered in the following (and most on the methods for functional data developed so far in the literature on functional data analysis) identify the infinite dimensional Hilbert space E with $L^2(T)$. However, we will here consider E to be a generic Hilbert space, since at least in one case the considered functional space will not be $L^2(T)$.

Note that, when $\mathcal{X}$ (respectively χ) denotes a random curve (respectively its observation, i.e. a fixed element of E), we are implicitly identifying $\mathcal{X}$ with the process $\{\mathcal{X}(t), t \in T\}$, where $T := (t_{min}, t_{max})$ (and respectively χ with $\{\chi(t), t \in T\}$); this is not the sole situation in which functional data express themselves. The situation in which the variable is a curve is necessary associated with an unidimensional set $T \in \mathbb{R}$, but it is important to remark that the notion of functional variable covers a larger area: it includes *multivariate curves*, i.e. vector of curves defined on $T \subset \mathbb{R} \to \mathbb{R}^d$ and considered in the following sections, or *random surfaces*, like for instance the grey levels of an image or the realization of a spatial process, and in this case $T \subset \mathbb{R}^2$. Even if we will often restrict our dissertation to random univariate (or multivariate) curves, all the methodologies presented in the following are potentially applicable to any kind of infinite dimensional mathematical object.

Once introduced the theoretical setting for functional random variables, we should also introduce what we mean for functional sample. Following Ferraty and Vieu [17], we give the definition

Definition 1.1.2. *A functional dataset $\chi_1, \ldots, \chi_n$ is the observation of n functional random variables $\mathcal{X}_1, \ldots, \mathcal{X}_n$ identically distributed as $\mathcal{X}$; $\mathcal{S} := \{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ is a functional sample.*

We will always assume, in this first part of the dissertation, the functional sample to be an independent sample. This will not be always the case: in the second part we will instead introduce proper methods to deal with samples of correlated curves.

We will instead *not investigate* the question on how functional data have been collected, which is a problem linked with the discretization (grid) on which data are observed. A prelimary stage of the analysis always consists in presenting data in a suitable way for functional processing, and according to the kind of data different smoothing methods can be invoked; we will detail the adopted technique where needed for a detailed description of the considered application, but we will not go into details on the theoretical aspects connected to data smoothing.

1.1.1 Centrality Notions for Functional Variables

We will now describe very standard features such as mean and median for the distribution of a f.r.v. $\mathcal{X}$, pointing out how these basic concepts can easily be extended from the multivariate setting to the infinite dimensional context. Let us focus on the univariate random curves case (the generalization to multivariate curves or random surfaces is immediate).

Mean

We call $\mathcal{X}$ weakly integrable if there exists a $\mu_{\mathcal{X}} \in E$ such that

$$\mathbb{E}\langle \mathcal{X}, y\rangle = \langle \mu_{\mathcal{X}}, y\rangle, \ \forall\, y \in E. \tag{1.1}$$

In this case, $\mu_{\mathcal{X}}$ defined in (1.1) is called the expected value of $\mathcal{X}$.

Some elementary results hold:

- $\mu_{\mathcal{X}}$ is unique;
- integrability (i.e., $\mathbb{E}[||\mathcal{X}||] < \infty$) implies weak integrability;
- $||\mu_{\mathcal{X}}|| < \mathbb{E}||\mathcal{X}||$.

In the special case when $E = L^2(T)$ one can show that

$$\mu_{\mathcal{X}}(t) := \mathbb{E}[\mathcal{X}(t)] = \int_{\Omega} \mathcal{X}(\omega, t) d\mathbb{P}(\omega), \ \forall\, t \in \mathbb{R}, \tag{1.2}$$

which implies that one can obtain the functional mean simply by point-wise evaluation; note that the expression in (1.2) is a direct generalization of the univariate random variables definition of the mean. Moreover, this mathematical definition gives rise to the well–known estimator of the mean through its empirical version

$$\bar{\mathcal{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i, \tag{1.3}$$

which is a point-wise sample estimator of the functional mean; the consistency of the estimator defined in (1.3) for independent and identically distributed samples is given in Hormann and Kokoszka ([32]).

A well-known property of univariate random variables consists in the fact that the mean μ_X can be equivalently defined as $\mu_X := \int_{\Omega} x(\omega) d\mathbb{P}(\omega)$, or as the solution (under existence and uniqueness) of the following optimization problem

$$\inf_{x \in \mathbb{R}} \mathbb{E}[(x - X)^2]. \tag{1.4}$$

Thus, we can give a further definition of the functional mean $\mu_{\mathcal{X}}(t)$, following the analogy with (1.4).

Definition 1.1.3. *A solution (under existence assumption) to the following optimization problem*

$$\inf_{\chi \in E} \mathbb{E}[d^2(\chi, \mathcal{X})] \tag{1.5}$$

will be called functional mean *of the f.r.v. $\mathcal{X}$ defined in the infinite dimensional functional Hilbert space (E, d).*

We will not discuss the details of the relationship between the definitions of the mean given in (1.1), (1.2) and (1.5) when E is a generic Hilbert space, since this is not the focus of the present dissertation. However, the equivalence of (1.2) and (1.5) can be proven when the infinite dimensional space we are considering is $L^2(T)$, and the chosen metric is the natural norm in $L^2(T)$, due to the regularity of this space.

We have to remark that the point-wise definition (1.3) of the functional mean should be used carefully, deeply considering the impact of this choice on the application at hand. In particular, it is implicit in this definition that functional data should be smoothed first, and only after a point-wise mean can be computed: in fact, in case of rough data this computation can be completely nonsense. Indeed, very often data smoothing is not sufficient to achieve a meaningful point-wise mean: the use of (1.3) to approximate (1.1) is not appropriated also in the presence of phase variability among data, a situation in which we can not assign to each fixed $t \in T$ a precise meaning among curves, due to the variability along the abscissa showed by the curves belonging to the functional dataset. When phase variability is present, other techniques should be first used to align data, and only after the functional mean can be properly estimated from the sample (details on alignment techniques will be given in Section 1.3).

Finally, we should mention that a robust version of the mean curve (called *trimmed mean*) is defined in Fraiman and Muniz [19] using a functional version of the notion of depth (see Lopez-Pintado and Romo [50] for further details). Interestingly, this robust version of the sample mean has already been used in many unsupervised classification algorithms, like in Garcia-Escudero and Gordaliza [22].

Median

In the univariate setting, one can define the median of a real random variable X as the solution (under existence and uniqueness) of the following optimization problem

$$\inf_{x \in \mathbb{R}} \mathbb{E}[|x - X|].$$

This definition is coherent with the definition of the mean of a univariate random variable given in (1.4). It is immediate to extend this idea to the functional case, by replacing (under existence of a solution) $\mathbb{R}$ with the infinite dimensional space E, and the metric $|\cdot|$ with $d(\cdot, \cdot)$, leading to the following definition

Definition 1.1.4. *A solution (under existence assumption) to the following optimization problem*

$$\inf_{\chi \in E} \mathbb{E}[d(\chi, \mathcal{X})] \tag{1.6}$$

will be called functional median *of the f.r.v. $\mathcal{X}$ defined in the infinite dimensional functional Hilbert space (E, d).*

An empirical estimation of the functional median $\mathcal{X}_{med}$ is given by the solution to the following approximated optimization problem

$$\inf_{\chi \in E} \sum_{i=1}^{n} d(\chi, \mathcal{X}_i), \tag{1.7}$$

and a computable sample version can be found via

$$\mathcal{X}_{med,\mathcal{S}} = \underset{\chi \in \mathcal{S}}{\operatorname{argmin}} \sum_{i=1}^{n} d(\chi, \mathcal{X}_i). \tag{1.8}$$

This sample approximation of the functional median motivates the development of clustering procedures like k-medoids, in which the sample estimate of the functional median, rather than the mean, is used as a prototype for the clusters (see Section 1.2 for further details).

1.1.2 Proximity

One of the most important definitions that has to be given, also in the multivariate setting, in order to develop a classification procedure, is the notion of *proximity*. Proximity measures between mathematical objects play a major role in all statistical methods. In many situations, a classical norm can be used to measure the closeness between two objects. Since in a finite dimensional euclidean space (typically $\mathbb{R}^p$) there is an equivalence between all norms, the choice in the mathematical sense of this kind of measure is not crucial apart from practical considerations (e.g. computational ones). Considering an infinite dimensional space, instead, the equivalence between norms fails and the problem becomes crucial. Even more, limiting ourselves to consider the natural measure of proximity (or of the distance) induced by the norm in the considered Hilbert space can become too restrictive. In some situations, and particularly those considered in Chapter 2, it appears that semi–metrics are better adapted than metrics to measure the distance among functional data: the shape of data, exogenous information on the cosidered application, and eventually the scope of the study can help to drive the semi–metric selection.

Let us first recall some basic definitions

Definition 1.1.5. *$\|\cdot\|$ is a semi–norm on some normed space F if and only if:*

1. $\forall\ (\lambda, x) \in \mathbb{R} \times F$, $\|\lambda x\| = |\lambda|\,\|x\|$;
2. $\forall\ (x, y) \in F \times F$, $\|x + y\| \leq \|x\| + \|y\|$.

Note that, in fact, a semi–norm $\|\cdot\|$ is a norm, except for the fact that $\|x\| = 0 \not\Rightarrow x = 0$. Similarly, a semi–metric d can be defined to be a metric but such that $d(x, y) = 0 \not\Rightarrow x = y$.

Definition 1.1.6. *d is a semi–metric on some metric space F if and only if:*

1. $\forall\ x \in F$, $d(x, x) = 0$;
2. $\forall\ (x, y, z) \in F \times F \times F$, $d(x, y) \leq d(x, z) + d(z, y)$.

We will now introduce some semi–metrics which will be used in the following. Consider a functional random sample $\mathcal{X}_1, \ldots, \mathcal{X}_n$ identically distributed as the f.r.v. $\mathcal{X} = \{\mathcal{X}(t), t \in T\}$.

Semi–metrics based on FPCA

In many multivariate situations, the classical Principal Components Analysis (PCA) is considered as a useful tool for displaying data in a reduced dimensional space. More recently, the PCA method has been extended to functional data and used in a variety of applications, since in this context this powerful tool helps reducing the dimension of the space from infinity to a subset of selected components.

The discussion about functional principal components can be set in any Hilbert space, using the definition of the internal product $\langle \cdot, \cdot \langle$ that is natural in the considered space. However, this generality is not necessary to the scopes of the present dissertation; for further details on the estimation of functional principal components in more general spaces see Hormann and Kokoszka [33, 31]. We thus assume the cosidered f.r.v. $\mathcal{X}$ to take values in the infinite dimensional Hilbert space $L^2(T)$; in this space, the basis composed by functional principal components, thanks to Karhunen–Love decomposition, is a complete orthonormal basis (see Ferraty and Vieu [17] for theoretical details). As long as $\mathbb{E}\left[\int_T \mathcal{X}^2(t)dt\right]$ is finite, the FPCA (see Ramsay and Silverman [61]) of the f.r.v. $\mathcal{X}$ allows us to write the following expansion

$$\mathcal{X}(t) = \mu_{\mathcal{X}}(t) + \sum_{q=1}^{\infty} \left(\int_T \mathcal{X}(s)\nu_q(s)ds \right) \nu_q(t), \tag{1.9}$$

where $\mu_{\mathcal{X}}(t)$ is the mean of the random function $\mathcal{X}$, $\{\nu_1(t), \nu_2(t), \ldots\}$ are the orthonormal eigenfunctions, and $\{\lambda_1, \lambda_2, \ldots\}$ the associated eigenvalues, of the nucleus of the covariance operator $\Gamma_{\mathcal{X}}(s,t)$

$$\Gamma_{\mathcal{X}}(s,t) = \mathbb{E}\left[(\mathcal{X}(s) - \mu_{\mathcal{X}}(s))(\mathcal{X}(t) - \mu_{\mathcal{X}}(t))\right]. \tag{1.10}$$

Each function ν_q detects an orthonormal direction in the functional space $L^2(T)$, explaining a decreasing portion of variability λ_q; the quantities $\int_T \mathcal{X}(t)\nu_q(t)dt$ are called *scores*, and correspond to the projections of the random function in the direction of the q^{th} principal component. Once the first Q eigenfunctions have been selected, according to problem–driven considerations, a truncated version of the previous expansion (1.9) of $\mathcal{X}$ can be given

$$\tilde{\mathcal{X}}_{(Q)}(t) = \mu_{\mathcal{X}}(t) + \sum_{q=1}^{Q} \left(\int_T \mathcal{X}(s)\nu_q(s)ds \right) \nu_q(t), \tag{1.11}$$

Thus, we can define a parametrized class of semi–norms extending the classical L^2 norm in the following way

$$\|\mathcal{X}\|_Q^{FPCA} = \sqrt{\int_T (\tilde{\mathcal{X}}_{(Q)}(t))^2 dt} = \sqrt{\sum_{q=1}^{Q} \left(\int_T \mathcal{X}(t)\nu_q(t)ds \right)^2}. \tag{1.12}$$

This definition leads to the following parametrized family of semi–metrics

$$d_Q^{FPCA}(\chi_i, \chi_j) = \sqrt{\sum_{q=1}^{Q} \left(\int_T [\chi_i(t) - \chi_j(t)]\nu_q(t)dt \right)^2}. \tag{1.13}$$

Note that Q is not a smoothing parameter, but rather a *tuning parameter*, indicating the resolution level at which the problem is considered. Note also that, in practice, $\Gamma_{\mathcal{X}}$ is unknown and the ν_q's too. However, the nucleus of the covariance operator can be estimated from its empirical version

$$\Gamma^n_{\mathcal{X}}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\left[(\chi_i(s)-\bar{\chi}(s))(\chi_i(t)-\bar{\chi}(t))\right], \tag{1.14}$$

and the eigenfunctions of $\Gamma^n_{\mathcal{X}}$ are consistent estimators of those of $\Gamma_{\mathcal{X}}$ (see Cardot *et al.* [11] for further details). Other approaches to the estimation of the Karhunen–Love decomposition are conceivable; see for example Horvath and Kokoszka [33].

Semi–metrics based on derivatives

A great advantage of the previously discussed semi–metric stands in the fact that it is not necessary to assume a certain level of smoothness for the considered f.r.v.; in other words, the infinite dimensional functional space E has to be a Hilbert space but needs not to be a Sobolev space. When, instead, the problem at hand (or, more often, the chosen smoothing technique) allows us to assume $E := H^p(T)$, i.e. $\chi \in \mathcal{C}^{p-1}(T)$ and $\chi^{(p)}(t) \in L^2(T)$, for $p \geq 1$, we can build a parametrized family of semi–metrics between curves by considering the distance between one among their derivatives $\chi^{(1)}(t), \ldots, \chi^{(p)}(t)$, where with $\chi^{(i)}$ we have denoted the i-th derivative of χ.

More precisely, given two observed curves χ_i and χ_j, we consider the following semi–metric

$$d_p^{deriv}(\chi_i,\chi_j) = \sqrt{\int_T \left(\chi_i^{(p)}(t)-\chi_j^{(p)}(t)\right)^2 dt} = ||\chi_i^{(p)}-\chi_j^{(p)}||_{L^2(T)}. \tag{1.15}$$

So far we have considered the case of univariate random curves. However, this semi–norm can straightforwardly be generalized to the vectorial functions case. Suppose $\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n$ to be the realization of a random sample from $\mathcal{X} = \{\mathcal{X}(t), t \in T \subset \mathbb{R} \to \mathbb{R}^d\}$. Then, given two sample elements $\boldsymbol{\chi}_i$ and $\boldsymbol{\chi}_j$, a natural generalization of the semi–metric (1.15) is the following

$$d_p^{deriv}(\boldsymbol{\chi}_i,\boldsymbol{\chi}_j) = \sqrt{\sum_{r=1}^{d}\int_T \left(\chi_{i,r}^{(p)}(t)-\chi_{j,r}^{(p)}(t)\right)^2 dt}, \tag{1.16}$$

where $\chi_{i,r}(t)$ is the r-th element of the d-dimensional vector of functions $\boldsymbol{\chi}_i(t)$, for $r = 1, \ldots, d$.

1.2 Clustering methods for functional data

The main purpose in performing *cluster analysis* (also called *data segmentation*) is finding the best grouping of a collection of objects into subsets, or "clusters",

which result homogeneous in some sense. In other words, the aim of such techniques is to classify a sample of data into homogeneous groups, such that those within the same cluster are more closely related to one another than objects assigned to different clusters, without having any a priori knowledge about the true underlying clustering structure.

This statement, although quite heuristic, helps us to elaborate some preliminary considerations

- this concept, which is the basis of any unsupervised classification method, does not imply any restriction on the nature of the objects forming the sample, and can thus be straightforwardly generalized to any infinite dimensional space where functional data take values;

- unlike the supervised setting, we have no information on the nature of the grouping structure, which can only be inferred from the observed data;

- to the scopes of any clustering method, any object can equivalently be described by a set of variables, or by its relation to all other objects in the sample.

Hence, the first relevant choice that defines the clustering method is the *measure of the proximity* (or equivalently, of the *distance*) between the individual objects being clustered. The clustering method finds then the best grouping based on the optimization of a measure-dependent criterion. This implies that the infinite dimensional nature of data should not change the complexity of the method, provided we can describe all the relations between infinite dimensional objects via a proper measure of their proximity.

We should remark, but this concept will be often pointed out in the following, that this choice can only come from subject matter considerations. This situation is somehow similar to the specification of a loss or cost function in prediction problems (supervised learning): the cost associated with an inaccurate prediction depends on considerations outside the data.

Another common issue in unsupervised classification is the choice of the method, i.e. hierarchical or not, prototype–based or not. This choice is often driven by the purpose of the analysis; for instance, when a hierarchical method is used, the output of the analysis is a tree providing the nested grouping structure in the considered functional data set: the knowledge of this structure can be useful for cluster interpretation. In the following, we will always cosider non–hierarchical prototype–based methods, due to the fact that the role played by the prototype is often crucial: indeed, the method proposed in Chapter 2 for performing joint clustering and alignment of curves, is based on the possibility to align, at each iteration of the method, each functional data to the prototype of its cluster. Moreover, in Chapter 3 a method for sparse clustering of functional data is described, in which the issue of selecting relevant functional features is carried out via an optimization problem with objective function based on the distance of each curve from the prototype of its cluster.

We will describe in the next subsections two possible clustering methods for functional data, originally developed in the multivariate setting, but extremely

open to an immediate generalization. Both the approaches can be included in the broader framework of iterative non–hierarchical clustering methods, and they are based on the same idea: each cluster is identified with a prototype (also called *template*, either one of the data or a new object), and each cluster includes those object which are closest to the prototype of that cluster than to any other prototype. In particular, k-mean clustering for multivariate data (whose generalization to functional data is described in subsection 1.2.1) is a well–established algorithm, first proposed in the last century by Hartigan [28] and widely used in statistical classification problems for its flexibility. The algorithm divides a set of n data in k subgroups, iteratively alternating the search for the templates of the k clusters, that minimize within-cluster variability, and the assignment of each of the n data to the cluster whose template is the nearest. Multivariate k-medoid (generalized to handle more general objects than multivariate data in subsection 1.2.2) is a less computationally intensive modification of the k-mean algorithm. The main difference involves the way each cluster template is estimated, with the k templates simply selected among the n data. They are both extremely popular in the clustering literature, and they directly assign each observation to a group or cluster without regard to a probability model describing the data. For both methods, a prespecified number of clusters $k < n$ is chosen, and each one is labelled by an integer $j \in \{1, \ldots, k\}$. Each observation is assigned to one cluster via the assignment function $C(i) = j$, that assigns the i-th observation to the j-th cluster.

1.2.1 Functional k-mean

Many generalizations of the multivariate k-mean algorithm to functional data have been proposed in the literature on functional data analysis. For example, Shimizu and Mizuta [78], Tarpey and Kinadeter [80] and Tokushige *et al.* [84] proposes a generalization of k-mean clustering algorithms for functional data, as a way to solve the problem of principal points estimation. In Cuesta-Albertos and Fraiman [14], a robust k-mean clustering procedure is developed, based on the idea of "impartial trimming", which proves to be useful for high dimensional data. Another k-mean algorithm for functional data can be found in Chiou and Li [12], where the efficiency of the clustering procedure is improved thanks to the use of a non-parametric random-effect model. In particular, a detailed definition of functional k-mean procedure and an introduction to its consistency properties can be found in Tarpey and Kinateder [80]. Consider a functional dataset $\{\chi_1, \ldots, \chi_n\}$, supposed to be a realization of the functional sample $\mathcal{X}_1, \ldots, \mathcal{X}_n$ in the infinite dimensional Hilbert space E.

Functional k-mean clustering algorithm is an iterative procedure, which alternates a step of *cluster assignment*, in which all curves are assigned to a cluster, and a step of *centroid identification*, in which a relevant functional representative (the *centroid*) for each cluster is identified. The procedure is initialized with a set of k centroids, chosen at random among the curves in the sample, and it is stopped at convergence (i.e., when cluster assignments do not change from one iteration to the next one).

More precisely, consider the set of cluster centroids at the current iteration

$\varphi_1, \ldots, \varphi_k$. In the cluster assignment step each curve is assigned to the cluster whose centroid is nearer according to the chosen measure of the proximity between two functional data; this measure of proximity between objects in E is the semi–metric d. This means that the cluster assignment is chosen via the following criterion

$$C(i) = \underset{j \in \{1,\ldots,k\}}{\operatorname{argmin}} \, d(\chi_i, \phi_j). \tag{1.17}$$

Given the cluster assignments $C(i)$, for $i = 1, \ldots, n$, the identification of centroids $\varphi_j(t)$ for $j = 1, \ldots, k$, is performed solving the following optimization problem

$$\varphi_j(t) = \underset{\varphi \in E}{\operatorname{argmin}} \sum_{i:C(i)=j} d(\chi_i, \varphi)^2, \tag{1.18}$$

The solution to this infinite dimensional optimization problem obviously depends on the choice of the Hilbert space E, since this choice influences the metric d induced by the natural norm in the chosen space. When we consider $(L^2, ||\cdot||_{L^2})$, it is immediate to prove that the solution to (1.18) is the point-wise mean, as it was defined in (1.3), of the objects belonging to the considered cluster

$$\varphi_j(t) = \frac{1}{N_j} \sum_{i:C(i)=j} \chi_i(t), \tag{1.19}$$

where $N_j = |\{i : C(i) = j\}|$, for $j = 1, \ldots, k$. This fact descends directly from the definition of the functional mean given in (1.5).

The interesting fact is that the optimal choice for the centroid according to (1.18) is the point-wise mean of the cluster elements given in (1.19) also in the case we are looking for the centroid in the Sobolev space H^1, with the metric given by the natural norm

$$||\chi||^2_{H^1(T)} = \int_T \chi^2(t)dt + \int_T (\chi')^2(t)dt.$$

In this case the proof is a little more tricky, and it uses the equivalence between the H^1 norm and the following one

$$||\chi||^2_{equiv} = \int_T (\chi')^2 dt + \chi^2(\bar{t}), \tag{1.20}$$

where $\bar{t} \in T$. The proof is as follows. Due to the equivalence between the norm in H^1 and the one expressed by (1.20), we have to prove that the point-wise mean of the curves belonging to the j-th cluster is the solution to the following problem

$$\underset{\varphi \in H^1}{\operatorname{argmin}} \sum_{i:C(i)=j} \left[(\chi_i^2(\bar{t}) - \varphi(\bar{t}))^2 + \int_T (\chi_i'(t) - \varphi'(t))^2 dt \right].$$

Now, the objective function is the sum of two positive terms: hence, we can consider them separately. The minimizers of the two terms taken separately are obvious if we keep in mind that we know the solution to (1.18) for $E = L^2(T)$

and the measure d given by the natural norm; indeed, the first one is minimized by choosing

$$\varphi_{j,1}(\bar{t}) = \frac{1}{N_j} \sum_{i:C(i)=j} \chi_i(\bar{t}),$$

while the second one by choosing

$$\varphi'_{j,2}(t) = \frac{1}{N_j} \sum_{i:C(i)=j} \chi'_i(t),$$

which means

$$\varphi_{j,2}(t) = \int_{t_{min}}^{t} \varphi'_{j,2}(x)dx = \int_{t_{min}}^{t} \frac{1}{N_j} \sum_{i:C(i)=j} \chi'_i(x)dx,$$

for all $t \in T$. Finally, note that the choice of the value of $\bar{t}$ is free, provided $\bar{t} \in T$; hence, by choosing $\bar{t} = t_{min}$, we finally obtain the global minimizer

$$\varphi_j(t) = \frac{1}{N_j} \sum_{i:C(i)=j} \chi_i(t_{min}) + \int_{t_{min}}^{t} \frac{1}{N_j} \sum_{i:C(i)=j} \chi'_i(x)dx. \tag{1.21}$$

The last step of the proof consists in showing that the prototype given by the expression in (1.21) is, in fact, a point-wise mean of the curves belonging to the j-th cluster. But this fact is immediate if we exchange the finite sum and the integral in the second term, and if we use the Fundamental Theorem of Calculus

$$\begin{aligned}
\varphi_j(t) &= \frac{1}{N_j} \sum_{i:C(i)=j} \chi_i(t_{min}) + \int_{t_{min}}^{t} \frac{1}{N_j} \sum_{i:C(i)=j} \chi'_i(x)dx, \\
&= \frac{1}{N_j} \sum_{i:C(i)=j} \chi_i(t_{min}) + \frac{1}{N_j} \sum_{i:C(i)=j} \int_{t_{min}}^{t} \chi'_i(x)dx, \\
&= \frac{1}{N_j} \sum_{i:C(i)=j} \chi_i(t_{min}) + \frac{1}{N_j} \sum_{i:C(i)=j} [\chi_i(t) - \chi_i(t_{min})], \\
&= \frac{1}{N_j} \sum_{i:C(i)=j} \chi_i(t).
\end{aligned}$$

1.2.2 Functional k-medoid

The optimization problem that gives the criterion to choose the cluster prototypes in functional k-mean, given in (1.18), is based on squared distances. This causes the procedure to lack robustness against outliers that produce very large distances. However, the only part of the functional k-mean procedure which assumes squared distances is the identification of centroids, which can be modified to handle the non–squared case. For $j = 1, \ldots, k$, we look for the cluster prototype such that

$$\phi_j(t) = \operatorname*{argmin}_{\varphi \in E} \sum_{i:C(i)=j} d(\chi_i, \varphi). \tag{1.22}$$

Note that the so obtained $\phi_j(t)$ is exactly the functional median of objects belonging to the same cluster, according to the definition given in (1.7).

The approximate solution to this problem defines the functional k-medoid method, which is exactly analogous to the k-mean algorithm, except for the indentification of the prototypes: this step of *medoids identification* is performed solving the following optimization problem

$$\phi_j(t) = \underset{\varphi \in \mathcal{S}}{\operatorname{argmin}} \sum_{i:C(i)=j} d(\chi_i, \varphi); \tag{1.23}$$

thus, each cluster medoid ϕ_j, for $j = 1, \ldots, k$, is the sample functional median of the objects belonging to the cluster, according to the definition given in (1.8).

We remark that alternating between (1.17) and (1.23) represents a particular heuristic search strategy for trying to solve

$$\min_{C, \{i_j\}_{j=1}^k} \sum_{j=1}^{k} \sum_{C(i)=j} d(\chi_i, \chi_{i_j}),$$

while k-mean algorithm is a way to solve the analogous problem with squared distances. In the book by Kaufman and Rousseeuw [36] many strategies are proposed for solving the same problem, which rely on the reduction of the same criterion based on the exchange between the current medoid and one of the objects.

1.3 Curves alignment

We have just described the generalizations to the functional case of two well–known algorithms for performing unsupervised classification of multivariate data. However, the clustering of functional data should not be approached without taking into account a problem of critical importance in functional data analysis, often encountered in the applications. This problem is the possible *misalignment* of the data. Data misalignment consists in the presence of a variability along the curves abscissa, which is often ancillary to the scopes of the analysis. Aim of alignment procedures is indeed to capture the phase variability shown by each curve in the dataset with a fuction, called *warping function*, which involves a transformation of the curve abscissa according to a proper criterion.

Many methods for curve alignment (or curve registration) have been proposed in the literature. For example, Lawton et al. [43] and Altman and Villarreal [1] deal with this problem using self-modelling non-linear regression methods, Lindstrom and Bates [46] develop non-linear mixed-effects models, and Ke and Wang [38] merge the above approaches in the unifying framework of semiparametric non-linear mixed-effects models.

A possible approach to curve alignment consists in finding a non–linear transformation of the abscissa $h(t),\ t \in T$, i.e. the desired warping, in order to line up important features, also called *landmarks*, which are detectable in all curves. This procedure is most suitable when information on these features, which have a precise meaning in the considered application, is collected together with the

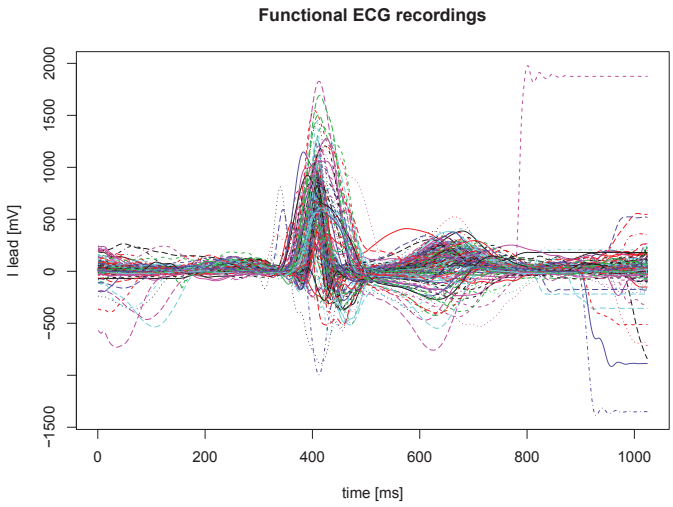

Figure 1.1: Sample of wavelets smoothed ECG recordings, on the I lead, of 198 patients, both healthy and affected by heart ischemic disease.

data. This is exactly the case of the ECG recordings shown in Figure 1.1: the picture shows the functional recostruction, via a multivariate wavelets smoothing technique (see the description of this analysis in Section 1.4 for further details), of a sample of 198 ECG traces (I lead). By inspection of the picture, we are able to detect a common shape for all ECGs, characterized by the typical peaks–waves alternation of the physiological ECG trace; however, each patient has his own heart beat timing, due to biological variability and to other clinical conditions (tachycardia, fibrillation, ...), which do not affect the shape but only the frequency of these peaks–waves. Indeed, the alignment of these curves can rely on a set of landmarks, which give the precise timing of each peak and wave start and ending in the ECG trace, and which are provided by the ECG recording system; this is an important piece of information we do not always have, which can influence the choice of the most suitable alignment procedure. Details on the analysis of these data will be given in Section 1.4.

And what if we do not have any landmark? A different line of research for curve alignment, advocated by J. O. Ramsay, is followed by Ramsay and Li [59], Ramsay and Silverman [61], James [35], Kaziska and Srivastava [37] and Sangalli et al. [65]. This technique, described in subsection 1.3.2, is based on the definition of a suitable measure of the proximity between curves, and performs curves alignment maximizing their proximity by means of a Procrustes procedure. The two possible approaches to perform curve alignment, landmarks alignment and the Procrustes continuous alignment procedure respectively, will be detailed in the next two subsections.

Figure 1.2 shows the growth curves of 93 children (39 boys and 54 girls) from Berkeley Growth Study data (see Tuddenham and Snyder [85]). This is a typical example, and a benchmark dataset for functional data analysis, considered by a

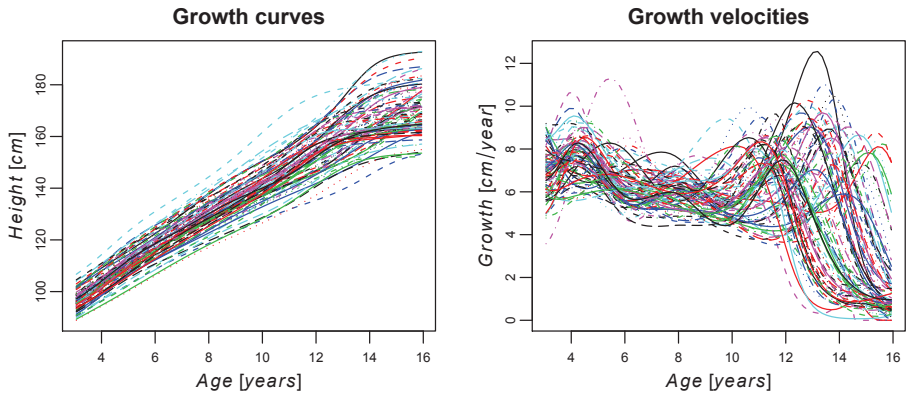

Figure 1.2: Growth curves of 93 children from Berkeley Growth Study data (left) and corresponding growth velocities (right).

number of authors (see for example Ramsay and Li [59], Sheehy et al. [76, 77], Ramsay and Silverman [61], James [35], Telesca and Inoue [81], Gervini and Gasser [24] and Sangalli *et al.* [67, 68]). Looking at the corresponding growth velocities, also displayed in Figure 1.2, it is apparent that all curves follow a similar course; this is characterized by a sharp peak of growth velocity between 10 and 16 years, the pubertal spurt, and a minor velocity peak between 2 and 5 years, the mid-spurt. However, different children have their growth spurts at different times, some take more time in their spurts, others less, each child following his/her personal biological clock. Thus, to learn something about the common growth path, it is first necessary to align the biological clocks of the children, eliciting the variability due to the different timings. This dataset will be considered in the next Chapter, as an example of a case study in which phase variability plays a crucial role in determining the grouping structure shown by the curves.

1.3.1 Feature or landmark alignment

A *landmark*, or feature, of a curve is some characteristic that one can associate with a specific argument value $t \in T$. These are typically maxima, minima, or zero crossing of curves, and may be identified at the level of some derivatives or at the level of the curves themselves. In the application we describe in the next Section, the procedure is even easier since landmarks are collected during data recording.

We set the problem of curve alignment in the general framework of estimating a possibly non–linear trasformation of the abscissa $h(t),\ t \in T$, and indicate how landmarks can help estimating this trasformation. Indeed, the application in the next section shows how vector–valued functional data (ECGs are multivariate functional data due to the presence of 8 leads) can be handled by obvious extensions of methods for scalar–valued functions.

The landmark alignment procedure requires for each functional data χ_i, $i =$

$1, \ldots, n$, the identification of the abscissa values $t_{i,f}$, $f = 1, \ldots, F$, associated with each of the F features. The goal is to construct a transformation h_i for each data such that the registered curves

$$\tilde{\chi}_i(t) := \chi_i(h_i(t)), \ i = 1, \ldots, n$$

show each of the F features in the same abscissa points $t_1, \ldots, t_F$. The so obtained transformation h_i will be called *warping function* of the abscissa.

Once a set of landmarks is given for each curve, the only task to accomplish is the resolution of a linear system defined by the following requirements: for all $i = 1, \ldots, n$

1. $h_i(t_{min}) = t_{min}$ and $h_i(t_{max}) = t_{max}$;
2. $h_i(t_f) = t_{i,f}$, for all $f = 1, \ldots, F$;
3. h_i is strictly monotonic, i.e. $s < t \ \Rightarrow \ h_i(t) < h_i(s)$.

We should clearly choose a basis to express the proper warping to satisfy the above mentioned constraints, e.g. piece–wise linear functions, polynomials, B-spline basis, ...; this choice depends both on the particular application at hand, and on the degrees of freedom the above requirements leave to the description of h_i.

1.3.2 The Procrustes continuous alignment procedure

The Procrustes fitting approach to alignment consists in finding, for each element $i = 1, \ldots, n$ of the functional sample, the continuous trasformation of the abscissa h_i that makes the curve the most similar to a sample prototype.

It has been originally developed in the multivariate setting, and involves the alternation of two steps:

- using the data to estimate a prototype φ with respect to which the transformation of the abscissa h_i for each observation $i = 1, \ldots, n$ can be defined;
- estimating the transformation itself, to make all curves the most similar to φ.

In the applications, as suggested in Ramsay and Li [59] and Kneip *et al.* [41], the point-wise sample mean of the unregistered curves is used as initial prototype for the estimation of each observation's warping function $h_i^{(1)}$; hence, $\varphi^{(0)}(t) = 1/n \sum_{i=1}^{n} \chi_i(t)$.

Once the first step warping functions have been estimated, an updated point-wise sample mean, obtained as

$$\varphi^{(1)}(t) = \frac{1}{n} \sum_{i=1}^{n} \chi_i(h_i^{(1)}(t)),$$

can be computed and used as a prototype for the computation of a revised version of warping functions, $h_i^{(2)}$. In this version of the Procrustes alignment procedure,

however, there is seldom need for the update from $h_i^{(1)}$ to $h_i^{(2)}$, since the change in the warping functions h_i from the first to the second iteration tends to be negligible (see Ramsay and Li [59], Ramsay and Silverman [61]).

An interesting remark is that, in case landmarks for the curves in the sample are provided (or whether they are obvious), the two procedures for curve alignment can be used in series: landmarks may be first used to align the curves before computing the initial prototype $\varphi_{align}^{(0)}(t)$, which is then used to initialize the Procrustes alignment procedure.

This fitting criterion for curve alignment is quite general, and needs not to be based on the curves sample mean: all notions of centrality can be used in the calculation of the prototype. Moreover, we shall see in Chapter 2 how this procedure can be integrated with a prototype based clustering procedure, to give birth to a *joint clustering and alignment procedure*; in this case, the most suitable choice for the prototype will depend on the adopted clustering method. See Section 2.2 for further details.

1.4 Case study: the analysis of ECG curves morphology

In this section we detail an application of the previously described functional clustering and alignment procedures to the analysis of ECG curves. The key point of the analysis consists in tuning and testing a real time procedure which enables a semi–automatic diagnosis of the patients' disease, not dependent on clinical considerations, but based only on ECG traces morphology. As detailed in the following, patients included in the sample are either healthy, or affected by right/left Bundle Branch Block (RBBB and LBBB respectively): Bundle Branch Block (BBB) is a cardiac conduction abnormality seen on the ECG, which causes activation of the left (right) ventricle to be delayed, resulting in the one ventricle contracting later than the other.

The main goal of this work is then to identify, from purely statistical considerations, specific ECG patterns which could benefit from an early invasive approach. Indeed, the identification of statistical tools capable of classifying curves using their shape only could support an early detection of heart failures, not based on usual clinical criteria. The functional dataset we are going to analyze ($n = 198$) is a sample of multivariate functional data describing patients' ECG: data are extracted from PROMETEO datawarehouse [1] , which contains all the ECG traces recorded on Milanese urban area by BLSs since the end of 2008. Each recording in PROMETEO datawarehouse consists in 8 curves (one for each ECG lead) for each patient, which represents his/her "Median" beat for that lead.

[1]PROMETEO (PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall' Extra Ospedaliero) is a project started in 2008 with the aim of spreading the intensive use of ECG as pre–hospital diagnostic tool, and of constructing a new database of ECGs with features never recorded before in any other data collection on heart diseases. Thanks to the partnerships of Azienda Regionale Emergenza Urgenza (AREU), Abbott Vascular and Mortara Rangoni Europe s.r.l., ECG recorder with GSM transmission have been installed on all Basic Life Supports (BLSs) of Milanese urban area.

Moreover, additional information can be collected in the datawarehouse, useful for signal processing and analysis: times of waves' repolarization and depolarization, which are landmarks indicating onset and offset times of main ECGs features, and an automatic diagnosis, established by Mortara-Rangoni VERITAS™ algorithm [2] , which we used to label ECG traces to validate the performances of our unsupervised clustering algorithm.

Details on Bundle Branch Blocks pathology and its impact on the shape of ECG signal will be treated in subsection 1.4.1, where also some clinical details about ECG signals will be given. Data preprocessing, and the chosen curves alignment and clustering methods are detailed in subsection 1.4.2, while in subsection 1.4.3 results of the analysis are discussed. Further details on this case study can also be found in Ieva *et al.*, [34].

1.4.1 Electrocardiography and Bundle Branch Block

Electrocardiography is a transthoracic recording of the electrical activity of the heart over time captured and externally recorded through skin electrodes. The ECG works mostly by detecting and amplifying the tiny electrical changes on the skin that are caused when the heart muscle depolarises during each heart beat (for further clinical details, see Lindsay [47]). Nowadays, the most commonly used clinical ECG system, the 12 lead ECG system, consists of the following 12 leads: I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, V6; of these 12 leads, only 8 are not dependent one from another, due to the configuration of measurement points on the patient's skin; hence, the ECG traces analyzed in the following will consist of leads I, II, V1, V2, V3, V4, V5 and V6 only.

Fig. 1.3 shows a scheme of the typical shape of a physiological single beat, recorded on ECG graph paper; main relevant points, segments and waves are highlighted. Deflections in this signal are denoted in alphabetic order starting with the letter P, which represents atrial depolarization. The ventricular depolarization causes the QRS complex, and repolarization is responsible for the T-wave. Atrial repolarization occurs during the QRS complex and produces such a low signal amplitude that it cannot be detected, with the exception of physiological ECGs (see Scher and Young [69]).

Bundle branch block pathology (also called fascicle injuries) result in altered pathways for ventricular depolarization. In this case, there is a loss of ventricular synchrony, ventricular depolarization is prolonged, and there may be a corresponding drop in cardiac output. From a clinical perspective, a RBBB typically causes prolongation of the last part of the QRS complex, while LBBB widens the entire QRS. Another usual finding with bundle branch block is appropriate T wave discordance: this means that the T wave will be deflected opposite the terminal deflection of the QRS complex. Unfortunately, some individuals will exhibit both left and right bundle branch blocks, thus showing a profoundly abnormal QRS interval.

In the following, we will focus our analysis on shape modifications induced on the ECG traces by BBB pathology, and we will investigate the grouping

[2] Mortara Rangoni Europe s.r.l. is the leading provider of ECG algorithms and components for various clinical applications, see `http://www.mortara.com`.

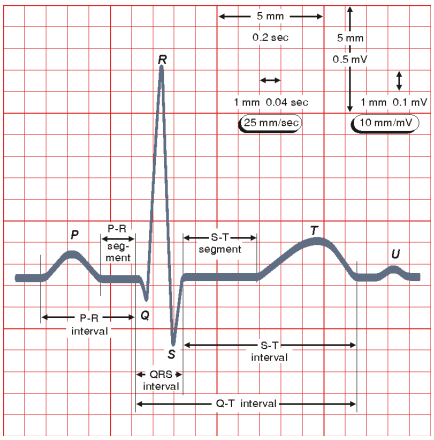

Figure 1.3: Scheme of the typical shape of a physiological single beat, recorded on ECG graph paper. Main relevant points, segments and waves are highlighted.

structure induced in the functional dataset by these shape modifications via a proper usupervised classification technique. The exploitation of these morphological modifications in the clustering procedure will be the focus of the following subsections.

1.4.2 Data smoothing, alignment and classification

The dataset coming from PROMETEO datawarehouse consists of the ECG signals of $n = 198$ subjects, among which 101 are Normal and 97 are affected by BBB (49 RBBB and 48 LBBB). As mentioned above, the aim of this work is exploring ECG curves morphology. Thus, the basic statistical unit is the multivariate functional data which describes heart dynamics, for each patient, on the eight leads.

However, in practice we have only a noisy and discrete observation of the function describing ECG trace for each patient. Moreover, each patient has his own "biological" time, i.e. the same event of the heart dynamics may happen at different time measurements for different patients (see Figure 1.1 and discussion): this is only misleading from a morphological point of view. These two very common problems can be addressed respectively with data smoothing and alignment.

Wavelets smoothing

A typical approach to perform data smoothing, i.e. to recostruct the functional form of the data which are noisily recorded, passes through the choice of a proper functional basis. Wavelet bases seem suitable for our data because every basis function is localized both in time and in frequency, being therefore able to capture ECG strong localized features (peaks, oscillations...). In particular we use a Daubechies wavelet basis with 10 vanishing moments (see Daubechies [15] for

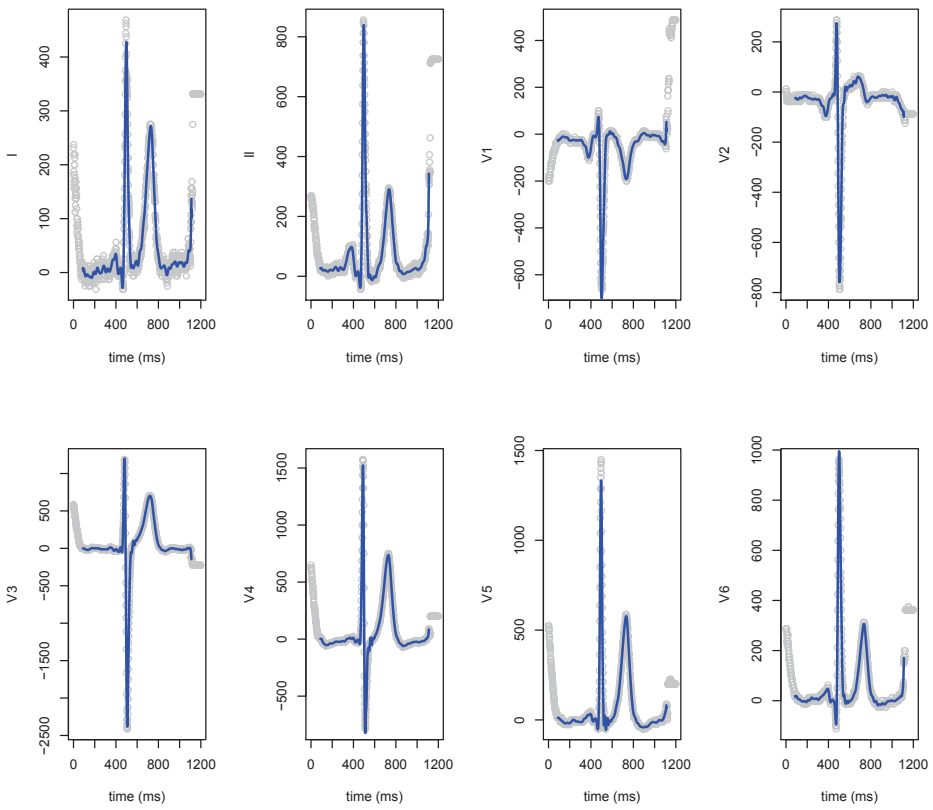

Figure 1.4: Raw data of the eight leads (black points) and wavelet functional estimates (blue) for a normal subject.

details), because we are interested also in derivatives of the ECG traces and thus we need a basis smooth enough for this purpose.

The wavelet smoothing procedure consists in changing over to wavelet domain, estimating basis coefficients and then obtaining a smooth estimate of the original signal thanks to wavelet basis expansion. As in most smoothing methods based on wavelet expansion, it is necessary to deal with a grid of 2^J points, $J \in \mathbb{N}$; in the further analysis we thus use only the central $2^{10} = 1024$ observation points. There is no loss of significant information: the portion of the signal on which we focus the analysis contains all the important features of the ECG trace. For this reason, we choose not to turn to non-decimated wavelets, which could be applied also to non dyadic grid but require a larger computational effort.

Since the eight leads (i.e. I, II, V1, V2, V3, V4, V5 and V6) jointly describe the complex heart dynamic, when smoothing these data it is appropriate to use a technique which takes into account all the eight leads simultaneously. This helps in detecting significant features, which reflect on more than one leads. To this aim in Pigoli and Sangalli [57] a wavelet based smoothing technique for

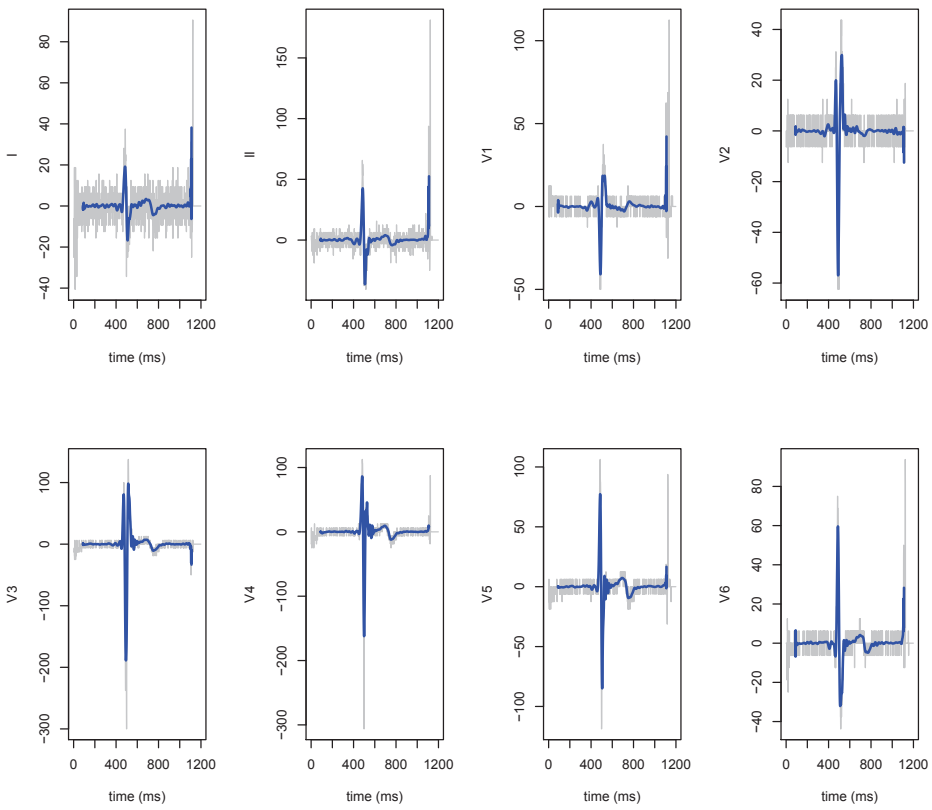

Figure 1.5: First central finite difference of the eight leads (gray) and wavelet estimates of the first derivatives (blue) for a normal subject.

multivariate curves is developed: this technique is used to obtain the estimation of 8 dimensional ECG signals and of their derivatives.

Thus, starting from the vectorial raw signal, we estimate the vectorial function

$$\boldsymbol{\chi}_i(t) = (I_i(t), II_i(t), V1_i(t), V2_i(t), V3_i(t), V4_i(t), V5(t), V6(t)),$$

and its derivatives, for each subject $i = 1, \ldots, n$. See Pigoli and Sangalli [57] for a detailed description of this smoothing procedure; in Figure 1.4 raw data and functional estimates obtained with this wavelet smoothing procedure for a normal subject are shown: observations are now in a functional form, suitable for performing classification. The smoothing procedure is essential also for an accurate derivative reconstruction, as shown in Figure 1.5, where the estimate of the first derivative is superimposed to the first central finite difference (i.e. a rough indication of first derivative behaviour).

Landmark alignment

As previously discussed in Section 1.3, functional ECG traces show both phase and amplitude variation, i.e. each patient has its own biological time, so that the same feature of the ECG recording can appear at different times among the patient. We address the problem of separating these two kinds of variabilities via a landmark based alignment procedure, similar to the one described in subsection 1.3.1. The landmarks we are going to use are provided by Mortara-Rangoni procedure, and identify the P wave (P_{onset} and P_{offset}), the QRS complex (QRS_{onset} and QRS_{offset}) and the T wave (T_{offset}); moreover, we decided to add one more landmark, indicating the R peak on the I lead ($Ipeak$), because this feature is clearly detectable on the I lead both in physilogical and in pathological ECG traces.

Since all the leads capture the same heart dynamics, biological time must be the same; thus, the same landmarks can be used to register all the leads. For the i-th patient, for $i = 1, \dots, n$, we look for a warping function h_i such that

$$\begin{aligned} h_i(P_{onset}) &= P^0_{onset} & h_i(P_{offset}) &= P^0_{offset} \\ h_i(QRS_{onset}) &= QRS^0_{onset} & h_i(Ipeak) &= Ipeak^0 \\ h_i(QRS_{offset}) &= QRS^0_{offset} & h_i(T_{offset}) &= T^0_{offset} \end{aligned}$$

where P^0_{onset}, P^0_{offset}, QRS^0_{onset}, $Ipeak^0$, QRS^0_{offset} and T^0_{offset} are the mean values of the correspondent landmarks for all the patients. These values are reported in Table 1.1, together with the associated standard deviations. We solve this problem using spline interpolation, and choosing h_i in the class of splines of degree 3; we choose non linear warping functions to align curves, since in this framework there is no simple affine transformation which can take into account the subject specific variability (we resort to Chapter 2 for the description of a methodology for curve clustering and alignment, in which affine warping is the most suitable choice).

We thus obtain the registered vectorial function

$$\mathbf{X}_i(t) = \boldsymbol{\chi}_i(h_i(t)),$$

for every patient $i = 1, \dots, n$. In Figure 1.6 both unregistered and registered I leads for all the 198 patients are shown.

This alignment procedure separates morphological information (i.e. amplitude variability) from duration of the different segments of ECG (i.e. phase variability): the former is captured by the registered ECG traces, while the latter is described by warping functions, determined by landmarks. The duration of the different segments of ECG has a great importance in clinical practice, and in particular the QRS complex length is one of the most important parameters to take into account in order to identify pathological situations. However, this kind of information is not able to distinguish among different pathologies, such as Right and Left BBB. This fact is confirmed in the considered functional dataset.

Table 1.1: Landmarks obtained at the end of the registration procedure, as the mean of landmarks of all the curves, and used to select the portion of smoothed and registered ECG curves relevant to our analysis (first line of the table); in the second line, landmarks standard deviations. Landmarks values are referred to a registered time in ms.

	P^0_{onset}	P^0_{offset}	QRS^0_{onset}	$Ipeak^0$	QRS^0_{offset}	T^0_{offset}
mean	184.3	298.2	354.8	407.2	476.9	755.8
standard deviation	39.7	37.4	18.9	15.4	21.4	44.2

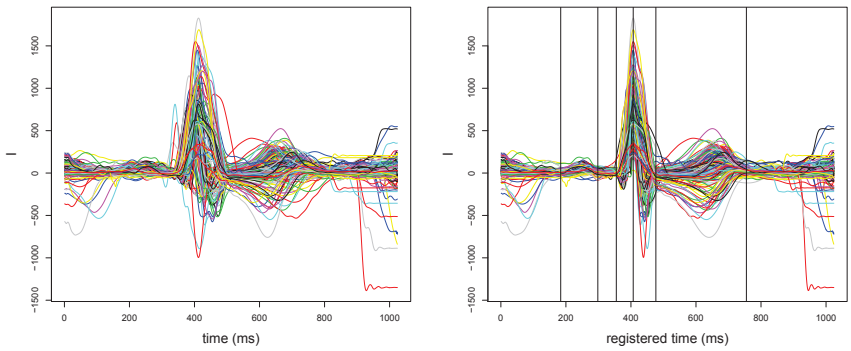

Figure 1.6: Original I leads for the n patients (left) and registered ones (right). Vertical lines indicate position of mean landmarks P^0_{onset}, P^0_{offset}, QRS^0_{onset}, $Ipeak^0$, QRS^0_{offset}, T^0_{offset}.

In fact, if we perform a multivariate 3-means clustering on interval lengths ($P_{offset} - P_{onset}$, $QRS_{onset} - P_{offset}$, $QRS_{offset} - QRS_{onset}$ and $T_{offset} - QRS_{offset}$), with the aim of identifying the existing 3 groups (Normal, RBBB and LBBB patients), we obtain the result shown in Table 1.2: this method correctly separates physiological traces from pathological ones, but it gives no information on the pathology. For this reason, in the following paragraph we will focus our analysis on the aligned curves, in the attempt to extract other diagnostic information from ECG morphology. In clinical practice, the result of our analysis should be considered together with traditional diagnostic tools based on segment lengths.

Table 1.2: Confusion matrix related to patients disease classification. Results are obtained performing 3-means clustering algorithm on interval lengths.

	Normal	RBBB	LBBB
Cluster 1	96	6	0
Cluster 2	2	17	25
Cluster 3	3	26	23

Unsupervised functional classification

Aim of the analysis is the development of a proper classification procedure, able to distinguish the grouping structure induced in the sample of ECGs by the presence of different pathologies, on the basis of the sole shape of the considered curves. The description of the chosen classification procedure is the focus of the present paragraph.

As previously discussed in subsection 1.4.1, ECG traces are very complex functional data, in which different portions of the domain can be analyzed in order to detect different pathologies. The main focus of our analysis stands in the investigation of BBB pathology, which mainly expresses in the ECG trace through a lengthening of the QRS complex and a modification of the T wave: we will thus focus our classification analysis on the QT-segment. Since ECG signals have been aligned, all the traces show relevant features at the same time points, corresponding to the reference landmarks P^0_{onset}, P^0_{offset}, QRS^0_{onset}, $Ipeak^0$, QRS^0_{offset}, T^0_{offset}: this fact allows us to select, for all the curves in the dataset, only the portion of ECG trace belonging to the interval $[P^0_{offset}, T^0_{offset}]$, which is relevant to our diagnostic purposes. In particular, we select only the portion of $\mathbf{X}(t)$ such that $t \in \tilde{T} := [P^0_{offset}, T^0_{offset}]$, where P^0_{offset} and T^0_{offset} are the values reported in the first line, second and sixth columns of Table 1.1.

We analyze the n patients according to a functional k-mean clustering procedure, in which all the eight leads $\mathbf{X}_i(t) : \tilde{T} \to \mathbb{R}^8$, for patients $i = 1, \ldots, n$, are simultaneously clustered. To develop this clustering procedure we suppose that $\mathbf{X}_i(t) \in E \equiv H^1(\tilde{T}; \mathbb{R}^8)$: we thus identify the infinite dimensional space in which the functional sample $\mathcal{X}_1, \ldots, \mathcal{X}_n$ takes values with the Sobolev space $H^1(\tilde{T}; \mathbb{R}^8)$; hence, the metric d_1 chosen to measure the proximity between objects in the sample is the most natural in this space, the H^1 norm $|| \cdot ||_{H^1(\tilde{T};\mathbb{R}^8)}$.

In order to perform comparisons, and to test the robustness of our clustering procedure, we consider two more ways to measure the distance between ECG traces; these two measures can be included in the previously described framework of semi–metrics based on derivatives: the semi–norm in the Sobolev space $H^1(\tilde{T}; \mathbb{R}^8)$, i.e. the L^2 norm of the first derivatives, and the norm in the Hilbert space $L^2(\tilde{T}; \mathbb{R}^8)$, which in the following we will indicate respectively with $\tilde{d}_1$ and d_2. These two measures are both considered in the clustering procedure not only to compare performances of multivariate functional k-mean under different specifications of the distance, but also to have an insight on the role of curves first derivatives: we claim that both the ECG trace and its first derivative are essential to distinguish more similar morphologies from less similar ones.

For details on the description of the functional k-mean clustering algorithm, and of the optimal centroid calculation with respect to the chosen metrics d_1, $\tilde{d}_1$ and d_2, we refer to Section 1.2, subsection 1.2.1. There are many different implementations of functional k-mean algorithm in the literature on functional data analysis, among which some procedures integrate data alignment in the classification steps. Here, instead, we chose to separate registration and clustering in two subsequent steps of the analysis, since the latter doesn't use any information beside morphology of the ECG traces, while the former is based on a strong clinical indication provided by landmarks supplied by the Mortara-Rangoni VERITAS™

algorithm.

The k-mean clustering procedure clearly depends not only on the choice of the distance, but also on the number of clusters k. Being the number of clusters a–priori unknown, we also consider a way to select the optimal number of clusters k^* via silhouette values associated to the final classification (see Struyf et al. [79]). In particular, the silhouette plot of a final classification consists in a bar plot of the *silhouette values* s_i, obtained for each patient $i = 1, \ldots, n$ as

$$s_i = \frac{b_i - a_i}{max\{a_i, b_i\}},$$

where a_i is the average distance, according to one of the chosen metrics d_1, $\tilde{d}_1$ and d_2, of the i-th patient to all other patients assigned to the same cluster, while

$$b_i := \min_{l=1,\ldots,k; l \neq C(i)} \frac{\sum_{j:C(j)=l} d(\mathbf{X}_i(t), \mathbf{X}_j(t))}{\#\{j : C(j) = l\}}$$

is the minimum average distance of the i-th patient from another cluster, where d is one of the chosen metrics. Clearly s_i always lies between -1 and 1, the former value indicating a misclassified patient, while the latter a very well classified one. Note that a patient which alone constitutes a cluster, has silhouette value equal to 1, but he is not considered in the silhouette plot for choosing k^*.

1.4.3 Results and discussion

Aim of the analysis is to detect the underlying grouping structure in our sample of 198 ECG traces. We thus perform clustering of the whole dataset via the multivariate functional k-mean algorithm previously described, using the different distances d_1, $\tilde{d}_1$ and d_2. The final silhouette plots obtained by clustering the sample of 198 ECG traces according to a multivariate functional k-mean procedure with distance d_1 and setting $k = 2, 3, 4, 5$ are shown in Figure 1.7: as we can appreciate from the picture, the grouping structure obtained setting $k = 3$ seems the best one, both in terms of silhouette profile, and in terms of wrong assignments. A similar result is obtained measuring the distance between curves via $\tilde{d}_1$ or d_2; however, the procedure seems to detect the best grouping structure when both the curves and their derivatives are considered in the distance. We thus set $k^* = 3$.

The final classification obtained with this choice of the distance, and setting $k = 3$, is shown in Figure 1.8, where the whole functional dataset is coloured according to cluster assignments; each panel corresponds to a different lead. From inspection of this picture a different shape of ECGs assigned to different clusters can be immediately appreciated, especially looking at the final centroids (functional mean) of each group, drawn in black in each panel of the picture. We shall now verify whether this difference in the ECGs morphology across clusters is due to the different pathology.

Since we have an indication of the different pathologies of the patients included in the sample, we can analyze the confusion matrix associated to the final cluster assignments, with respect to the Mortara-Rangoni algorithm classification

Table 1.3: Confusion matrices related to patients disease classification. Results are obtained by application of multivariate functional 3-mean clustering algorithm to smoothed and registered QT-segment of ECG curves, with different choices of the distance between ECGs: H^1 norm (top), H^1 semi-norm (centre) and L^2 norm (bottom). In the left table, cluster 1,2,3 respectively correspond to orange, green and red in Fig. 1.8.

	Normal	RBBB	LBBB
1	95	7	1
2	6	42	3
3	0	0	44

	Normal	RBBB	LBBB
1	71	12	0
2	30	36	5
3	0	1	43

	Normal	RBBB	LBBB
1	94	6	2
2	7	43	3
3	0	0	43

(Normal, RBBB and LBBB). The confusion matrices obtained via multivariate functional k-mean with different choices of the distance between curves (given by d_1, $\tilde{d}_1$ or d_2) are shown in Table 1.3. We remark that the final cluster assignments are based on the sole shape of the smoothed and registered ECG curves and their first derivatives, analyzed via a unsupervised classification procedure.

Both choosing the H^1 norm and the L^2 norm, the results seem appreciable, and slightly better in the former case: the final grouping structure traces out quite coherently the patients disease classification, with only few cases wrongly assigned. Moreover, we remark the improvement in the results obtained via multivariate functional 3-mean with respect to the results of 3-mean clustering algorithm on interval lengths (see Table 1.2): we are now able not only to detect pathological subjects, but also to distinguish between the two different pathologies present in the dataset. The result obtained via multivariate functional 3-mean clustering with H^1 semi-norm, instead, is not so positive, since cluster 1 and 2 apparently merge physilogical traces with ECGs of patients affected by RBBB.

The effectiveness of the clustering procedure in detecting the grouping structure among data suggests the definition of a semi–automatic diagnostic procedure based on the multivariate functional k-mean algorithm: in fact, the final result of our clustering procedure is a set of k centroids, representative of each cluster, which can be used as reference signals to compare a new ECG trace. Suppose a new ECG signal is available: we could have an immediate hint on the new patient's diagnosis by smoothing its ECG trace, registering it and finally assigning it to the group characterized by the nearest centroid. It is thus important to eval-

Table 1.4: Mean misclassification cost (first row) and standard deviation (second row) computed over 20 repetitions of the cross-validation procedure via equation (1.24).

distance	d_1	$\tilde{d}_1$	d_2
mean $cost_{CV}$	0.1227563	0.2286588	0.1275316
std dev $cost_{CV}$	0.1112663	0.1050911	0.1220574

uate the *misclassification cost* for this procedure, with the choice of the different functional distances. To this aim, we perform a *cross-validation analysis*.

We randomly choose among ECGs a training set of 80 Normal subjects, 40 RBBBs and 40 LBBBs, for a total of $n_{training} = 160$ curves. A multivariate functional 3-mean clustering is performed on the selected training set; we then consider the remaining $n_{test} = 38$ curves, and we assign each of them to the cluster whose centroid is nearer, according to distances d_1, $\tilde{d}_1$ and d_2. Given the patients disease classification, we compute misclassification cost using the following index

$$cost_{CV} = \frac{\lambda_1 \cdot misc_N + \lambda_2 \cdot (misc_{RN} + misc_{LN}) + \lambda_3 \cdot (misc_{RL} + misc_{LR})}{n_{test}}, \tag{1.24}$$

$misc_N$ being the number of healthy patients assigned to a pathological cluster [3], $misc_{RN}$ and $misc_{LN}$ the number of patients respectively affected by RBBB and LBBB which are assigned to the cluster of healthy patients, while $misc_{RL}$ and $misc_{LR}$ the number of patients whose ECGs are detected as pathological, but whose pathology is wrong. The parameters λ_1, λ_2 and λ_3 are misclassification weights: they are chosen according to the suggestion of the clinicians, who believe that assigning a BBB patient to the cluster of healthy patients is approximately 4 times more serious than treating as pathological a normal subject, which indeed is two times more serious than assigning a RBBB patient to the LBBB cluster (or viceversa); in order to determine the values of the weights we introduce a further request: $cost_{CV} = 1$ in the worst case, i.e. when all Normal subjects are classified as BBB and all BBB subjects are classified as Normal. This led to the choices $\lambda_1 = 0.4270$, $\lambda_2 = 1.7079$ and $\lambda_3 = 0.2135$.

We repeat this procedure 20 times, computing each time the misclassification cost according to equation (1.24): the mean and standard deviation computed along the 20 cross-validation repetitions are shown in Table 1.4. Even if all the distances provide good results, we notice that the norm in the Hilbert space $H^1(\tilde{T}; \mathbb{R}^8)$ seems to give best results, thus confirming our initial claim: both registered curves and first derivatives are needed to accurately compare ECGs morphology.

[3] given the final cluster assignments, the cluster of healthy patients is detected as the one that includes the most physiological traces. The pathological ones are subsequently chosen, first the one that contains the more RBBB traces, while the cluster that remains is the LBBB one.

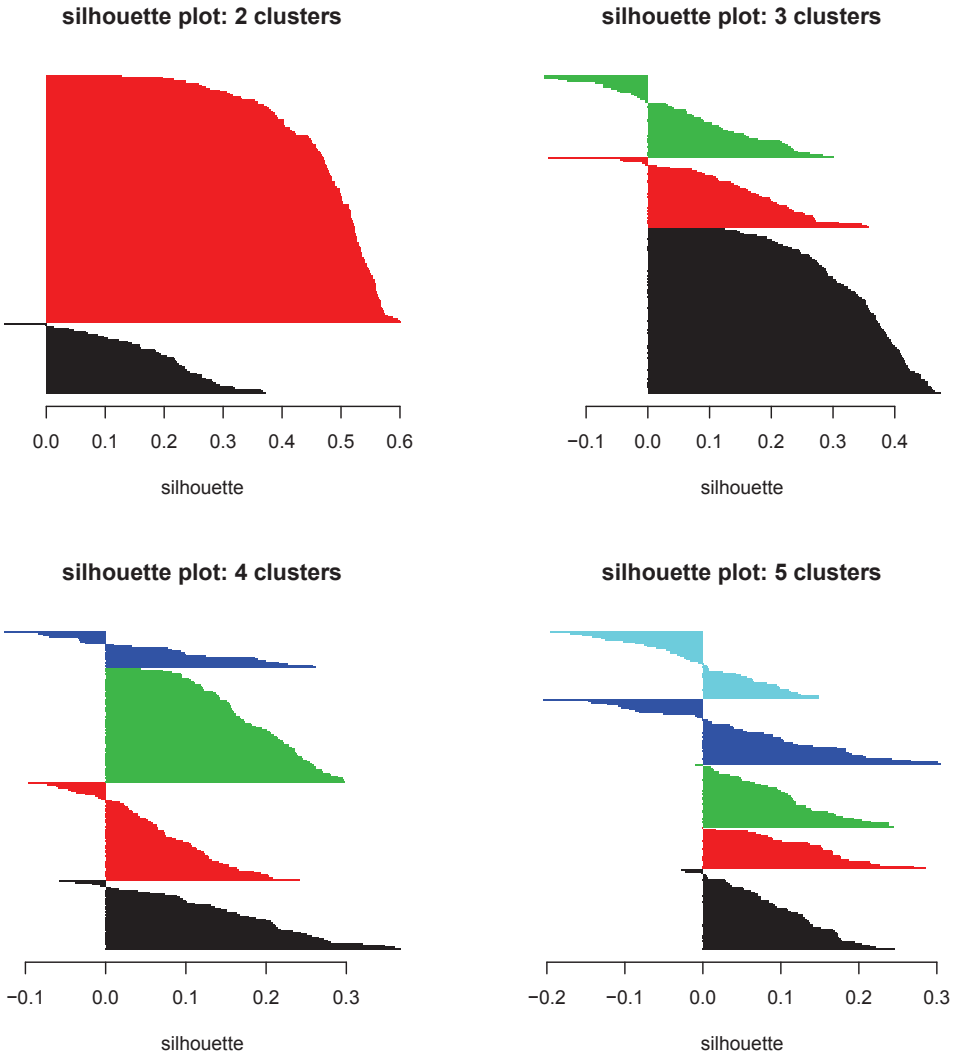

Figure 1.7: Silhouette plots of the clustering result obtained via multivariate functional k-mean procedure, setting $k = 2, 3, 4, 5$ and with distance given by d_1; data are ordered according to an increasing value of silhouette within each cluster, and are coloured according to the cluster assignment.

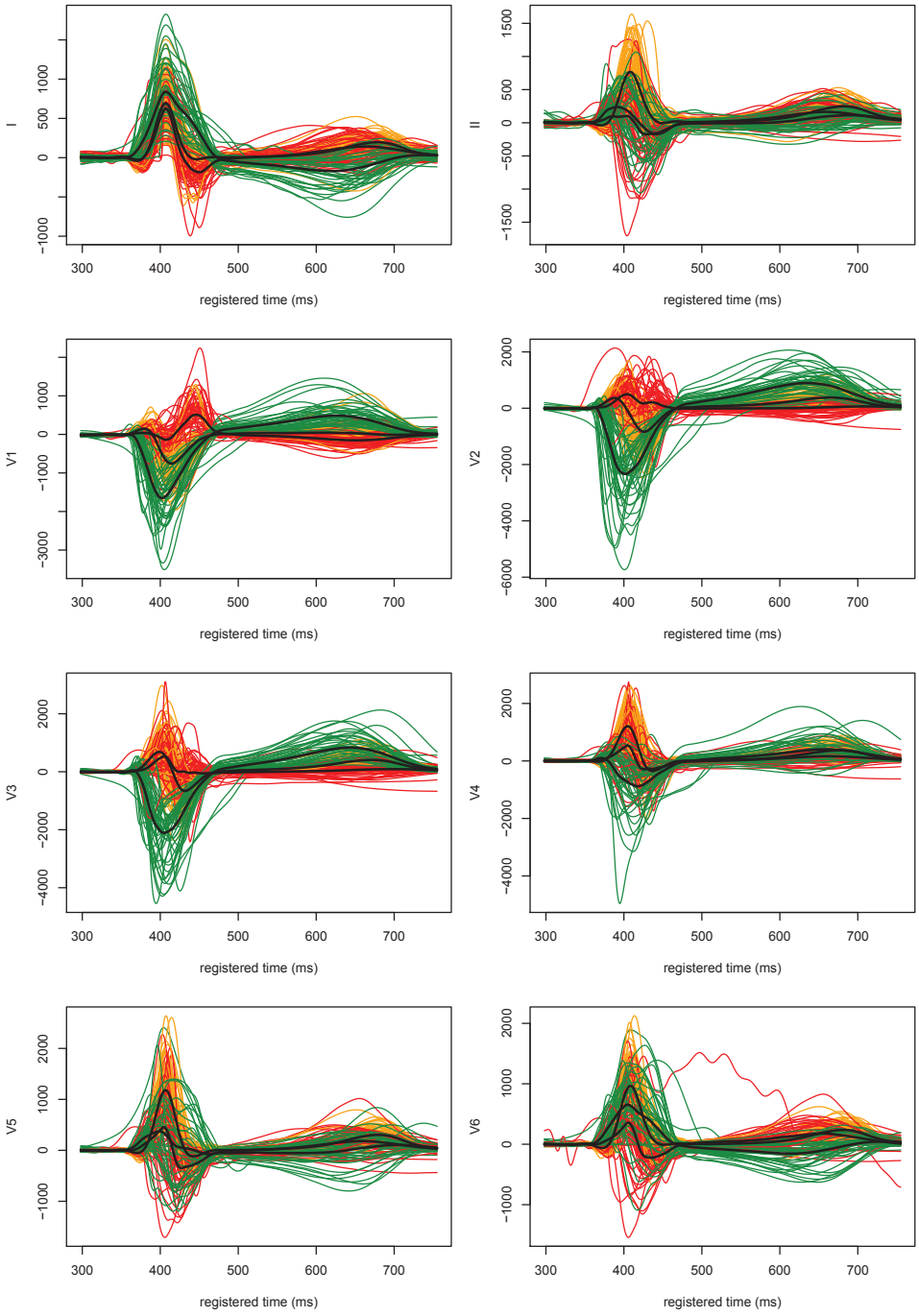

Figure 1.8: Smoothed and registered ECG traces (QT-segment): the whole dataset is coloured according to the final cluster assignments of multivariate functional 3-mean clustering, with distance given by d_1; the superimposed black lines are the three final cluster centroids (functional means). Each panel correspond to a different lead of the ECG traces.

Chapter 2

Joint Clustering and Alignment of Functional Data

The problem of *clustering of functional data* has been handled via different methods in the statistical literature on functional data analysis, most of which are based on generalizations of the multivariate k-mean algorithm. Indeed, when dealing with clustering of curves, we need to consider a problem which is peculiar to functional data, namely the possible *misalignment* of the data. An instance of this issue is given by the data in Figure 1.2, discussed in the previous Chapter when introducing the problem of misalignment of functional data, and describing the growth pattern (growth curves and associated growth velocities) in a sample of 93 children included in the Berkeley Growth Study (our benchmark dataset; see Tuddenham and Snyder [85] for further details on this study). Clustering of these data is of interest for the identification of different growth patterns, but the evident misalignment of the data could act as a confounding factor, and the above cited k-mean algorithms fail to give efficient results. This highlights the need for a clustering procedure which is able to jointly deal with data alignment, decoupling the variability due to data misalignment (phase variability) and the variability due to the shape (amplitude variability).

For a more detailed introduction to the problem of curve alignment (or curve registration), and to the possible approaches adopted in this dissertation, we refer to Chapter 1, Section 1.3. The introduction to the problem we gave in that context included two real data examples, the already mentioned growth curves, and the multivariate functional dataset of ECG traces, whose analysis is fully described in Section 1.4. These two application are deeply different: ECG curves have a specific pattern, associated with the electrophysiology of the heart, and fully understood from a clinical perspective; thus, when aligning these data, it is very important to include clinical information, e.g. landmarks associated to the ECGs relevant features, in order to remove phase variability consistently with the clinical interpretation of ECGs. The growth curves shown in Figure 1.2 are a quite different example of misaligned data: here we have no prior information on the curve pattern, nor we have landmarks associated to identifiable features (such as the pubertal spurt); moreover, it is clear that phase variability is induced in the sample by the biological variability among

children, an intrinsically continuous process on a macroscale. Thus, it seems more appropriate to approach the alignment of these data via a Procrustes procedure, described in subsection 1.3.2, which iteratively seeks the most suitable continuous warping of each curve in the functional dataset with respect to a given prototype (also called *template*).

Following the Procrustes approach to curve alignment, described in the previous Chapter, in Sangalli et al. [68] we proposed a procedure which is able to jointly cluster and align a set of functional data. We stated the problem of joint clustering and alignment of functional data as an optimization problem, and we proposed an iterative procedure for its solution. This procedure was thus specified in a k-mean algorithm. In the present Chapter, we describe the method proposed in Sangalli et al. [68] for joint clustering and aligning functional data in a broader framework, depending on the identification of a set of ideal cluster templates, solution to the clustering and alignment optimization problem. This leads to the definition of an alternative specification of the procedure in a k-medoid algorithm fashion; this new version approximates more directly the original optimization problem, and is potentially less sensitive to the presence of anomalous data. Moreover, this alternative version gives us the possibility of testing the robustness [1] of our alignment and clustering procedure with respect to the different specifications of the prototype for the clusters.

For details on the methods for joint clustering and alignment of functional data described in the present Chapter, and for other data analysis examples, see the works by Sangalli *et al.*, [67, 68]. In particular, these procedures have been successfully applied also to multivariate functional data describing cerebral vascular geometries: in this application data consists in the three spatial coordinates of 65 Internal Carotid Artery (ICA) centerlines, and functional clustering is of interest for the identification of ICA's with different morphological shapes, possibly associated with the presence or absence of a cerebral aneurysm; this work contributed to improve upon the exploratory statistical analyses performed on the same dataset in the AneuRisk Project [2] (previous analyses are detailed in Sangalli et al. [64, 65, 66]). See Boudaoud et al. [9], Liu and Yang [49] and Liu and Muller [48] for other recent approaches to the problem of clustering of misaligned functional data.

This Chapter is organized as follows. In Section 2.1 we introduce a proper framework for the problem of joint clustering and alignment, defining phase and amplitude variability. In Section 2.2 we state the problem of curve clustering and alignment as an optimization problem, and we propose an iterative procedure

[1] In the present dissertation, we shall call *robust* a procedure whose final results are not influenced by different specifications in the algorithm; in particular, we will refer to a robust clustering procedure when different approximations of the optimization problem which identifies cluster ideal templates do not affect the final result.

[2] The AneuRisk Project is a joint research program that aims at evaluating the role of vascular geometry and hemodynamics in the pathogenesis of cerebral aneurysms. The project involves MOX Laboratory for Modeling and Scientific Computing (Dip. di Matematica, Politecnico di Milano), Laboratory of Biological Structure Mechanics (Dip. di Ingegneria Strutturale, Politecnico di Milano), Istituto Mario Negri (Ranica), Ospedale Niguarda Ca' Granda (Milano) and Ospedale Maggiore Policlinico (Milano), and is supported by Fondazione Politecnico di Milano and Siemens Medical Solutions Italia.

for its solution. In Section 2.3 we describe two approaches to the template identification step in the procedure proposed in Section 2.2, obtaining a k-mean and a k-medoid specification of the iterative procedure. The subsequent Section 2.4 is devoted to the application of the two algorithm versions to the Berkeley Growth Study dataset.

2.1 Phase and Amplitude variability

The variability shown by the curves in a functional dataset can be decomposed into two components: *phase variability* and *amplitude variability*. Heuristically, phase variability is the one that we capture, and remove, when we perform curve alignment, while amplitude variability is the one that remains among the curves once they have been aligned. We will now specify a possible way to capture phase variability.

Consider an infinite dimensional semi–metric functional space E, and a curve in this space $\boldsymbol{\chi}(t)\colon \mathbb{R} \to \mathbb{R}^d$, $\boldsymbol{\chi} \in E$. Aligning $\boldsymbol{\chi}_1 \in E$ to $\boldsymbol{\chi}_2 \in E$ means finding a warping function $h(t)\colon \mathbb{R} \to \mathbb{R}$, of the abscissa t, such that the two curves $\boldsymbol{\chi}_1 \circ h$ and $\boldsymbol{\chi}_2$ are the most similar (with $(\boldsymbol{\chi} \circ h)(t) := \boldsymbol{\chi}(h(t))$). This is the typical setting for the Procrustes approach to curve alignment (see Section 1.3).

To properly define the alignment problem, it is thus necessary to specify a semi–metric $d(\cdot,\cdot)$ on E that measures the proximity between two curves, and a class W of warping functions h (such that $\boldsymbol{\chi} \circ h \in E$, for all $\boldsymbol{\chi} \in E$ and $h \in W$) indicating the allowed transformations for the abscissa. Hence, aligning $\boldsymbol{\chi}_1$ to $\boldsymbol{\chi}_2$ according to (d, W) means finding $h^* \in W$ that minimizes $d(\boldsymbol{\chi}_1 \circ h, \boldsymbol{\chi}_2)$. This procedure decouples phase and amplitude variability without loss of information: phase variability is captured by the optimal warping function h^*, whilst amplitude variability is the remaining variability between $\boldsymbol{\chi}_1 \circ h^*$ and $\boldsymbol{\chi}_2$. Note that the choice of the couple (d, W) defines what is meant by phase variability and amplitude variability. The choice of the semi–metric $d(\cdot,\cdot)$ is a crucial point, since the well-posedness and coherence of the optimization problem associated to clustering and alignment heavily depend on its definition. Many semi–metrics for measuring the proximity between functions have been considered in the literature on functional data analysis, and we have given a quick glance on some of the possibilities in subsection 1.1.2; for a proficient mathematical introduction to the issue see the book by Ferraty and Vieu [17].

Note that the same problem can equivalently be described in terms of a semi–metric $d(\cdot,\cdot)$, or in terms of a *similarity index* $\rho(\cdot,\cdot)$; the only difference between the two approaches, is that all minimization problems with respect to $d(\cdot,\cdot)$ become maximization problems with respect to $\rho(\cdot,\cdot)$. Sangalli et al. [65, 68] proposed the following bounded similarity index between two functions $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2 \in E$, where $E \equiv \{\boldsymbol{\chi} : \boldsymbol{\chi} \in H^1(\mathbb{R};\mathbb{R}^d), \boldsymbol{\chi}'(t) \neq 0 \text{ for } t \in T \subset \mathbb{R}, |T| > 0\}$

$$\rho(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2) = \frac{1}{d} \sum_{p=1}^{d} \frac{\int_{\mathbb{R}} \chi'_{1p}(t)\chi'_{2p}(t)dt}{\sqrt{\int_{\mathbb{R}} \chi'_{1p}(t)^2 dt}\sqrt{\int_{\mathbb{R}} \chi'_{2p}(t)^2 dt}}, \tag{2.1}$$

with χ_{ip} indicating the pth component of $\boldsymbol{\chi}_i$, $\boldsymbol{\chi}_i = \{\chi_{i1}, \ldots, \chi_{id}\}$; geometrically, (2.1) represents the average of the cosines of the angles between the derivatives

of homologous components of $\boldsymbol{\chi}_1$ and $\boldsymbol{\chi}_2$. The two curves are said to be similar when the index assumes its maximal value 1; for the similarity index defined in (2.1), this happens when the two curves are identical except for shifts and dilations of their components

$$\rho(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2) = 1 \quad \Leftrightarrow \quad \begin{array}{l} \text{for } p = 1, \ldots, d, \ \exists \theta_{0p} \in \mathbb{R}, \theta_{1p} \in \mathbb{R}^+ : \\ \chi_{1p}(s) = \theta_{0p} + \theta_{1p} \chi_{2p}(s). \end{array} \tag{2.2}$$

The choice of this similarity index comes along with the following choice for the class W of warping functions of the abscissa

$$W = \{h : h(t) = mt + q \ \text{ with } \ m \in \mathbb{R}^+, q \in \mathbb{R}\} \tag{2.3}$$

i.e., the group of strictly increasing affine transformations.

The couple (ρ, W) defined in (2.1) and (2.3) satisfies the following properties, that we deem to be minimal requirements for coherence:

1. The similarity index $\rho(\cdot, \cdot)$ is bounded, with maximum value equal to 1. Moreover, $\rho(\cdot, \cdot)$ is

   - *reflexive*: $\rho(\boldsymbol{\chi}, \boldsymbol{\chi}) = 1, \ \forall \ \boldsymbol{\chi} \in E$;
   - *symmetric*: $\rho(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2) = \rho(\boldsymbol{\chi}_2, \boldsymbol{\chi}_1), \ \forall \ \boldsymbol{\chi}_1, \boldsymbol{\chi}_2 \in E$;
   - *transitive*: $\left[\rho(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2) = 1 \wedge \rho(\boldsymbol{\chi}_2, \boldsymbol{\chi}_3) = 1\right] \Rightarrow \rho(\boldsymbol{\chi}_1, \boldsymbol{\chi}_3) = 1$ $\forall \ \boldsymbol{\chi}_1, \boldsymbol{\chi}_2, \boldsymbol{\chi}_3 \in E$.

2. The class of warping functions W is a convex vector space and has a group structure with respect to function composition $\circ$.

3. The index $\rho(\cdot, \cdot)$ and the class W are consistent in the sense that, if two curves $\boldsymbol{\chi}_1$ and $\boldsymbol{\chi}_2$ are simultaneously warped along the same warping function $h \in W$, their similarity does not change

   $$\rho\left(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2\right) = \rho\left(\boldsymbol{\chi}_1 \circ h, \boldsymbol{\chi}_2 \circ h\right), \quad \forall \ h \in W. \tag{2.4}$$

   This guarantees that it is not possible to obtain a fictitious increment of the similarity between two curves $\boldsymbol{\chi}_1$ and $\boldsymbol{\chi}_2$ by simply warping them simultaneously to $\boldsymbol{\chi}_1 \circ h$ and $\boldsymbol{\chi}_2 \circ h$.

Together, 2. and 3. imply the following property

4. For all h_1 and $h_2 \in W$,

   $$\rho\left(\boldsymbol{\chi}_1 \circ h_1, \boldsymbol{\chi}_2 \circ h_2\right) = \rho\left(\boldsymbol{\chi}_1 \circ h_1 \circ h_2^{-1}, \boldsymbol{\chi}_2\right) = \rho\left(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2 \circ h_2 \circ h_1^{-1}\right).$$

   This means that a change in similarity between $\boldsymbol{\chi}_1$ and $\boldsymbol{\chi}_2$ obtained by warping simultaneously $\boldsymbol{\chi}_1$ and $\boldsymbol{\chi}_2$ along h_1 and h_2 respectively, can also be obtained by warping the sole $\boldsymbol{\chi}_1$ or the sole $\boldsymbol{\chi}_2$ along $h_2 \circ h_1^{-1}$ or along $h_1 \circ h_2^{-1}$.

Moreover, the couple (ρ, W) defined in (2.1) and (2.3) satisfies the additional auxiliary property

5. Let W^d be the set of all transformations $\mathbf{r} : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ such that

$$\mathbf{x} = (x_1, ..., x_d) \in \mathbb{R}^d \quad \longmapsto \quad \mathbf{r}(\mathbf{x}) = (r_1(x_1), \ldots, r_d(x_d)) \in \mathbb{R}^d,$$

with $r_1, \ldots, r_d \in W$. Then, for all $\mathbf{r}_1$ and $\mathbf{r}_2 \in W^d$,

$$\rho(\mathbf{r}_1(\boldsymbol{\chi}_1), \mathbf{r}_2(\boldsymbol{\chi}_2)) = \rho(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2) \ .$$

In words, the similarity index between two curves is unaffected by strictly increasing affine transformations of one or more components of the curves.

2.1.1 Shape Invariant Model (SIM)

For $\boldsymbol{\chi} \in E$, assume the existence of $\boldsymbol{\varphi} \in E$ and of a parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \theta_2, \theta_3)$ (with $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ and $\boldsymbol{\theta}_1 \in \mathbb{R}^{d,+}$, $\theta_2 \in \mathbb{R}$, $\theta_3 \in \mathbb{R}^+$), such that

$$\boldsymbol{\chi}(t) = \boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 \cdot \boldsymbol{\varphi}(\theta_2 + \theta_3 t), \tag{2.5}$$

where with $\cdot$ we denote the component-wise multiplication of vectors in $\mathbb{R}^d$. If equation (2.5) holds, we shall simply write $\boldsymbol{\chi} \in \mathrm{SIM}(\boldsymbol{\varphi})$, since the condition (2.5) means that $\boldsymbol{\chi}$ falls within a *shape invariant model* (SIM), with *characteristic shape curve* $\boldsymbol{\varphi}$. For $d = 1$, SIM models were introduced by Lawton et al. [43]. For further details, see Kneip and Gasser [40].

SIM models are strongly connected with the couple (ρ, W) defined in (2.1) and (2.3). Indeed,

$$\exists h \in W : \rho(\boldsymbol{\chi} \circ h, \boldsymbol{\varphi}) = 1 \quad \Leftrightarrow \quad \boldsymbol{\chi} \in \mathrm{SIM}(\boldsymbol{\varphi}); \tag{2.6}$$

this follows directly from (2.2) and (2.3). Note that, thanks to property 4., the roles of $\boldsymbol{\chi}$ and $\boldsymbol{\varphi}$ can be swapped.

Now, consider a n-dimensional functional dataset $\{\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n\} \subset E$, such that $\boldsymbol{\chi}_i \in \mathrm{SIM}(\boldsymbol{\varphi})$ for $i = 1, \ldots, n$; then, the following property follows immediately:

6. For all $\boldsymbol{\chi}_i, \boldsymbol{\chi}_j$, with $i, j = 1, \ldots, n$, $\exists h_i \in W,\ h_j \in W$ such that

$$\rho\left(\boldsymbol{\chi}_i \circ h_i, \boldsymbol{\chi}_j \circ h_j\right) = \rho\left(\boldsymbol{\chi}_i \circ h_i, \boldsymbol{\varphi}\right) = \rho\left(\boldsymbol{\chi}_j \circ h_j, \boldsymbol{\varphi}\right) = 1 \quad \forall\ i, j = 1, \ldots, n.$$

Hence, in the present framework we can describe the situation in which only phase variability (and no amplitude variability) is present in the functional dataset by assuming all functional data to be drawn from the same shape invariant model.

2.2 Joint clustering and alignment of functional data

We will now integrate the previously introduced framework for alignment with a prototype clustering method, in order to develop an innovative approach to joint clustering and alignment of functional data. Consider the problem of clustering and aligning a functional dataset $\{\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n\}$ with respect to a set of k template

curves $\underline{\boldsymbol{\varphi}} = \{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_k\}$ (with $\{\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n\} \subset E$ and $\underline{\boldsymbol{\varphi}} \subset E$). For each template curve $\boldsymbol{\varphi}_j$ in $\underline{\boldsymbol{\varphi}}$, define the domain of attraction

$$\Delta_j(\underline{\boldsymbol{\varphi}}) = \{\boldsymbol{\chi} \in E : \sup_{h \in W} \rho(\boldsymbol{\varphi}_j, \boldsymbol{\chi} \circ h) \geq \sup_{h \in W} \rho(\boldsymbol{\varphi}_r, \boldsymbol{\chi} \circ h), \forall\, r \neq j\}, \quad j = 1, \ldots, k. \tag{2.7}$$

Moreover, define the labeling function

$$\lambda(\underline{\boldsymbol{\varphi}}, \boldsymbol{\chi}) = \min\{r : \boldsymbol{\chi} \in \Delta_r(\underline{\boldsymbol{\varphi}})\}. \tag{2.8}$$

Note that $\lambda(\underline{\boldsymbol{\varphi}}, \boldsymbol{\chi}) = j$ means that the similarity index obtained by aligning $\boldsymbol{\chi}$ to $\boldsymbol{\varphi}_j$ is at least as large as the similarity index obtained by aligning $\boldsymbol{\chi}$ to any other template $\boldsymbol{\varphi}_r$, with $r \neq j$. Thus $\boldsymbol{\varphi}_{\lambda(\underline{\boldsymbol{\varphi}}, \boldsymbol{\chi}_i)}$ indicates a template the curve $\boldsymbol{\chi}$ can be best aligned to and $\lambda(\underline{\boldsymbol{\varphi}}, \boldsymbol{\chi})$ a cluster which $\boldsymbol{\chi}$ should be assigned to. Indeed, $\boldsymbol{\chi} \in SIM(\boldsymbol{\varphi}_j)$ implies that $\lambda(\underline{\boldsymbol{\varphi}}, \boldsymbol{\chi}) = j$.

Case of known templates. If the k templates $\underline{\boldsymbol{\varphi}} = \{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_k\}$ were known, then clustering and aligning the set of n curves $\{\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n\}$ with respect to $\underline{\boldsymbol{\varphi}}$ would simply mean to assign $\boldsymbol{\chi}_i$ to the cluster $\lambda(\underline{\boldsymbol{\varphi}}, \boldsymbol{\chi}_i)$ and align it to the corresponding template $\boldsymbol{\varphi}_{\lambda(\underline{\boldsymbol{\varphi}}, \boldsymbol{\chi}_i)}$, for $i = 1, \ldots, n$.

Here we are interested in the more complex case of unknown templates.

Case of unknown templates. In the previously described setting, the problem of clustering and aligning the set of n curves $\{\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n\}$ with respect to k unknown templates means finding the solution (under existence and uniqueness) to the following optimization problem

(i) find $\underline{\boldsymbol{\varphi}} = \{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_k\} \subset E$ and $\underline{\mathbf{h}} = \{h_1, \ldots, h_n\} \subset W$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \rho(\boldsymbol{\varphi}_{\lambda(\underline{\boldsymbol{\varphi}}, \boldsymbol{\chi}_i)}, \boldsymbol{\chi}_i \circ h_i) \geq \frac{1}{n} \sum_{i=1}^{n} \rho(\boldsymbol{\psi}_{\lambda(\underline{\boldsymbol{\psi}}, \boldsymbol{\chi}_i)}, \boldsymbol{\chi}_i \circ g_i), \tag{2.9}$$

for any other set of k templates $\underline{\boldsymbol{\psi}} = \{\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k\} \subset E$ and any other set of n warping functions $\underline{\mathbf{g}} = \{g_1, \ldots, g_n\} \subset W$,

and then, for $i = 1, \ldots, n$,

(ii) assign $\boldsymbol{\chi}_i$ to the cluster $\lambda(\underline{\boldsymbol{\varphi}}, \boldsymbol{\chi}_i)$ and warp $\boldsymbol{\chi}_i$ along h_i.

The optimization problem (i) describes a search both for the set of optimal k templates, and for the set of optimal n warping functions. Note that the solution $(\underline{\boldsymbol{\varphi}}, \underline{\mathbf{h}})$ to (i) has mean similarity $\frac{1}{n} \sum_{i=1}^{n} \rho(\boldsymbol{\varphi}_{\lambda(\underline{\boldsymbol{\varphi}}, \boldsymbol{\chi}_i)}, \boldsymbol{\chi}_i \circ h_i)$ equal to 1 if and only if it is possible to perfectly align and cluster in k groups the set of n curves, i.e. if and only if there exists $\underline{\mathbf{h}} = \{h_1, \ldots, h_n\} \subset W$ and a partition $\mathcal{P} = \{P_1, \ldots, P_k\}$ of $\{1, \ldots, n\}$ in k elements, such that $\rho(\boldsymbol{\chi}_i \circ h_i, \boldsymbol{\chi}_j \circ h_j) = 1$ for all i and j belonging to the same element of $\mathcal{P}$. Because of (2.6), this is equivalent to the existence of k characteristic shape curves, $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_k$, such that

$$\forall\, i = 1, \ldots, n, \quad \exists\, l_i \in \{1, \ldots, k\} : \ \boldsymbol{\chi}_i \in \mathrm{SIM}(\boldsymbol{\varphi}_{l_i}). \tag{2.10}$$

In this case the optimization problem (i) is solved by setting $\boldsymbol{\varphi}_{\lambda(\underline{\boldsymbol{\varphi}},\boldsymbol{\chi}_i)} \equiv \boldsymbol{\varphi}_{l_i}$. It should also be noted that, thanks to property 3., if $\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_k\}$ and $\{h_1, \ldots, h_n\}$ provide a solution to (i), then also the quantities $\{\boldsymbol{\varphi}_1 \circ g_1, \ldots, \boldsymbol{\varphi}_k \circ g_k\}$ and $\left\{h_1 \circ g_{\lambda(\underline{\boldsymbol{\varphi}},\boldsymbol{\chi}_1)}, \ldots, h_n \circ g_{\lambda(\underline{\boldsymbol{\varphi}},\boldsymbol{\chi}_n)}\right\}$ are a solution to (i), for any $\{g_1, \ldots, g_k\} \subset W$. In other words, if we have a solution of the maximization problem (i), we can obtain another solution by simply warping each cluster (i.e. the functions belonging to the cluster and the relevant template curve) along an arbitrary affine transformation in W. Moreover, this solution identifies the same clusters (i.e., is associated to the same partition $\mathcal{P} = \{P_1, \ldots, P_k\}$ of $\{1, \ldots, n\}$).

The non-linear optimization problem (i) is not analytically solvable in its complete generality. Numerical methods can be proposed for its solution, for instance employing mixing of discrete and continuous optimization algorithms. Following a statistical perspective, we here prefer to give an approximate solution to the optimization problem, integrating a classical functional k-mean or k-medoid clustering algorithm (described in Section 1.2) and a Procrustes procedure for curve alignment (described in Section 1.3). The resulting iterative procedure simultaneously deals with (i) and (ii) alternating *template identification steps*, *cluster assignment steps* and *alignment steps* (see [67, 68] for further details). In the template identification step we estimate the set of k templates associated to the k clusters identified at the previous assignment and alignment step; in the assignment and alignment step, we align the n curves to the set of the k templates obtained in the previous template identification step, and we assign each of the curves to one of the k clusters. As we shall see, the proposed clustering and alignment procedure also considers the problem of non-uniqueness of the solution, by targeting a specific solution via a *normalization step*.

2.2.1 Clustering and alignment iterative procedure

Let $\underline{\boldsymbol{\varphi}}^{[q-1]} = \{\boldsymbol{\varphi}_1{}^{[q-1]}, \ldots, \boldsymbol{\varphi}_k{}^{[q-1]}\}$ be the set of templates after iteration $q-1$, and $\{\boldsymbol{\chi}_1{}^{[q-1]}, \ldots, \boldsymbol{\chi}_n{}^{[q-1]}\}$ be the n curves aligned and clustered to $\underline{\boldsymbol{\varphi}}^{[q-1]}$. At the qth iteration the algorithm performs the following steps.

Template identification step. For $j = 1, \ldots, k$, the template of the jth cluster, $\boldsymbol{\varphi}_j{}^{[q]}$, is estimated using all curves assigned to cluster j at iteration $q-1$, i.e. all curves $\boldsymbol{\chi}_i{}^{[q-1]}$ such that $\lambda(\underline{\boldsymbol{\varphi}}^{[q-1]}, \boldsymbol{\chi}_i{}^{[q-1]}) = j$. Ideally, the template $\boldsymbol{\varphi}_j{}^{[q]}$ should be estimated as the curve $\boldsymbol{\varphi} \in E$ that maximizes the within-cluster total similarity

$$\boldsymbol{\varphi}_j{}^{[q]} = \arg\max_{\boldsymbol{\varphi} \in E} \sum_{i:\lambda(\underline{\boldsymbol{\varphi}}^{[q-1]}, \boldsymbol{\chi}_i{}^{[q-1]})=j} \rho(\boldsymbol{\varphi}, \boldsymbol{\chi}_i{}^{[q-1]}), \tag{2.11}$$

i.e., $\boldsymbol{\varphi}_j{}^{[q]}$ should be the estimate from the functional dataset of the functional median, or Frchet median, associated to the similarity ρ according to the definition given in (1.7).

Assignment and alignment step. The set of curves $\{\boldsymbol{\chi}_1{}^{[q-1]}, \ldots, \boldsymbol{\chi}_n{}^{[q-1]}\}$ is clustered and aligned to the set of templates $\underline{\boldsymbol{\varphi}}^{[q]} = \{\boldsymbol{\varphi}_1{}^{[q]}, \ldots, \boldsymbol{\varphi}_k{}^{[q]}\}$: for $i =$

$1, \ldots, n$, the i-th curve $\boldsymbol{\chi}_{i^{[q-1]}}$ is aligned to $\boldsymbol{\varphi}_{\lambda(\underline{\boldsymbol{\varphi}}^{[q]}, \boldsymbol{\chi}_{i}{}^{[q-1]})}$ and the aligned curve $\tilde{\boldsymbol{\chi}}_{i^{[q]}} = \boldsymbol{\chi}_{i^{[q-1]}} \circ h_{i^{[q]}}$ is assigned to cluster $\lambda(\underline{\boldsymbol{\varphi}}^{[q]}, \boldsymbol{\chi}_{i}{}^{[q-1]}) \equiv \lambda(\underline{\boldsymbol{\varphi}}^{[q]}, \tilde{\boldsymbol{\chi}}_{i}{}^{[q]})$.

Normalization step. After each assignment and alignment step, we also perform a normalization step. In detail, for $j = 1, \ldots, k$, all the $N_{j^{[q]}}$ curves $\tilde{\boldsymbol{\chi}}_{i^{[q]}}$ assigned to cluster j are warped along the warping function $(\bar{h}_{j^{[q]}})^{-1}$, where

$$\bar{h}_{j^{[q]}} = \frac{1}{N_{j^{[q]}}} \sum_{i:\lambda(\underline{\boldsymbol{\varphi}}^{[q]}, \tilde{\boldsymbol{\chi}}_{i}{}^{[q]})=j} h_{i^{[q]}} \tag{2.12}$$

obtaining $\boldsymbol{\chi}_{i^{[q]}} = \tilde{\boldsymbol{\chi}}_{i^{[q]}} \circ (\bar{h}_{j^{[q]}})^{-1} = \boldsymbol{\chi}_{i^{[q-1]}} \circ h_{i^{[q]}} \circ (\bar{h}_{j^{[q]}})^{-1}$. In this way, at each iteration, the average warping undergone by curves assigned to cluster j is the identity transformation $h(t)=t$. Indeed:

$$\frac{1}{N_{j^{[q]}}} \sum_{i:\lambda(\underline{\boldsymbol{\varphi}}^{[q]}, \boldsymbol{\chi}_{i}{}^{[q]})=j} \big(h_{i^{[q]}} \circ (\bar{h}_{j^{[q]}})^{-1}\big)(t) = t, \qquad j = 1, \ldots, k. \tag{2.13}$$

The normalization step is thus used to select, among all candidate solutions to the optimization problem, the one that leaves the average locations of the clusters unchanged, thus avoiding the drifting apart of clusters or the global drifting of the overall set of curves. Note that the normalization step preserves the clustering structure chosen in the maximization step, i.e., $\lambda(\underline{\boldsymbol{\varphi}}^{[q]}, \tilde{\boldsymbol{\chi}}_{i}{}^{[q]}) = \lambda(\underline{\boldsymbol{\varphi}}^{[q]}, \boldsymbol{\chi}_{i}{}^{[q]})$ for all i.

The algorithm is initialized with a set of initial templates, $\underline{\boldsymbol{\varphi}}^{[0]} \subset E$, and with $\{\boldsymbol{\chi}_{1}{}^{[0]}, \ldots, \boldsymbol{\chi}_{n}{}^{[0]}\} = \{\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n\}$, and stopped when, in the assignment and alignment step, the increments of the similarity indexes are all lower than a fixed threshold.

2.3 Template identification

Whilst the assignment and alignment step and the normalization step are less troublesome, the template identification step raises more issues from a theoretical and computational point of view. In fact, the identification of the template $\boldsymbol{\varphi}_{j^{[q]}}$, as the curve $\boldsymbol{\varphi} \in E$ that maximizes the total similarity (2.11), cannot be easily dealt with. For this reason, in [68] we proposed to estimate the template $\boldsymbol{\varphi}_{j^{[q]}}$ as a loess, with Gaussian kernel and appropriate smoothness parameter, of the curves assigned to cluster j at iteration $q-1$ (i.e. all curves $\boldsymbol{\chi}_{i^{[q-1]}}$ such that $\lambda(\underline{\boldsymbol{\varphi}}^{[q-1]}, \boldsymbol{\chi}_{i}{}^{[q-1]}) = j$). See [68] for details on the implementation. The algorithm obtained with this specification for the template identification step was named *k-mean alignment*, in analogy with the functional k-mean clustering algorithm described in subsection 1.2.1. Note that the identification of the template via a point–wise functional mean of the elements belonging to the cluster would not be a good choice for k-mean alignment algorithm, due to the fact that when computing the template a proper alignment for the curves has not been achieved yet: a more stable procedure such as loess provides far better estimates.

On the other hand, the fact that the template is estimated by loess of the curves assigned to the cluster, instead of the curve that maximize the total

similarity (2.11), raises doubts about a possible bias of the algorithm. Moreover, estimating the template by loess of the curves assigned to the cluster might make this step sensitive to the presence of anomalous data.

For this reason, in [67] we proposed an alternative specification of the template identification step, that constitutes a direct approximation to the maximization of the total similarity (2.11). In particular, we restrict the set over which the maximization is carried out, limiting the search to functions of the sample; this is exactly the same approximation proposed in (1.8) for the computation of the functional median starting from a given functional sample (see subsection 1.1.1 for details). The template $\boldsymbol{\varphi}_{j\,[q]}$ is thus estimated as the curve $\boldsymbol{\varphi}$, among all curves assigned to cluster j at iteration $q-1$ (i.e. all curves $\boldsymbol{\chi}_{i\,[q-1]}$ such that $\lambda(\underline{\boldsymbol{\varphi}}_{[q-1]}, \boldsymbol{\chi}_{i\,[q-1]}) = j$), that maximizes the total similarity (2.11)

$$\boldsymbol{\varphi}_{j\,[q]} = \underset{\boldsymbol{\chi}_{i\,[q-1]}:\lambda(\underline{\boldsymbol{\varphi}}_{[q-1]},\boldsymbol{\chi}_{i\,[q-1]})=j}{\arg\max} \sum_{k:\lambda(\underline{\boldsymbol{\varphi}}_{[q-1]},\boldsymbol{\chi}_{k\,[q-1]})=j} \rho(\boldsymbol{\chi}_{i\,[q-1]}, \boldsymbol{\chi}_{k\,[q-1]}).$$

Note that the k curves selected as templates, in the template identification step, shall skip the subsequent assignment and alignment step and normalization step.

The algorithm obtained with this specification for the template identification step will be named *k-medoid alignment*, in analogy with the functional k-medoid clustering algorithm described in subsection 1.2.2, and with multivariate k-medoid clustering described for instance by Kaufman and Rousseeuw [36] or Hastie *et al.* [29, 30].

The same analogy suggests that this novel specification of the template identification step is less sensitive to the presence of anomalous data. In particular, the comparison of the clustering results obtained with the two alternative algorithm versions, the k-mean and the k-medoid version, might indicate accidental anomalous data, as we shall see in the applications to real data described in the following sections. Moreover, thanks to these alternative specifications, we will have the possibility of highlighting the robustness of our alignment and clustering procedure when different algorithm specifications are considered.

2.4 Case study: the analysis of growth curves

In this section we apply k-mean and k-medoid alignment to the analysis of a benchmark data set in the functional data analysis literature: the 93 growth curves from Berkeley Growth Study (see Tuddenham and Snyder [85]). These data have been previously considered by a number of authors (see for example Ramsay and Li [59], Ramsay and Silverman [61], James [35], and references therein). In the present discussion we will give some interesting insight on joint clustering and alignment with respect to sole clustering or sole alignment.

The heights (in cm) of the 93 children in the data set are measured quarterly from 1 to 2 years, annually from 2 to 8 years and biannually from 8 to 18 years. The growth curves are estimated by means of monotonic cubic regression splines (see Ramsay and Silverman [61]), implemented using the R function `smooth.monotone` available in the `fda` package [62]. Figure 1.2 shows the estimated growth curves and their derivatives, the growth velocities. Looking at the

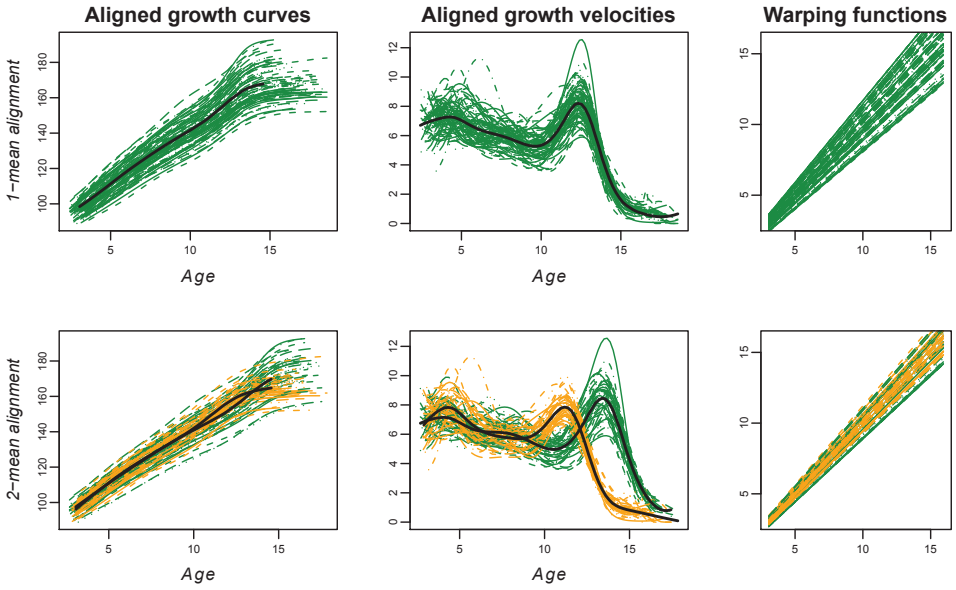

Figure 2.1: Results of k-mean alignment of growth curves, for k=1; 2: aligned growth curves (with superimposed estimated templates, in black) and corresponding growth velocities (with superimposed first derivatives of estimated templates, in black), together with warping functions. The colors of aligned curves and warping functions depend on cluster assignment.

growth velocities, it is apparent that the children follow a similar growth course, but that each child has a personal biological clock.

Figure 2.1 shows 1-mean and 2-mean aligned growth curves, the corresponding growth velocities and warping functions; Figure 2.2 shows the corresponding results obtained by k-medoid alignment. Results seem quite robust with respect to the different algorithm specifications.

Figure 2.3, left, displays in orange (blue), the boxplots of the similarity indexes between the original growth curves and their mean (medoid) curve, indicated with "orig", and the boxplots of the similarity indexes between the k-mean aligned (k-medoid aligned) growth curves and their estimated templates, for $k = 1, 2, 3$. The right panel of Figure 2.3 displays the corresponding means of similarity indexes. From inspection of the similarity indexes, both k-mean and k-medoid alignment suggest the presence of just one characteristic curve, since the choice of $k=2$ is not payed off by a reasonable gain in the similarities.

Since, out of the 93 children, 39 are boys and 54 are girls, we might wonder if the analysis points out some differences among them (notice that here we are not performing any supervised classification of boys and girls). Figure 2.4 is obtained from Figure 2.2 (top panels) displaying in blue the 1-medoid aligned growth curves of boys, and the corresponding growth velocities and warping functions, and in pink the ones of girls. The warping functions show a pretty neat separation of boys and girls in the phase; this highlights that the biological clocks of boys and

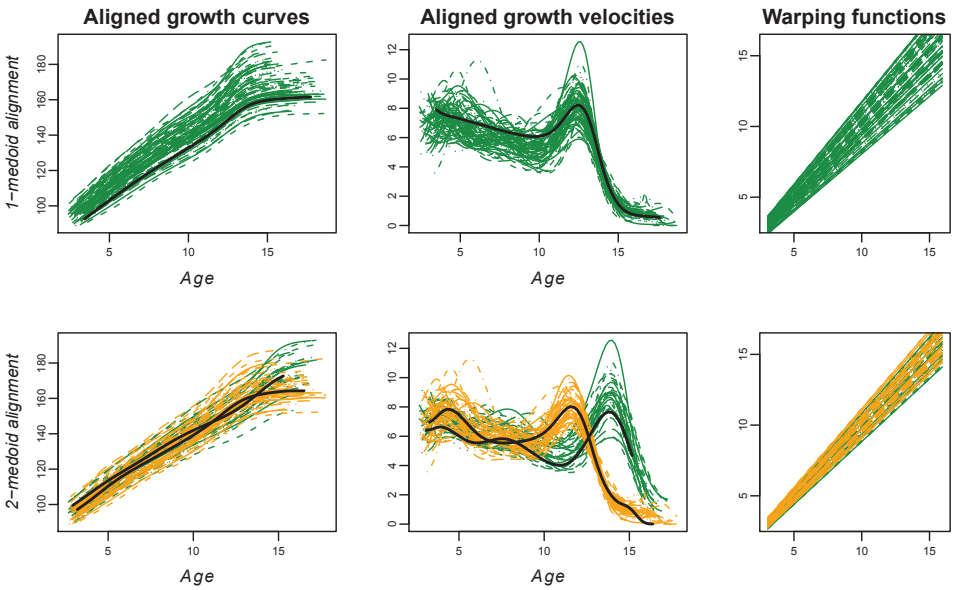

Figure 2.2: Results of k-medoid alignment of growth curves, for k=1; 2: aligned growth curves (with superimposed estimated templates, in black) and corresponding growth velocities (with superimposed first derivatives of estimated templates, in black), together with warping functions. The colors of aligned curves and warping functions depend on cluster assignment.

girls run at different speeds, and in particular that boys start to grow later, having warping functions with smaller intercepts, and grow slowler, having warping functions with smaller slopes. The left panel of Figure 2.4 also shows that, once the biological clocks of the children have been aligned, the height of boys stochastically dominates that of girls for any registered biological age. Finally, boys seem also to have a more pronounced growth, especially during puberty, as highlighted by their more prominent growth velocity peak. All these features are in complete agreement with the results obtained by 1-mean alignment, which are displayed in Figure 2.5.

Table 2.1: Left: results of 2-mean bivariate clustering of slopes and intercepts of the warping functions obtained by 1-mean alignment. Right: results of 2-medoid bivariate clustering of slopes and intercepts of the warping functions obtained by 1-medoid alignment. Cluster assignment vs gender.

2-mean		clusters	
		1	2
gender	F	44	10
	M	1	38

2-medoid		clusters	
		1	2
gender	F	43	11
	M	1	38

The grouping structure of the warping functions obtained by 1-mean and 1-medoid alignment of the growth curves, can be explored by a coherent unsu-

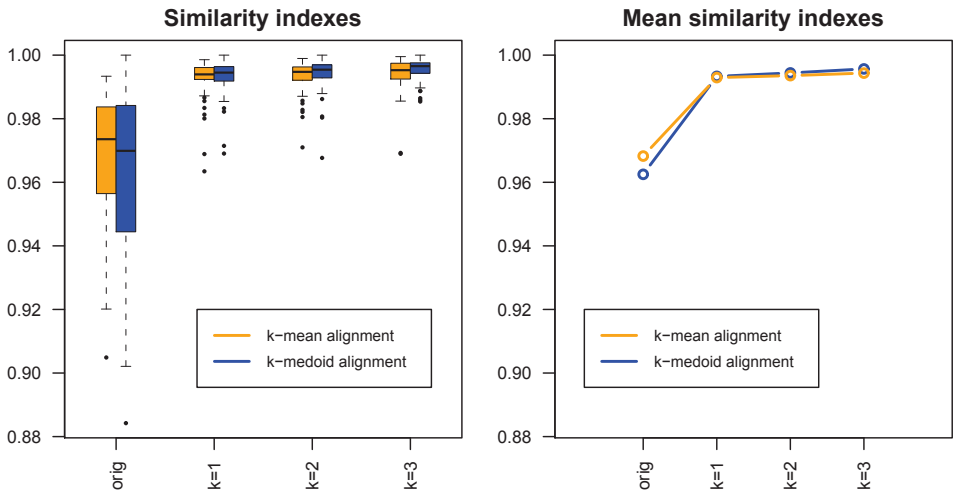

Figure 2.3: Left: in orange (blue), boxplots of similarity indexes between the original growth curves and their mean (medoid) curve, "orig", and boxplots of similarity indexes between the k-mean aligned (k-medoid aligned) growth curves and their estimated templates, for $k = 1, 2, 3$. Right: corresponding means of similarity indexes.

pervised classification of their slopes and intercepts, i.e. by 2-mean and 2-medoid bivariate clustering respectively. The results of these unsupervised classifications are shown in Table 2.1. Note that both 2-mean clustering of the warping functions obtained by 1-mean alignment of growth curves, and 2-medoid clustering of the warping functions obtained by 1-medoid alignment of growth curves, assign 1 boy to the female prevalent cluster, and respectively 10 and 11 girls to the male prevalent cluster. From inspection of Table 2.2, which compares the cluster assignments of the two procedures, we conclude that this boy and the 10 girls are exactly the same. Thus, both algorithms agree that these ten girls have biological clocks closer to those of boys, and that the boy has a biological clock closer to those of girls. This fact is evident in Figure 2.6, which displays the slopes and intercepts of the warping functions (pink for girls and blue for boys), and highlights in red the ones of the ten girls and in green the one of the boy. Figure 2.6 also displays in black the only mismatch between the two algorithms, which has in fact a biological clock borderline between the two groups.

It is interesting to remark that in this application, despite the presence of two groups in the dataset (boys and girls), k-mean alignment apparently suggests growth curves being realizations of the same shape invariant model, while a grouping structure is clearly detectable in the warping functions describing phase variability. This conclusion is quite different from the one discussed in the morphological analysis of ECG traces (see Section 1.4): in that application of clustering and alignment, indeed, the different pathologies induced a grouping structure both in the phase and in the amplitude variability of ECGs.

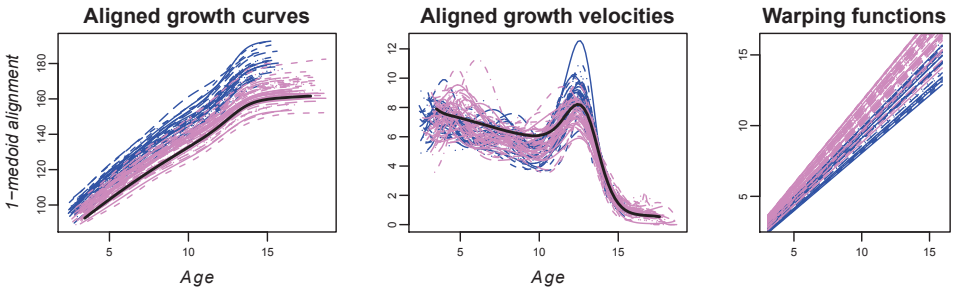

Figure 2.4: Figure obtained from Figure 2.2 (top panels) displaying in blue the growth curves, growth velocities and warping functions of boys and in pink the ones of girls.

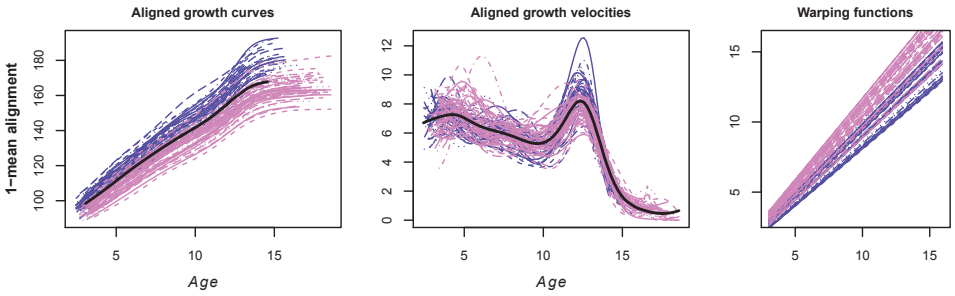

Figure 2.5: Figure obtained from Figure 2.1 (top panels) displaying in blue the growth curves, growth velocities and warping functions of boys and in pink the ones of girls.

Table 2.2: A comparison between cluster assignments obtained by 2-mean and 2-medoid bivariate clustering of slopes and intercepts of the warping functions obtained by 1-mean alignment and 1-medoid alignment of growth curves.

		2-medoid	
		1	2
2-mean	1	44	1
	2	0	48

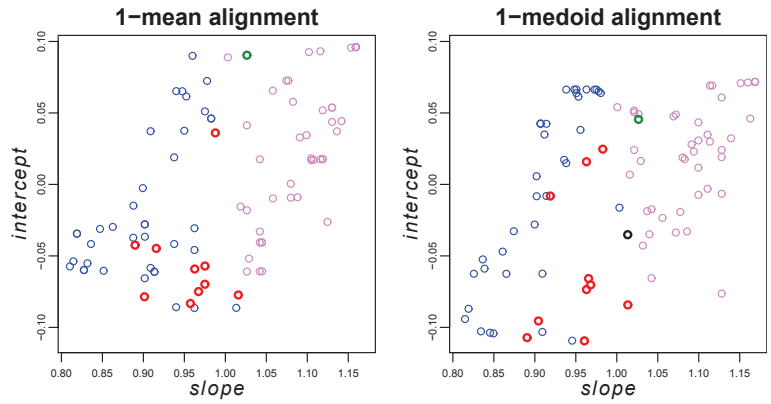

Figure 2.6: Slopes and intercepts of the warping functions resulting from 1-mean alignment (left) and 1-medoid alignment (right) of growth curves. Blue circles correspond to boys and pink circles to girls. Red (green) circles are girls (boys) which in both cases have been assigned to the male (female) prevalent cluster when the corresponding bivariate clustering procedure is applied. The black circle is the only mismatching case between the two procedures (see Table 2.2).

Chapter 3

Sparse Clustering of Functional Data

The focus of this Chapter is on *sparse clustering of functional data*, i.e. on the definition of a proper methodology able to cluster functional data and jointly to select the features which are the most relevant to the clustering scope. Sparse clustering methods for data belonging to a finite dimensional space have already been proposed: indeed, sparse clustering is a widely studied problem in the statistical literature, especially in the unsupervised learning. When dimensional reduction is the major focus, a common approach to non-parametric classification is based on Principal Component Analysis; indeed, sparse PCA clustering methods for vector data have been recently proposed, see for instance [51] and references therein. Model-based approaches have also been considered, for instance, by McLachlan and Peel [52] and Fraley and Raftery [20], who assume data to be generated from a mixture of Gaussian distributions with k components. Many non–parametric sparse clustering methods are instead based on modifications of non–sparse ones, with the introduction of a weighting vector for feature selection. Among the others, Friedman and Meulman [21] proposed *Clustering Objects on Subsets of Attributes (COSA)*, in which feature selection is based on a weighting vector, and different feature weights within each cluster are allowed. Finally, Tibshirani and Witten [83], introduced a weighted k-mean method, which seems to perform quite well in most cases. This last recent work is particularly interesting to the scopes of the present dissertation, since as most of the methods previously considered, it is a modification of the k-mean algorithm.

Tibshirani and Witten [83] propose, in fact, a generalization of the weighted k-mean algorithm, in which a vector of positive weights is responsible for the identification of the most relevant features for the classification. This weighting vector is identified as the solution to a constrained optimization problem, in which the objective function depends on the between cluster sum of squares of the weighted variables (with the drivng idea that the greater is the discriminatory power associated to the variable, the greater will be the corresponding weight to maximize the objective). Tibshirani and Witten prove existence and uniqueness of the solution of the sparse clustering optimization problem, and they are also able to find an analitical solution closely related to LASSO method.

Many clustering methods suited for the case of functional data have already been discussed in the previous Chapters, e.g. to capture phase variability while clustering functional data. In the spirit of Tibshirani and Witten [83], we here introduce a method for the sparse classification of functional data, based on a weighted k-mean. The generalization of the sparse clustering problem to the functional case leads us to consider a positive weighting function $\omega(\cdot)$, instead of a positive weighting vector $\mathbf{w}$. The constrained optimization problem which defines the sparse clustering for functional data can then be described as a variational problem, whose solution exists and is unique. Many are the reasons to deepen ou insight on clustering functional data in this direction; among the others, we cite only the most relevant considerations:

- improving clustering results: assuming that a small number of features separates the clusters, sparse clustering might result in a more accurate identification of the groups when compared with standard clustering, since it removes ancillary information;
- dimensional reduction: especially in the functional case, the role played by the weighting function is fundamental to select the most relevant fetures;
- improving the interpretation of the results: the weighting function $\omega(\cdot)$ also helps the interpretation of the procedure that leads to the formation of different clusters, and it thus gives a simpler way to recognize them in future applications.

The Chapter is organized as follows. In Section 3.1 we introduce the problem of sparse clustering considering first the fnite dimensional case. We will focus, in particular, on the method proposed by Tibshirani and Witten in [83]. In Section 3.2 we define the theoretical setting and the assumptions leading to our method, focussing on the meaning of features selection for functional data, and we give the main result concerning functional sparse clustering, i.e. the existence and uniqueness of the solution to the variational problem defining sparse k-mean for functional data; moreover we here present a representation for the optimal weighting function in the case of continuous data. Finally, an application to the Berkeley Growth Study data is illustrated in Section 3.3, where we compare the classification of growth curves obtained by sparse k-mean clustering to that illustrated in the previous Chapter (see Sangalli *et al.* [67, 68] for details) as a result of a k-mean alignment procedure.

3.1 Multivariate Sparse Clustering

It is not an straightforward to extend to functional data the sparse k-mean method proposed in Tibshirani and Witten [83] for finite dimensional data. To grasp differences and analogies in the two mathematical settings, we deem it important to briefly recall the method proposed by Tibshirani and Witten.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with $\mathbf{x}_i \in \mathbb{R}^p$, be the data and set

$$d_{i,i',j} = \left(x_{i,j} - x_{i',j}\right)^2,$$

to be the distance relative to the j-th feature, between observation $\mathbf{x}_i$ and observation $\mathbf{x}_{i'}$. A k-mean problem can be restated as the maximization of the distance between clusters

$$\max_{(C_1,\dots,C_k)} \sum_{j=1}^{p} \left(\frac{1}{2n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{i,i',j} - \sum_{h=1}^{k} \frac{1}{2n_h} \sum_{i,i' \in C_h} d_{i,i',j} \right), \tag{3.1}$$

where n_h is the cardinality of the h-th cluster and the maximum is taken over the set of all possible clusters $C_1, \dots, C_k$ grouping data in the sample. *Sparse k-mean Clustering* aims at selecting only the features relevant for grouping the data; it is the solution to the following maximization problem:

$$\begin{aligned} \max_{(C_1,\dots,C_k),\mathbf{w}} & \sum_{j=1}^{p} w_j \left(\frac{1}{2n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{i,i',j} - \sum_{h=1}^{k} \frac{1}{2n_h} \sum_{i,i' \in C_h} d_{i,i',j} \right) \\ \text{subject to} \quad & \|\mathbf{w}\|_{\ell_2}^2 \leq 1,\ \|\mathbf{w}\|_{\ell_1} \leq s,\ w_j \geq 0\ \forall j. \end{aligned} \tag{3.2}$$

Here w_j is the j-th component of a weighting vector $\mathbf{w}$ that has the role of feature selector, achieved by the constraint on the ℓ_1 norm. The other constraints are obvious: $\mathbf{w}$ must have at most unitary Euclidean norm and be nonnegative. The *tuning parameter* s is specified by outer information or is to be found through other methods (see Tibshirani and Witten [83] for further details).

In Tibshirani and Witten [83], an iterative algorithm for the solution to the optimization problem (3.2) is proposed. The authors propose to optimize (3.2) by alternatively optimizing with respect to $(C_1, \dots, C_k)$, holding $\mathbf{w}$ fixed, and optimizing with respect to $\mathbf{w}$, holding $(C_1, \dots, C_k)$ fixed. In general, using iterative approaches like this one, a global optimum of (3.2) will not be achieved; however, we are at least guaranteed that each iteration increases the objective function.

The first optimization involves application of a standard clustering procedure to a weighted version of the data (the authors propose to use both k-mean and hierarchical procedures). More interestingly, to optimize (3.2) with respect to $\mathbf{w}$ with $(C_1, \dots, C_k)$ held fixed, the problem can be rewritten as

$$\begin{aligned} & \max_{\mathbf{w}} \mathbf{w}^T \mathbf{a} \\ \text{subject to} \quad & \|\mathbf{w}\|_{\ell_2}^2 \leq 1,\ \|\mathbf{w}\|_{\ell_1} \leq s,\ w_j \geq 0\ \forall j, \end{aligned} \tag{3.3}$$

where $\mathbf{a} \in \mathbb{R}^p$ and its j-th coordinate has the following expression $a_j = \frac{1}{2n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{i,i',j} - \sum_{h=1}^{k} \frac{1}{2n_h} \sum_{i,i' \in C_h} d_{i,i',j}$, for $j = 1, \dots, p$.

In Tibshirani and Witten [83] an analytical solution to the optimization problem (3.3) is presented, which is based on Soft-thresholding. For $x \in \mathbb{R}$ and $c \geq 0$, define

$$S(x, c) = \text{sign}(x)(|x| - c)_+$$

to be the Soft-thresholding operator; for ease of notation we allow for the operator S to be applied component-wise to vectors of $\mathbb{R}^p$.

Proposition 3.1.1 (Tibshirani and Witten [83]).
For $1 \leq s \leq \sqrt{p}$, a solution to problem (3.3) is

$$\boldsymbol{w} = \frac{S(\boldsymbol{a}_+, \Delta)}{\|S(\boldsymbol{a}_+, \Delta)\|_{\ell^2}},$$

where x_+ denotes the positive part of $x \in \mathbb{R}$ and $\Delta = 0$ if that results in $\|\boldsymbol{w}\|_{\ell_1} \leq s$; otherwise, $\Delta > 0$ is chosen to yield $\|\boldsymbol{w}\|_{\ell_1} = s$.

The proof of this result is almost straightforward: problem (3.3) could be restated as a linear maximization problem over a convex domain. Then, Karush-Kuhn-Tucker Theorem is applied (see Tibshirani [82] for details). We stress the fact that unicity could not be achieved; indeed, the authors assume that there exists a unique maximal disposition of the data in the clusters.

3.2 Functional Sparse Clustering

We aim now at focussing on the functional case. Consider a functional dataset $\chi : T \to \mathbb{R}$, where T is a Borel subset of $\mathbb{R}$ and $\chi \in L^2(T)$ with respect to a measure μ defined on the Borel subsets of $\mathbb{R}$ and such that $\mu(T) < \infty$ (most often μ is the Lebesgue measure). We here assume that data need not be aligned or have are already been aligned: for details on alignment procedures see Chapter 1, and for a focus on clustering and alignment of functional data see Chapter 2 [1].

3.2.1 Theoretical Setting

For functional data we need to give a meaning to the concept of feature selection. When data are elements of $\mathbb{R}^p$, in its complete generality feature selection means finding a subset D of the set of p features, such that D contains only those that discriminate the most between clusters; in principle one could inspect the $2^p - 1$ non void subsets of the p features and find an optimum subset where, for instance, BCSS is maximized. We can't follow this lead for functional data, since the set of features T is usually a continuum.

However, in the case of finite dimensional data, feature selection is eventually carried out by an optimal weighting vector $\mathbf{w}$ (Proposition 1). In the functional setting the role of the weighting vector in performing feature selection can be played by a weighting function

$$w : T \to \mathbb{R},$$

satisfying some constraints that will follow after setting the variational problem that defines analytically *Functional Sparse k-mean Clustering.*

[1] The problem of sparse clustering and alignment of functional data has not yet been considered in the literature on functional data analysis. Indeed, its mathematical formulation could result in a unified definition of the clustering problem for functional data, including both the selection of a proper level of sparsity for the clusterization, and the decoupling of phase and amplitude variability, the former being ancillary and the latter relevant to the scopes of classification

Let $\chi_1, ..., \chi_n$ represent the functional dataset and define the Total Sum of Squares to be

$$TSS = \frac{1}{2n} \sum_{i,j=1}^{n} \int_T (\chi_i - \chi_j)^2 \tag{3.4}$$

(here and in the following, all integrals are taken with respect to the measure μ). When data labels are partitioned into the clusters $C_1, ..., C_k$, let the Within Clusters Sum of Squares be

$$WCSS = \sum_{h=1}^{k} \frac{1}{|2C_h|} \sum_{i,j \in C_h} \int_T (\chi_i - \chi_j)^2 , \tag{3.5}$$

where $|C_h|$ is the cardinality of the cluster C_h. Finally define the Between Clusters Sum of Square as the difference

$$BCSS = TSS - WCSS.$$

The standard k-mean algorithm for functional data looks for a partition of the data labels that solves the following maximization problem (see Tarpey and Kinateder [80] for details on the mathematical definition of the functional k-mean):

$$\max_{(C_1,...,C_k)} \int_T \left\{ \frac{1}{2n} \sum_{i,j=1}^{n} (\chi_i - \chi_j)^2 - \sum_{h=1}^{k} \frac{1}{|2C_h|} \sum_{i,j \in C_h} (\chi_i - \chi_j)^2 \right\}. \tag{3.6}$$

The natural generalization to sparse k-mean clustering is thus the following variational problem:

$$\max_{w(x),(C_1,...C_k)} \int_T w(x) \left(\frac{1}{2n} \sum_{i,j=1}^{n} (\chi_i - \chi_j)^2 - \sum_{h=1}^{k} \frac{1}{|2C_h|} \sum_{i,j \in C_h} (\chi_i - \chi_j)^2 \right), \tag{3.7}$$
$$\text{subject to } \|w(x)\|^2_{L^2(T)} \leq 1, \ \|w(x)\|_{L^1(T)} \leq s, \ w(x) \geq 0 \ \mu - a.e.$$

We shall now discuss the constraints on the weighting function w. First we assume w to belong to the closed unitary ball of $L^2(T)$. Indeed, if the L^2 norm of w were not bounded, the solution to problem (3.7) would not exist: in fact, given a tentative solution w, one could always increase the value of the functional in (3.7) by multiplying w for an increasing constant. Hence every partition of the data labels would be optimal.

Moreover, w is defined on the same domain T of the functions belonging to the data set and it is nonnegative almost everywhere on T; ideally, if a Borel set $B \subset T$ is not relevant for cluster identification, i.e. data are equally distributed on that set, then $w(x) = 0$ for $x \in B$. On the contrary, if a subset of T has a role in the clustering, then w should be strictly positive on the subset and as much greater as that set is important for partitioning the data.

Finally, the constraint on w that truly characterizes the sparsity of the method: L^1 boundedness. Note that, by Hölder's inequality,

$$\|w(x)\|_{L^1(T)} = \int_T w(x) \leq \sqrt{\mu(T)} \|w(x)\|_{L^2(T)} \leq \sqrt{\mu(T)}. \tag{3.8}$$

As in the finite dimensional case, the constraint on the L^1 norm of the weight function is what generates sparsity in the clustering. Indeed, for $s > 0$ define the set

$$W_s = \left\{ u \in L^2(T) : \|u(x)\|_{L^1(T)} \leq s,\ \|u(x)\|^2_{L^2(T)} \leq 1, u(x) \geq 0\ \mu - a.e. \right\}, \tag{3.9}$$

and note that

$$W_t \subset W_s,\ \text{if } 0 < t < s. \tag{3.10}$$

By letting

$$W = \bigcup_s W_s,$$

where the union is taken over all s such that $0 < s \leq \sqrt{\mu(T)}$, we obtain

$$W = \left\{ u \in L^2(T) : u(x) \geq 0\ \mu - a.e.,\ \|u(x)\|^2_{L^2(T)} \leq 1 \right\},$$

which, in particular, contains the non-sparse standard k-mean solution. From this we deduce, that, the greater is s, the less sparsity will result in the final clustering and that the solution of non-sparse standard k-mean can be obtained as a limiting case of sparse clustering. Therefore, the tuning parameter s assumes the role of "index of sparsity" of the resulting clustering. This happens because the L^1 norm controls the amplitude and the extension, over the domain T, of the weighting function w.

3.2.2 Main Result

Our main result concerning the problem of sparse clustering for functional data states that the solution to the functional sparse k-mean clustering problem exists and is unique.

Theorem 3.2.1. *Fix the number of cluster $k \geq 1$ and let the data $\chi_1, ..., \chi_n$ be uniformly bounded. Then the solution to problem (3.7) exists and is unique.*

We omit the proof for the sake of brevity, leaving instead space to add some considerations; all the details can be found in Floriello *et al.* [18]. Theorem 3.2.1 is not only important because it proves the well-posedness of the variational problem (3.7), but also because its proof is constructive: the proof is based on the construction of a sequence of simple functions w_n, which is proven to converge to the solution w of the problem of sparse clustering defined in (3.7); indeed, this sequence of w_n's tells us how to build an algorithm performing sparse functional k-mean by iteratively searching for a solution to problem (3.7).

Moreover, we have an asymptotical form for the optimal weight function in the functional setting if we slightly strengthen the hypotheses on the data.

Theorem 3.2.2.
Let χ_i's be functions such that: $\chi_i \in L^2(T)$, $\chi_i \in L^\infty(T)$, with $\|\chi_i\|_{L^\infty(T)} \leq M$, $\forall i = 1, \ldots, n$. Suppose, moreover, $\chi_i \in C(T)$ for every $i = 1, \ldots, n$. Define

$$b(x) := \frac{1}{2n} \sum_{i,j=1}^{n} (\chi_i - \chi_j)^2 (x) - \sum_{h=1}^{k} \frac{1}{|2C_h|} \sum_{i,j \in C_h} (\chi_i - \chi_j)^2 (x), \tag{3.11}$$

which is well defined thanks to the hypotheses on the χ_i's and $b(x) \in L^2(T)$. Then

$$w(x) = \gamma b(x), \text{ for an appropriate } \gamma \in \mathbb{R}^+. \tag{3.12}$$

Thanks to the properties of the solution, we can say that Functional Features Selection, given a clustering problem and a measure space $(T, \mathcal{G}, \mu)$, means to find subsets D_i, i belonging to an index set I, $D_i \subset \mathcal{G}$, $\forall i \in I$, such that if we call $\bar{\chi}_1, \ldots, \bar{\chi}_k$ the mean functions of the k clusters, we have that for every pair of functions $\bar{\chi}_r$ and $\bar{\chi}_t$, $r, t = 1, \ldots, k$ with $r \neq t$ and for every $x \in D_i$, there exist sets A_r and A_t, not connected, such that $\bar{\chi}_r(x) \in A_r$ and $\bar{\chi}_t(x) \in A_t$.

Finally, observe that Theorem 3.2.1 is also useful in case of supervised learning. Indeed, the case in which we already know which functions belong to the different clusters is a simpler optimization problem, which is a particular case of problem (3.7): keeping the cluster assignments $C_1, \ldots, C_k$ fixed, we immediately get the right weight function discriminating between the groups.

3.2.3 An iterative algorithm implementing functional sparse k-mean clustering

We here propose a possible iterative algorithm implementing functional sparse k-mean clustering. The criterion used in our algorithm is that a weight should be assigned to a subset of the domain basing on the importance assumed by that particular subset in the classification, i.e. the increase in the BCSS that it could generate. Note that we have to maximize over the set of weighting functions and the set of all possible cluster assignments. To do this, we find the maximum of the objective functional (3.7) over the set of admissible functions, holding fixed the cluster assignments, then we search for the maximizing cluster assignments holding fixed the weight function w and proceed in this way iteratively. Unfortunately, this kind of procedure only assures that the value of the objective functional is increased at every step, but it does not assure that we achieve the global optimum in (3.7).

The algorithm needs an initial condition on $(C_1, \ldots, C_k)$. Then, at every step, it divides the domain into subsets, trying to detect the interval where the difference between the clusters is localized. According to this difference, it builds the maximizing weighting function and finds the maximizing cluster assignements.

Step 1

Initialize the algorithm by choosing $(C_1, \ldots, C_k)$ to be the solution of the functional k-mean clustering with the same data.

Step n

Divide the domain D into n subsets A_l of measure $\frac{\mu(D)}{n}$ and build the maximizing weighting function as a piecewise constant function on the n subdomains:

$$w_{n,\alpha}(x) = \alpha \sum_{l=1}^{n} c_l I_{A_l}(x), \tag{3.13}$$

where the coefficients c_l have the form:

$$c_l = \frac{1}{\mu(A_l)} \left(\frac{1}{2n} \sum_{i,j=1}^{n} \int_{A_l} (\chi_i - \chi_j)^2 - \sum_{h=1}^{k} \frac{1}{|2C_h|} \sum_{i,j \in C_h} \int_{A_l} (\chi_i - \chi_j)^2 \right) \quad (3.14)$$

where C_h is the h-th cluster coming from the $n-1$-th step of the algorithm. Having thus constructed the optimal weighting function, we seek for the optimal cluster assignment taking, as w, the function just found.

This iterative procedure is run until a stopping criterion, based on the difference between the weighting functions computed at two subsequent steps, is attained, or until a maximum number of iterations is reached.

3.3 Case study: the analysis of growth curves

In this section we illustrate the results obtained with sparse functional clustering on the growth curves included in the Berkeley Growth Study, a benchmark dataset for functional data analysis which is also provided in the `fda` package (see [58, 62]). We will here deepen the discussion on the role of the tuning parameter, s, and we will detail a possible criterium to estimate it. Moreover, we will compare the results obtained via sparse functional clustering with the ones obtained with the joint alignment and clustering method, k-mean alignment, first introduced in Sangalli *et al.* (2010) [68] and described in Chapter 2.

The Berkeley Growth Study dataset describes the heights (in cm) of 93 children, 54 girls and 39 boys, measured quarterly from 1 to 2 years, annually from 2 to 8 years and then biannually from 8 to 18 years. The reconstructed functional data we are going to consider in this Section are the same we have analyzed in Chapter 2: we will here consider both the reconstructed functional growth patterns (left panel in Figure 1.2) and the estimated growth first derivatives (right panel in Figure 1.2).

After having obtained regular (i.e. twice differentiable) curves, we can analyze them using sparse functional clustering techniques. It is reasonable that children of different sex will show different growth patterns, hence we expect the grouping structure will mainly reflect the gender membership. In other words, we are supposing that the height curves distinguish between girls and boys, and we want to test whether our algorithm is able to detect this grouping structure and to find the age in which the differences between boys and girls are more relevant.

The most important issue to be considered before running the sparse functional clustering algorithm consists in finding the right tuning parameter, defining sparsity. In all our discussion we have supposed s fixed; however, the way to determine the most suited value for this parameter in real problems deserves great attention. In lucky cases, a value for s could be assigned relying on specific outer information about the analyzed phenomenon. For instance, we may know the size of the subset of the domain where the functions have evident different shapes. In almost all other cases, we must turn to computational methods.

In Tibshirani and Witten [83] the authors propose an algorithm based on Gap Statistics to evaluate the optimal tuning parameter. This method is substantially

a comparison between the objective function of the sparse clustering result and the distribution of the same quantity after application of the procedure to B bootstrap samples, obtained by independently permuting the observations within each feature. Then one looks for the s^* which maximizes the Gap Statistics with respect to s. In this way, we look for the value of s that mostly expresses a sort of intrinsic clusterization the data are thought to contain. We can keep this method also in the functional setting, but with some modifications. The functional dataset, in practice, is a discretization of the observed continuous phenomenon; thus, we can apply the same method based on Gap Statistics, but we have to take into account the functional nature of the data by considering of the right inequalities between norms.

It is important to observe that, the more we increase s, the greater will be the value of the objective functional and the less the sparsity will result, leading to a method that is near to classical K-means. Therefore, we can give two interpretations of the tuning parameter: first of all, s is a measure of the sparsity of the data; moreover, $\frac{1}{s}$ is connected to the quantity of information at our disposal. Concerning the latter interpretation, indeed if little is known about the osberved phenomenon, or if the Gap Statistics method gives a high value for s (meaning that there are no evident intrinsic clusterizations), then $\frac{1}{s}$ is low, and the method is near to classical K-Means. On the contrary, if s has a low value, this means that there are evident clusterizations and that they are located in particular subsets. As a result, there may need a little effort in choosing the right L^1 constraint. A good weight function solution to the problem (3.7) should, besides helping the classification, identify subsets of the domain, if any, where clustering the functions is meaningless.

In the analysis of the growth curves from Berkeley Growth Study, we have no outer information helping us to choose s: it is not evident which time interval distinguishes the clusters the most, and also clinical considerations on the pubertal spurt can not provide useful insight, since this event is not easy to determine, it happens at different ages for different children and it could have different durations[2] . Therefore, we need other quantitative methods to estimate the right s. To this aim, we define an equally spaced grid of 100 points from 0 to 18 years, we evaluate the curves on this grid and collect the values in a matrix belonging to $\mathbb{R}^{93\times 100}$. We then apply a modified method with respect to the Gap Statistic discussed in Tibshirani and Witten [83], in order to take into account the functional nature of the original data and the right connections between the norms of functions involved. This procedure provides $s^* = 4.123106$, a value very close to the maximum allowable $\sqrt{18} \approx 4.242641$. This fact could suggest the absence of a portion of the domain where the clusters are well separated, as we could also observe looking at Figure 1.2, left panel. As a consequence, the weight function will be interpreted as the difference, in every point of the domain, between the two groups, meaning that eventual peaks or valleys indicate where the two clusters differ, respectively, the most or the least.

[2] The pubertal spurt is a sharp peak of growth velocity between 10 and 16 years, while the mid-spurt is a minor velocity peak between 2 and 5 years. Typically, different children have their growth spurts at different times and for different time periods, each child following his/her personal biological clock.

After fixing the tuning parameter s to the previously detected optimal value, we can run the sparse functional k-mean algorithm. The results are impressive. The resulting clusterization is coincident with the gender membership, as expressed by the perfectly diagonal confusion matrix in Table 3.1: the algorithm correctly groups in one cluster of cardinality 54 the growth curves of the girls, and in the other, of cardinality 39, the growth curves of the boys. In Figure 3.1 the mean functions obtained via sparse functional k-mean are reported for each of the clusters, in blue for boys and in pink for girls: as we could have expected, the mean curve of boys starts becoming higher than the one of girls after the pubertal spurt. Moreover, looking carefully at the Figure, a difference in the shape could be observed. Indeed, there is a small interval, centered just after 10 years, where the mean function of the girls has a negative curvature, whereas the mean function of the boys has a positive one.

	Boys	Girls
Cluster 1	39	0
Cluster 2	0	54

Table 3.1: Cluster assignments obtained via sparse functional k-mean algorithm ($k = 2$) vs gender membership.

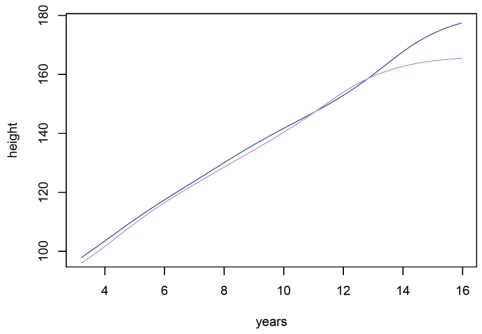

Figure 3.1: Mean functions for the two clusters obtained via sparse functional k-mean algorithm: the blue curve is the mean growth curve for boys, while the pink curve is the mean growth curve for girls.

In order to have a better insight on this characteristic, in Figure 3.2 the function $b(x)$ found via sparse functional k-mean ($k = 2$) on Berkeley Growth Study data is reported. It is interesting to consider the shape of this function: $b(x)$ has a nearly-constant pattern, and starts a monotonic growth in correspondence of the abscissa value of 13 years. This fact highlights the increase in the differences among children growth patterns after 13 years, a behaviour we also find in the medical literature on the dynamic of growth in children, characterized by the presence of the pubertal spurt. As confirmed by auxologists and evolutionists, girls grow faster than boys, and the peak of the process happens at about 12-13 years, age at which girls undergo to radical changes in their body. The same process happens later to boys, usually at 14-15 years. Thus, the increase in $b(x)$

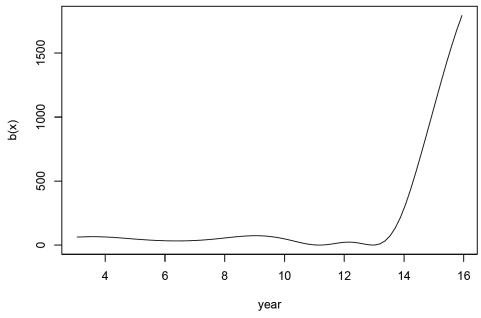

Figure 3.2: $b(x)$ obtained via sparse functional k-mean algorithm ($k = 2$) on Berkeley Growth Study data.

indicates the presence of the pubertal spurt, and suggests that, while girls are going through changes, boys are still rather late in their evolution. When time passes the differences given by gender in the dynamic of growth smooth, and this probably happens for greater ages with respect to the ones we are considering (3–16 years). We can only guess that $b(x)$ has a right horizontal asymptote, having a value given by the difference between the mean height of boys and girls. Indeed the monotonically increasing shape of $b(x)$ after 13 years suggests that height diversify between sexes during puberty because of a different type of evolution, which lasts until at least 16 years.

Finally, we discuss a comparison with the analysis of the Berkeley Growth Study dataset already reported in Sangalli *et al.* [67, 68] and in Chapter 2, where the clustering of growth curves is performed jointly with alignment of the curves themselves, via an innovative procedure named k-mean alignment.

The results of application of k-mean alignment to the Berkeley Growth Study dataset are reported in Figure 3.3, where the $k = 1$ and $k = 2$ aligned growth curves, growth velocities and warping functions are reported, in blue for boys and in pink for girls. As already commented in the last Section of Chapter 2, the boxplots of the similarity indexes and the corresponding mean similarities attained by k-mean alignment for $k=1, 2, 3$, reported in Figure 2.3, suggest that one cluster is the best choice for this dataset. Indeed, if we look at the results of 1-mean alignment (top panels in Figure 3.3) the warping functions show a pretty neat separation of boys and girls in the phase: this highlights that the biological clocks of boys and girls run at different speeds, as already pointed out by the previous analysis via sparse functional k-mean (boys start to grow later, having warping functions with smaller intercepts, and grow more slowly, having warping function with smaller slopes). Moreover, looking at the growth curves (top left panel), we note that the height of boys stochastically dominates the one of girls for any registered biological age.

Hence, the analysis via k-mean alignment reported in Chapter 2 suggested that the most relevant clustering structure in the Berkeley Growth Study dataset, induced by the gender membership, expresses itself in the phase variability, and can thus be captured by the alignment procedure. One can thus wonder if,

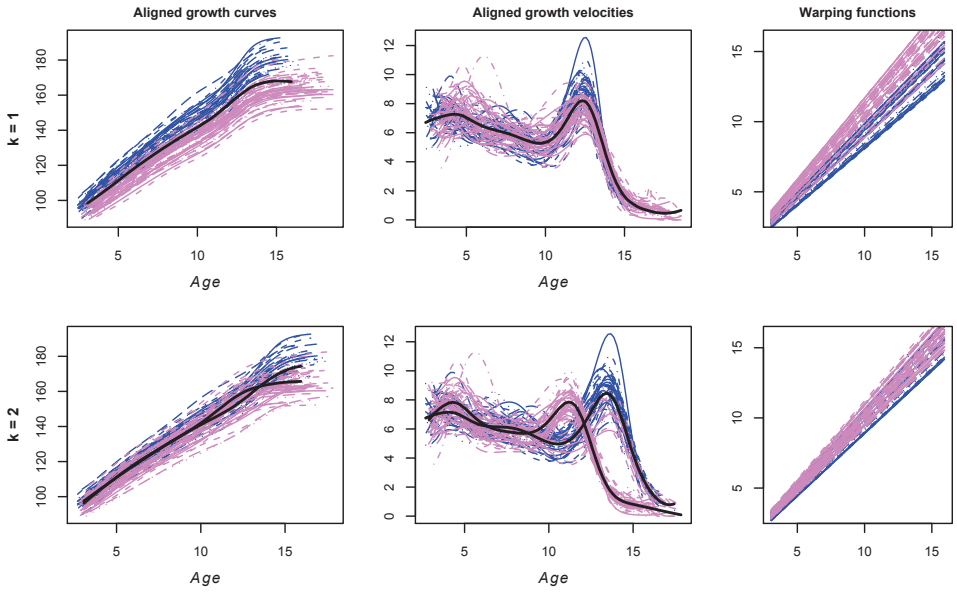

Figure 3.3: Results of k-mean alignment of growth curves, for $k = 1, 2$: aligned growth curves (with superimposed estimated templates, in black) and corresponding growth velocities (with superimposed first derivatives of estimated templates, in black), together with warping functions. The colors of aligned curves and warping functions depend on the gender: the growth curves, growth velocities and warping functions of boys are drawn in blue and the ones of girls in pink.

once the growth curves have been aligned via 1-mean alignment, other relevant clustering structures can emerge in the growth patterns. This problem can be addressed by functional sparse clustering. Indeed, in order to have a better insight on the clustering structure in the Berkeley Growth Study dataset, we also run the sparse functional k-mean algorithm on the set of 1-mean aligned growth curves shown in Figure 3.3 (top left panel).

Results are shown in Table 3.2 and in Figures 3.4 and 3.5. The resulting clusterization is again coincident with the gender membership (Table 3.2). In Figure 3.4 the mean functions obtained via sparse functional k-mean on 1-mean aligned curves are reported for each of the clusters, in blue for boys and in pink for girls: now that the children growth patterns have been registered, the height of boys is always higher than the one of girls, and the effect of the pubertal spurt has disappeared.

	Boys	Girls
Cluster 1	39	0
Cluster 2	0	54

Table 3.2: Cluster assignments obtained via sparse functional k-mean algorithm ($k = 2$) on 1-mean aligned growth curves vs gender membership.

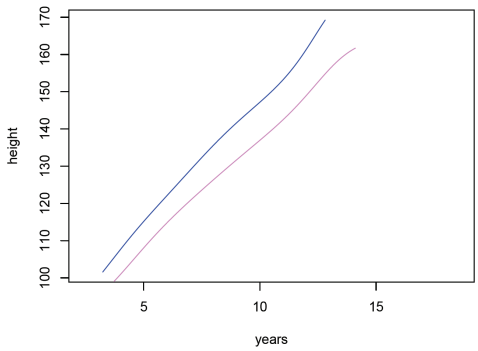

Figure 3.4: Mean functions for the two clusters obtained via sparse functional k-mean algorithm on 1-mean aligned growth curves: the blue curve is the mean growth curve for boys, while the pink curve is the mean growth curve for girls.

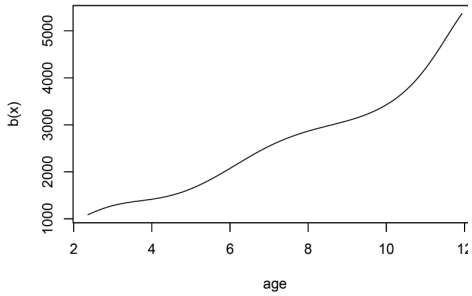

Figure 3.5: $b(x)$ obtained via sparse functional k-mean algorithm ($k = 2$) on 1-mean aligned growth curves.

It is interesting to comment also on the resulting $b(x)$, shown in Figure 3.5. The monotonic increasing pattern of $b(x)$ for 1-mean aligned curves confirms the fact that the most relevant difference between boys and girls, the pubertal spurt, has been removed by the aligning procedure. Indeed, once the variability due to the different growth timings has been removed, the difference given by gender can still be detected, but it increasingly expresses in the whole domain. This increase is of course due to the most relevant phenomenon distinguishing the height of boys and girls, which has not been removed by the alignment procedure: the fact that the height of boys stochastically dominates the one of girls. This characteristic is common to all registered ages, and becomes more and more relevant going towards adulthood.

Part II

Classification of Spatially Dependent Functional Data

Introduction to Part II

In this second part of the dissertation we will consider unsupervised classification of functional data with spatial dependence. The problem that motivated the investigation was originally posed by our industrial partner (eni S.p.A.), in the context of irradiance data remotely sensed on a high resolution lattice.

The analysis of functional data with spatial dependence is a quite novel area of statistical research: some efforts have been already made to generalize the problem of spatial prediction (kriging) to geo-referenced functional data, among others see the interesting review by Delicado *et. al* [16]. However, most interesting for us is the case of functional data indexed by the sites of a spatial lattice. In this framework, a possible line of research consists in developing a bayesian hierarchical model for spatially correlated functional data, as in Baladandayuthapani *et al.* [4] and references therein; the drawback of these approaches is that they often involve semiparametric modeling of hierarchical functions, and are most suited for small datasets due to the computational effort associated to MCMC procedures. Concerning bayesian hierarchical modeling, spatial-temporal models have also been developed for meteorological applications (see Handcock and Wallis [27], and Beltran *et al.* [6]). However, in the direction of finding proper models for classification rather than prediction of functional spatially-dependent data, no effort has been made for generalizing classical model–based methods for image segmentation, based on Hidden Markov Random Fields models (see Geman and Geman [23], Kunsch *et al.* [42], Leung and Lang [45], and the book by Cressie [13]), to the case of functional data.

We propose a non–parametric unsupervised classification method, suited for the case of functional high–dimensional data indexed by the sites of a lattice. The method is an exploratory tool, which can nevertheless be included in the latter line of research, since its efficacy has been proven via simulation when a Hidden Markov Random Field model is assumed for data generation. Moreover, the fields of application of the proposed method are indeed more general than classification problems: we are proposing a general framework, capable of hadling functional data indexed by a lattice, which makes all functional data analysis techniques (suited to independent data) applicable to spatially dependent data, for a variety of scopes (classification, dimensional reduction, prediction, ...).

The general scheme of the proposed method, together with simulation studies carried out to test its accuracy, will be presented in Chapter 4. In this Chapter an application to environmental satellite data will also be illustrated, concerning solar irradiance measured on a non–uniform low–resolution lattice of sites covering the whole earth; scope of the analysis is clustering the sites of the lattice, in order to gain insight on the possible locations of solar power plants. In Chapter 6 the motivating application posed by eni S.p.A., concerning again a clustering problem, will be presented: the considered data are reflectance spectra, which are functions of wavelength, remotely sensed on a huge high–resolution lattice of sites, uniformly covering an area of some ten of square kilometres. Finally, in Chapter 5 the analysis of telephonic time patterns, recorded on a fine uniform grid covering the whole area of Milan, will be described: the scope here is dimensional reduction, rather than clustering, in order to describe via a proper

functional basis the typical behaviour of the mobile phone network user, and in order to catch his impact on the total mobile traffic in various areas of the city, for city planning purposes. Hence, we have to remark that, even if the three analyses described in the three chapters of this second part are unified by the general framework in which the method of analysis resides, they are deeply different both in the application at hand, and in their final scopes.

Chapter 4

Bagging Voronoi Classifiers for Clustering Spatial Functional Data

Many methods for the analysis of high–dimensional data have been proposed in recent years: some of them fall within the framework of functional data analysis (see Ferraty and Vieu [17]; Ramsay and Silverman [61]), and only few among these consider spatially dependent functional data (see the review by Delicado *et. al* [16]). With a non-parametric approach, we here consider the problem of unsupervised classification of spatially dependent functional data, where each curve is indexed by the sites of a spatial finite lattice S_0 contained in S, the region of interest for the analysis. In particular, our motivating application in this Chapter consists in analyzing a global data set concerning irradiance data along time: we examine the annual patterns of the maximum amount of energy needed to backup a photovoltaic system in 47520 worldwide non-polar districts (a non uniform lattice S_0 covering the whole earth surface S). The analysis of these patterns is closely related to the sizing of power emergency generators needed in case of consecutive no-sun days (see Richter *et. al* [63]). The problem consists in associating to each site $\mathbf{s} \in S_0$ a latent label $l \in \{1, \ldots, L\}$, such that sites homogeneous with respect to the distribution generating the functional data are labeled the same. In our motivating application, we aim at identifying different homogeneous macro-areas, interpretable in terms of the observed phenomenon and not captured by customary unsupervised classification procedures that do not take into proper account the spatial dependence among data.

A tentative approach to this problem consist in the use of standard functional clustering procedures, such as the functional k-mean or functional k-medoid algorithm described in Chapter 1, that do not properly account for spatial dependence; indeed, while assigning a site to a cluster, information carried by neighboring sites is not considered. We thus expect these standard non-spatial approaches to provide good results when the true clusters are associated to very different distributions and spatial dependence is weak. Instead, in less trivial situations where the distributions associated to different clusters may be very similar, exploiting information carried by neighbors can lead to more accurate

results.

We propose a new bagging algorithm for unsupervised classification that exploits spatial dependence by repeatedly generating random connectivity maps and by clustering, at each replicate, local representatives of neighboring functional data. Performance of our algorithm is tested in various situations, and compared with standard clustering techniques. The new algorithm is completely non-parametric, since no explicit assumption is made neither on the distribution generating the latent field of labels, nor on the conditional distribution generating functional data. A great advantage of this approach is its flexibility in the exploitation of further information on the considered region, which is not paid off by an increment of computational cost. For a given number k of clusters, the proposed spatial clustering procedure generates and analyzes bootstrap samples in three basic steps: generation of a spatial random Voronoi tessellation, identification of a representative for each of the n elements of the tessellation, p-dimensional reduction and clustering of the representatives. For each site of the lattice S_0, the final output is the frequency distribution of cluster assignment to each of the k clusters; this distribution can be summarized in a classification map by means of its mode via a majority vote on cluster assignment. The fact that our data is functional is not irrelevant to the computational cost of standard procedures for the analysis of lattice data. This further motivates our method, which implicitly performs a reduction both in the dimension of the sample (by clustering a small number n of representatives) and in the infinite dimension of functional data (through the p–dimensional reduction of the representatives). Moreover, most algorithms for image classification based on Hidden Markov Random Field models, which perform classification via a maximum a posteriori – MAP – criterion (such as *simulated annealing*, or *iterated conditional modes*, see Besag [7]; Geman and Geman [23] for details on these procedures) heavily depend on strong assumptions on the distribution of the observed signal, typically assumed to be Gaussian.

The scheme of the Chapter is the following. In Section 4.1, the Bagging Voronoi classifiers algorithm for clustering spatially dependent functional data is introduced and described. In Section 4.2, some technical issues concerning the algorithm are detailed. In particular, we describe our strategy for capturing spatial dependence and for reducing the size of the data set and our approach to dimension reduction of functional data in a clustering perspective. In Section 4.3 the properties of the algorithm are explored through a battery of simulations including different scenarios. Finally, in Section 4.4 our motivating application is fully illustrated, and results of the application of the Bagging Voronoi classifiers algorithm to irradiance data are shown. Further details on the proposed clustering strategy, and on the considered application, can be found in Secchi *et al.* [71].

4.1 Bagging Voronoi classifiers: a general framework for the analysis of spatially dependent functional data

Suppose a latent field of labels $\Lambda_0 : S_0 \to \{1, \dots, L\}$ is defined on the lattice S_0; $\Lambda_0(\mathbf{s})$ is the true unknown label associated to the site $\mathbf{s} \in S_0$, where $S_0 \subset S$ and S is a measurable subset of $\mathbb{R}^d$. The field Λ_0 sums up some characteristics of the considered area which are interesting for the scopes of the analysis. In addition, suppose that a functional datum is observed in each site $\mathbf{s} \in S_0$: given Λ_0, the functional data are independently generated in each site $\mathbf{s} \in S_0$ from a distribution indexed by $\Lambda_0(\mathbf{s})$. The main object of the present paper is a classification procedure aiming at the reconstruction of the unknown field Λ_0 of labels, based on the clustering of the functional data indexed by the sites of S_0. Hence, the final result of the procedure is a label assignment for each site of the lattice, in accordance to the observed functional data.

The procedure is a bagging-inspired clustering algorithm. Indeed, it is composed by a *bootstrap* sampling phase, articulated in three basic steps, and by an *aggregation* phase (see Breiman [10] for details on bagging procedures): at each replicate of the three steps, a single weak classifier is found, which exploits a specific structure of spatial dependence, thus obtaining a coarse estimate of the unknown latent field of true labels Λ_0. A more accurate global classifier is obtained after B replicates, by *bagging* together all single classifiers. Higher values of B, imply a higher accuracy of the final estimate (the reconstruction of the latent field of labels Λ_0), which includes all the estimates of the B weak single classifiers. Moreover, the advantage of such an approach stands in the embarrassingly parallel nature of the bootstrap, whose computational cost can be reduced by parallel programming.

The procedure is implemented through an algorithm that we sketch via a pseudo-code scheme showed in Figure 4.1. In `step 1` of the bootstrap sampling part of the algorithm, the random Voronoi tessellation of the lattice isolates neighboring groups of data to capture potential spatial dependence. In `step 2` a local representative for each element of the tessellation is identified to sum up local information: indeed, neighboring data are most likely drawn from the same functional distribution. Functional dimensional reduction of local representatives is performed in `step 3` to select relevant functional features in the data. The projections of local representatives on the space spanned by the obtained basis are then clustered in k groups according to a suitable unsupervised method, depending on the application (e.g. k-mean clustering, Partitioning Around Medoids – PAM, ...). A final classification map of the lattice S_0 is obtained in the aggregating phase where results of each replicate are bagged together: `cluster matching` is needed to ensure coherence of cluster assignments across replicates. Then, the frequency distribution of assignment of each site to each of the k clusters along the bootstrap replicates is considered. Indeed, for each site in S_0, a final assignment to *one* of the k clusters can be obtained by selecting the label corresponding to a mode of this distribution. Specifications of each phase of the algorithm will be detailed in Section 4.2.

Algorithm. Bagging Voronoi classifiers.

Bootstrap:

Initialize B, n, p, K. Choose a metric $d(\cdot, \cdot)$.

`for` $b := 1$ `to` B `do`

`step 1.` randomly generate a set of nuclei $\Phi_n^b = \{\mathbf{Z}_1^b, \ldots, \mathbf{Z}_n^b\}$ among the sites in S_0: for $i = 1, \ldots, n$, $\mathbf{Z}_i^b \overset{i.i.d.}{\sim} \mathcal{U}(S_0)$, where $\mathcal{U}$ is the uniform distribution on the lattice. Obtain a random Voronoi tessellation of S_0, $\{V(\mathbf{Z}_i^b | \Phi_n^b)\}_{i=1}^n$, by assigning each site $\mathbf{x} \in S_0$ to the nearest nucleus $\mathbf{Z}_i^b$, according to the specified distance $d(\cdot, \cdot)$;

`step 2.` for $i = 1, \ldots, n$, compute the function g_i^b, acting as local representative, by summarizing information carried by the functional data associated to sites belonging to the i-th element of the tessellation $V(\mathbf{Z}_i^b | \Phi_n^b)$;

`step 3.` perform dimensional reduction of the local representatives $\{g_1^b, \ldots, g_n^b\}$ by projecting them on the space spanned by a proper p–dimensional functional orthonormal basis, thus generating the p-dimensional scores vectors $\{\mathbf{g}_1^b, \ldots, \mathbf{g}_n^b\}$, which are then clustered in K groups according to a suitable unsupervised method.

`end for`

Aggregation:

perform `cluster matching`: for $k = 1, ..., K$, and $b = 1, ..., B$, indicate with C_k^b the set of $\mathbf{x} \in S_0$ whose label is equal to k, and match the cluster labels across bootstrap replicates, to ensure identifiability.

`for` $\mathbf{x} \in S_0$ `do`

- calculate the frequencies of assignment of the site to each of the K clusters along iterations, i.e., $\pi_{\mathbf{x}}^k = \#\{b \in \{1, ..., B\} : \mathbf{x} \in C_k^b\}/B$, $\forall\, k = 1, \ldots, K$;
- compute spatial entropy $\eta_{\mathbf{x}}^K$ for each site $\mathbf{x} \in S_0$.

`end for`

Figure 4.1: Pseudo-code scheme of the Bagging Voronoi classifiers algorithm.

Note that the procedure depends on a number of choices which initialize the algorithm, e.g. the parameters B, n, p and k have to be chosen in advance. While B should be chosen large enough to ensure the desired algorithm accuracy, p depends on the particular problem at hand, and its choice will be discussed in Section 4.2.

We shall now discuss the choice of the dimension n of the Voronoi tessellation, and of the correct number k of clusters. As a tentative approach to both issues, we propose an innovative *spatial entropy* criterion for the evaluation of the quality of the final classification. Consider the frequency distribution of assignment $\boldsymbol{\pi}_{\mathbf{s}} = (\pi_{\mathbf{s}}^1, \ldots, \pi_{\mathbf{s}}^k)$ of each site $\mathbf{s} \in S_0$ to each of the k clusters. The entropy associated to the final classification in the site $\mathbf{s} \in S_0$ is obtained as

$$\eta_{\mathbf{s}}^k = -\sum_{j=1}^{k} \pi_{\mathbf{s}}^j \cdot \log(\pi_{\mathbf{s}}^j), \tag{4.1}$$

which assumes minimum value 0 when there is an r such that $\pi_{\mathbf{s}}^r = 1$ while $\pi_{\mathbf{s}}^j = 0$ for all $j \neq r$, with $j, r \in \{1, \ldots, k\}$, and maximum value $\log(k)$ when $\pi_{\mathbf{s}}^j = \frac{1}{k}$ for $j = 1, \ldots, k$. The use of index (4.1) as a criterion for evaluating the final classification is based on the idea that the more the frequency distribution $\boldsymbol{\pi}_{\mathbf{s}}$ is concentrated on one particular label, the more the classification is precise and stable along replicates. Conversely, when frequencies are more uniformly spread over all labels, uncertainty associated to the final classification in $\mathbf{s}$ is higher. Spatial entropy can be visualized by plotting its value in each site of the lattice. A global evaluation index can also be computed as the *average normalized entropy*

$$\eta^k = \frac{\sum_{\mathbf{s} \in S_0} \eta_{\mathbf{s}}^k}{\log(k) \cdot |S_0|}, \tag{4.2}$$

including the contribution to the final classification quality of all sites in S_0. For comparisons over different choices of k, the quantity $\eta_{\mathbf{s}}^k$ in equation (4.2) has been normalized by its maximum value.

Since the index expressed in (4.1) is a measure of the uncertainty associated to the final classification, we expect the value of η^k to be low if n is properly chosen in accordance to the (unknown) spatial dependence in the latent field of labels: for an optimal choice of n, few elements of the Voronoi tessellation will cross the borders among regions associated to different labels in the latent field, and high values of the entropy will be very localized along these borders.

One might guess that spatial entropy is a good criterion also for the selection of the most proper value for k: indeed, when k is optimally chosen, we expect a neat data classification leading to a map of the spatial entropy mostly equal to zero, with values significantly different from zero only in sites at the boundaries between regions associated to different labels in the latent field. In fact, in Section 4.3 we will elaborate on the matter by means of simulation studies, and we will conclude that the entropy criterion generally leads to a choice for k more parsimonious than necessary. Indeed, the problem of the choice of an optimal k is a well–known issue in cluster analysis: many strategies have been proposed to solve it, and none of them so far proved to be conclusive. In Section 4.3 we will discuss a possible approach, based on the evaluation of the following index associated to the final classification:

$$\theta = \frac{tr(B)}{tr(B+W)}, \tag{4.3}$$

where B and W are the final *between* and *within* cluster sum of squares matrix, respectively.

4.2 Details on the algorithm

We will now expand on the details of each step of the Bagging Voronoi classifiers algorithm described through the pseudo-code scheme in Figure 4.1. Note that the extreme generality of the proposed algorithm makes each step flexible to different specifications eventually motivated by the application at hand.

4.2.1 Step 1: Voronoi tessellations

We will here give motivations for the treatment of spatial dependence in lattice data via Voronoi tessellations. This comes from a consistency result proved by Penrose [56] in the framework of stochastic geometry.

Consider $S \subset \mathbb{R}^d$. To univocally define a random tessellation it is necessary to select a set of points $\mathbf{s} \in S$ as nuclei for the Voronoi tessellation. Thus, let $\Phi_n = \{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n\}$ be a set of n points in S sampled from a proper distribution F defined on S: this will be the set of nuclei of the Voronoi tessellation. For each $\mathbf{Z}_i \in \Phi_n$, define the polyhedron

$$V(\mathbf{Z}_i|\Phi_n) = \{\mathbf{s} \in S : d(\mathbf{s}, \mathbf{Z}_i) \leq d(\mathbf{s}, \mathbf{Z}_j), \text{ for all } \mathbf{Z}_j \in \Phi_n,\ i \neq j\}, \tag{4.4}$$

to be the closed Voronoi cell with nucleus $\mathbf{Z}_i$ for the Voronoi tessellation induced by Φ_n, i.e. the set of $\mathbf{s} \in S$ lying at least as close to $\mathbf{Z}_i$, in the sense of the metric $d(\cdot, \cdot)$, as to any other point of Φ_n: the collection $\{V(\mathbf{Z}_i|\Phi_n)\}_{i=1}^n$ forms a Voronoi tessellation of S. There is no strong restriction on the choice of the metric $d(\cdot, \cdot)$, but the final tessellation will clearly depend both on the choice of the metric, and on the distribution F. Voronoi tessellations have many interesting properties, which make them good tools for partitioning a general domain (see Mller [53] for further details on Voronoi tessellations in Euclidean spaces). However, the most interesting property for our purposes is undoubtedly a *coverage* property. Consider the collection of Lebesgue measurable sets $\{A_l\}_{l=1}^L$, $A_l \subset S$ for $l = 1, \ldots, L$, and let $V_i = V(\mathbf{Z}_i|\Phi_n)$; moreover, let

$$A_l^n = \bigcup_{\mathbf{Z}_i \in A_l} V_i,$$

be an approximation of A_l given by the Voronoi cells whose nuclei belong to A_l. The coverage property states that, for $l = 1, \ldots, L$, the set A_l^n is a consistent estimator of the unknown set A_l in the sense that

$$\mu(A_l^n \Delta A_l) \xrightarrow{a.s.} 0, \quad n \to \infty, \tag{4.5}$$

where Δ denotes the symmetric difference between two sets and μ is the Lebesgue measure. The coverage property for Voronoi tessellations expressed in (4.5) has been proved by Penrose in [56] under the reasonable assumption that the support of F includes S, and it represents a strong law of large numbers in the context of Voronoi tessellations.

The coverage property (4.5) is essential for the validity of our algorithm, since it states that, when the tessellation becomes less and less coarse, subsets of the domain S associated to the same label are well approximated by suitable sets

of Voronoi elements. Indeed, with a view to our classification problem, let the collection of sets $\{A_l\}_{l=1}^{L}$ be defined by setting, for $l = 1, \ldots, L$,

$$A_l = \{\mathbf{s} \in S : \Lambda(\mathbf{s}) = l\},$$

where $\Lambda : S \to \{1, \ldots, L\}$ is a function whose restriction to S_0 is Λ_0.

4.2.2 Step 2: functional local representatives

Consider the situation where T is a bounded interval of $\mathbb{R}$ and a realization $\chi_{\mathbf{s}} : T \to \mathbb{R}$ of a functional random variable is observed in each site $\mathbf{s}$ of the lattice S_0. Following a completely non-parametric approach, we do not make assumptions on the distribution generating $\chi_{\mathbf{s}}$. The Voronoi tessellation provides a partition of S_0 in random neighborhoods and induces a partition in the sample of functional data $\{\chi_{\mathbf{s}}\}_{\mathbf{s}\in S_0}$. For each element V_i of the Voronoi tessellation, we sum up the information contained in the sub-sample $\{\chi_{\mathbf{s}}\}_{\mathbf{s}\in V_i}$ by estimating a *functional local representative* through a method that exploits spatial dependence of neighboring data. In the literature on spatial statistics, this is called *spatial smoothing* (see Banerjee *et al.* [5] for further details). We now describe an approach to the estimation of functional local representatives, adopted in all data analyses considered in the following sections. It must be stressed that this approach is by no means the only possibility; all schemes apt to estimate a local centroid (e.g. a functional median or mean, a locally weighted polynomial regression curve, ...) could be used in place of it.

For $i = 1, \ldots, n$, the functional local representative g_i of the sub-sample $\{\chi_{\mathbf{s}}\}_{\mathbf{s}\in V_i}$ is computed as a weighted mean with a Gaussian kernel:

$$g_i(t) = \frac{\sum_{\mathbf{s}\in V_i} w_{\mathbf{s}}^i \chi_{\mathbf{s}}(t)}{\sum_{\mathbf{s}\in V_i} w_{\mathbf{s}}^i}, \tag{4.6}$$

where $w_{\mathbf{s}}^i$ is a Gaussian weight centered in $\mathbf{Z}_i$ and decreasing with respect to $d(\mathbf{s}, \mathbf{Z}_i)$. Intuitively we assume that spatial dependence between two sites decreases when the distance between them increases. In this sense the local representative already accounts for spatial dependence: a bigger contribution to its computation, in fact, will be given by functional data associated to sites nearer to the nucleus of the element. The kernel covariance matrix is $\sigma^2 \mathbb{I}_2$, where $\sigma = d_{max}/d_{min}$, being d_{max} and d_{min} the maximum and minimum distance between two nuclei of the tessellation, respectively. This choice connects σ to the mean dimension of the tessellation element via an estimator of the elements mean diameter (see Mller [53] for a proof in Euclidean spaces). The setting is general and can be easily adapted to different situations arising in applications, when in presence of prior information: for example, a non–diagonal covariance matrix in the Gaussian kernel could be used to account for anisotropy in the latent random field.

The choice of n, which sets the tessellation dimension and thus the number of local representatives to be computed, has great influence on the algorithm behavior, since the latent field of labels is unknown. If the labels were known, we would choose a tessellation perfectly matching the cluster borders. Thus,

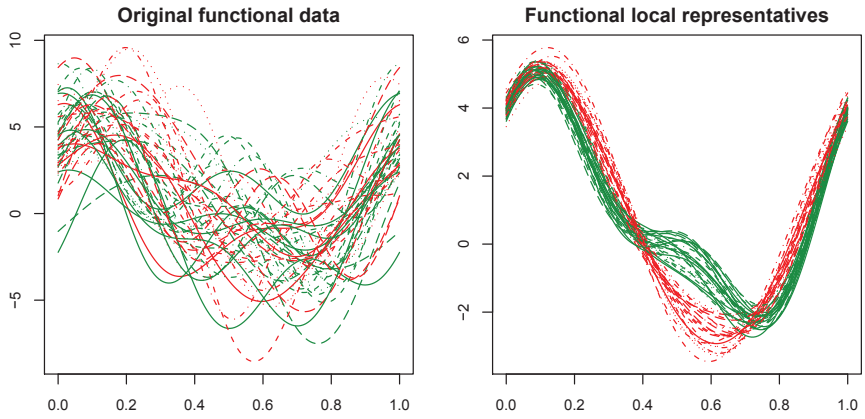

Figure 4.2: In the left panel, a sample of 50 synthetic data randomly selected from the realization of a Hidden Markov Random Field with Gaussian emission probability function; in the right panel, the sample of functional local representatives obtained using a Voronoi tessellation with $n = 50$. Different colors correspond to different labels.

independently from the choice of n, each local representative would be computed using data drawn from the same distribution. Indeed, labels are always unknown, inducing a bias-variance trade-off which determines the existence of an optimal choice of n. Consider the example in Figure 4.2. In the left panel a sample of 50 curves randomly selected from a synthetic data set is shown: they are generated according to a Hidden Markov Random Field with Gaussian emission probability function. The latent field of labels is the realization of a Ising field on a 100×100 lattice of sites and parameter $\beta = 3$. Functional data are obtained using a Fourier basis with fixed ($p = 5$) dimension and coefficients taken from the realization of the emitted random field. The mean vectors of the emission probability function are $\boldsymbol{\mu}_{-1} = (1, 2, 2.25, 0, 0)$ and $\boldsymbol{\mu}_1 = (1, 1.5, 1.25, 0.75, 0.75)$, respectively, and covariance matrices are the identity in $\mathbb{R}^5$ in both cases. Thus the functional distribution generating data is a mixture, with components associated to the two labels of the latent field. In the right panel the functional local representatives corresponding to the same data set are depicted: the tessellation dimension is chosen equal to $n = 50$ and local representatives are computed using equation (4.6). Note that in the left panel we can hardly distinguish a grouping structure, since the variability within the two groups is confounding the one between the groups. In the right panel of the picture, instead, we can distinguish curves belonging to two different groups thanks to the evident reduction of the variance in the sample of local representatives. Moreover, the portion of the domain in which the mean curves of the two clusters are the most different is also evident in the right panel; this is another effect of the good balance between bias and variance in the computation of local representatives.

In general, a certain number of representatives will be optimal in terms of misclassification error. The optimal choice of n is the one that finds a good compromise between variance and bias:

- as n decreases, noise is reduced in the local representatives sample, since local representatives are weighted sample means calculated on sub-samples that are larger on average (minimal variance). However, at the same time the associated Voronoi tessellation follows less accurately the boundaries in the true latent field of labels, thus including different mixture components in the computation of local representatives (maximal bias). The limiting case is $n \equiv 1$, when all sites in the finite lattice belong to the same Voronoi element, and are thus used to compute a single representative.

- As n increases, the resulting Voronoi tessellation approximates more accurately the boundaries of the latent field of labels (minimal bias), but at the same time the variability reduction due to spatial smoothing is smaller (maximal variance). The limiting case is $n \equiv |S_0|$, when all sites in the finite lattice are nuclei, and thus no spatial smoothing is performed.

The optimal value of n determined by this trade-off depends both on the strength of spatial dependence, and on the mixture components of the distribution generating the functional data. In Section 4.3 a simulation study aimed at evidencing the existence of an optimal value for n in some realistic situations is detailed.

4.2.3 Step 3: dimensional reduction and clustering

The third step of the algorithm aims at performing data dimensional reduction and at clustering the dimensionally reduced data, to obtain a classification map. We thus introduce functional data analysis techniques aimed at catching the most relevant features in the functional distribution of the sample of local representatives $\{g_1, \ldots, g_n\}$ (see Ramsay and Silverman [61] for a thorough exposition of functional data analysis). We remark that the third and fourth steps of the proposed procedure are the most open to different specifications, according to the aim of the application at hand: we will give in Chapter 5 the details of an application in which the Bagging Voronoi strategy is used to select a reference basis for the sample of spatially dependent functional data: thanks to bagging, the estimated basis is able to capture the different temporal patterns shown by the sample of data, while exploiting spatial dependence.

For dimensional reduction of functional data, we need to find the best projection of data onto the space generated by a proper functional basis of finite dimension p. The choice of this basis is extremely open, and heavily depends on the application. In general, we shall distinguish two possible situations. Either the functional basis used for p-dimensional reduction of the sample of local representatives is fixed along replicates (e.g., a wavelet basis, or a Fourier basis) or the functional basis is data–driven, adding greater flexibility with respect to the functional features possibly arising in applications.

Among the latter approaches to p–dimensional reduction, one of the possibilities consists in functional principal component analysis (FPCA), i.e., Karhunen-Love decomposition. This is the choice adopted in all data analyses illustrated in the following sections. Indeed, the basis composed by functional principal components is a complete orthonormal basis of $L^2(T)$ (see Ferraty and Vieu [17] for theoretical details). Note that we must assume that the representatives

$g_1, \ldots, g_n$ are independently generated by a distribution with finite second order moments, in order to consistently estimate the covariance structure needed to find principal components. However, we expect that spatial dependence among local representatives has been highly reduced so that this assumption does not seem unreasonable.

For an introduction to Karhunen–Love decomposition of functional data we refer to subsection 1.1.2. When the scope of the decomposition stands in performing dimensional reduction of the sample of local representatives $\{g_1, \ldots, g_n\}$, we can select only the first p eigenfunctions to represent data; in particular, keeping in mind our classification problem, we can select those basis functions which, according to a graphical inspection, are explaining features mostly associated to a grouping structure. Alternatively, p could be determined by fixing a given portion of the variability to be explained by selected components.

Having chosen the p principal components, only projections of data along them are analyzed, thus obtaining dimensional reduction. To meet the final task of the classification procedure, we then perform k-mean clustering on the sample of the p-vectors of the scores $\{\mathbf{g}_1, \ldots, \mathbf{g}_n\}$. Note that this is exactly equivalent to performing functional k-mean clustering with a semi–metric induced by the first p eigenfunctions of the estimated covariance operator. We also remark that other clustering methods can be used to obtain a final classification map, e.g., hierarchical methods, PAM or sparse clustering, depending on the scopes of the analysis and on the application.

4.2.4 Aggregation: cluster matching

The output of the b-th replicate of the bootstrap phase of the algorithm is a label assignment for each local representative estimated during the b-th run of `step 2`. Let $\hat{l}_1^b, \ldots, \hat{l}_n^b$ denote the labels of the local representatives at the b-th replicate, i.e. for $i = 1, \ldots, n$, the label $\hat{l}_i^b \in \{1, 2, \ldots, k\}$ is the cluster assignment of the representative g_i^b to one of the k clusters. All sites $\mathbf{s}$ in V_i^b get the same label $\hat{l}_i^b$. For $j = 1, ..., k$, indicate with C_j^b the set of sites $\mathbf{s} \in S_0$ whose label is equal to j at the bth replicate. To obtain a classification map of the lattice S_0, we consider the frequency distribution of assignment of each site to each of the k clusters along the B replicates. To compute this frequency distribution we need in turn to assume that cluster labels $\{C_1^b, \ldots, C_k^b\}$ are coherent along the B replicates. More specifically, we want cluster labels $\{C_1^b, \ldots, C_k^b\}$ to be coherent with $\{C_1^m, \ldots, C_k^m\}$, for all $m < b$, and $b \geq 2$; this is obtained through *cluster matching*, which we perform only on subsequent replicates. The algorithm looks for the label permutation that minimizes the total sum of the off-diagonal frequencies in the contingency table describing the joint distribution of sites along the classifications at two subsequent replicates. Other different procedures for cluster matching are conceivable.

4.3 Simulation studies on synthetic data

We now describe two simulation studies aimed at testing the Bagging Voronoi classifiers algorithm on synthetic data. In the first simulation we focus on spatial

dependence and we explore its influence on the algorithm performance and on the choice of the parameter n setting the dimension of the Voronoi tessellation. In the second simulation we tackle the problem of selecting the right number k of clusters. Since the focus is on spatial dependence and only secondly on the functional nature of data, in all simulation studies the emitted random field belongs to a finite dimensional vector space.

4.3.1 Simulation 1: the role of spatial dependence in the optimal choice of n

The first simulation study aims at testing the algorithm performance with respect to the choice of the number n, setting the dimension of the Voronoi tessellation, under different conditions characterizing the spatial dependence of the latent field of labels.

Set S_0 to be a two-dimensional square lattice of 50×50 sites and generate the latent field of labels as a Potts Markov Random field $\Lambda_0 : S_0 \to \{1, \dots, L\}$, where L is the number of labels in the latent field. The Potts model extends the Ising model to the case $L > 2$; both models have been extensively studied in statistical physics to describe the behavior of magnetic materials (see Wu [86]). The probability of a specific configuration for the sites in S_0 depends on its energy:

$$\mathbb{P}\left(\Lambda_0(S_0) = \boldsymbol{\lambda}\right) = \frac{1}{Z} \exp\left\{\beta \sum_{\mathbf{s} \in S_0} \sum_{\mathbf{x}' \in \mathcal{N}_\mathbf{s}} \delta(\lambda(\mathbf{s}), \lambda(\mathbf{x}'))\right\},$$

where Z is a normalizing constant, $\boldsymbol{\lambda} = \{\lambda(\mathbf{s}) \in \{1, \dots, L\} : \mathbf{s} \in S_0\}$ collects the realizations of the field on each site of the lattice S_0, and $\mathcal{N}_\mathbf{s}$ is a first-order neighborhood of $\mathbf{s} \in S_0$. The function δ is such that $\delta(\updownarrow_1, \updownarrow_2) = 1$ if $\updownarrow_1 = \updownarrow_2$ and 0 otherwise. The strength of spatial dependence is controlled by the parameter β, a physical constant characterizing the influence of neighboring sites on the realization of the field in a particular site: higher values of β imply a stronger spatial dependence (see Kunsch [42] for more details on this model). In the following simulations, we set $\beta \in \{\frac{1}{4}, \frac{1}{2}, 1, 2, 4\}$ and $L \in \{2, 3, 4\}$.

Conditionally on the realization of the latent field, in each site $\mathbf{s}$ of S_0 we generate independently a random multivariate vector of dimension $p = 5$ from a multivariate Gaussian distribution; the distribution of the random vector depends exclusively on the site label

$$\mathbf{Y}_\mathbf{s} | (\Lambda_0(\mathbf{s}) = l) \sim N_p(\boldsymbol{\mu}_l, 2\mathbb{I}_p).$$

Hence the conditional distribution of the emitted signal given the label differs only in the mean for different values of the label. In particular, for $L = 2$ we choose $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2 = (-2, -1, 0, 1, 2)$ as mean values for the emitted field. When L increases to 3 we add $\boldsymbol{\mu}_3 = (-1.76, 1.81, 1.11, 0.796, -0.136)$ (each component of $\boldsymbol{\mu}_3$ has been independently generated from a uniform distribution on the interval [-2,2]). Finally, when $L = 4$ we add $\boldsymbol{\mu}_4 = (0, 0, 0, 0, 2)$.

The parameters controlling the algorithm are fixed as follows: $B = 50$, $k = L$ and $n \in \{1, 5, 10, 25, 50, 125, 250, 500, 1000, 2500\}$. The Voronoi tessellations are

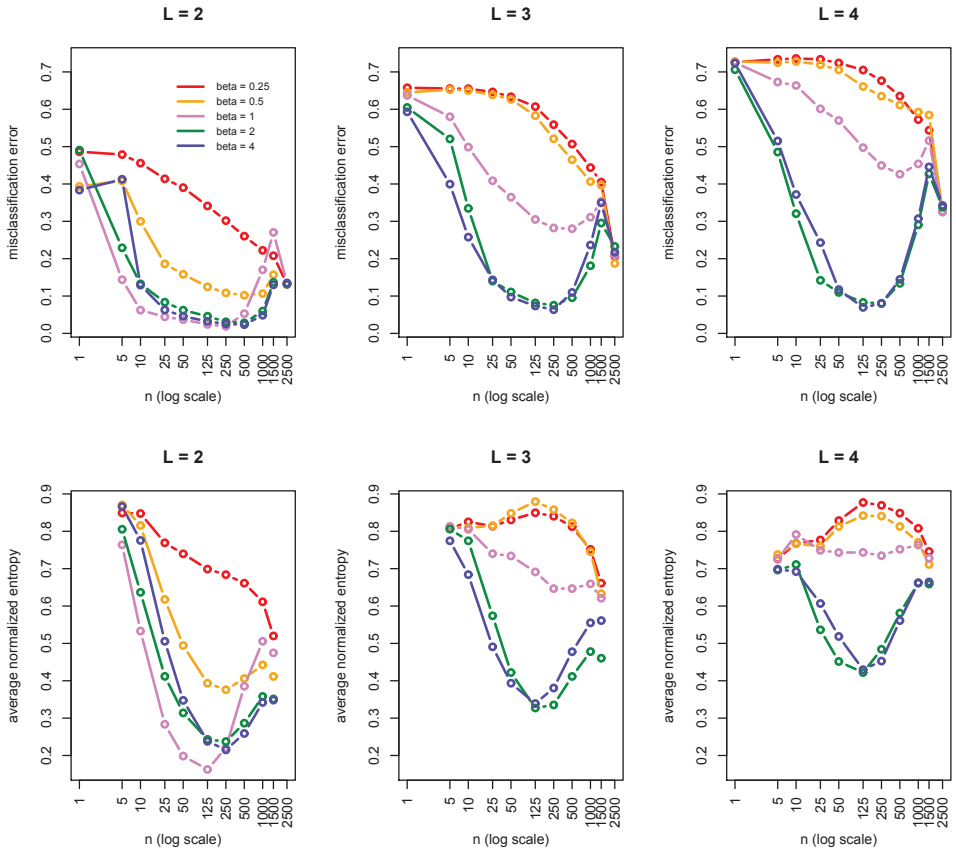

Figure 4.3: Results of the first simulation study: in the top panels, misclassification error obtained via Bagging Voronoi classifiers algorithm over different choices for n and β – average over 30 replicates of the procedure; each panel corresponds to a different number L of labels in the latent field. In the bottom panels, the results of Bagging Voronoi classifiers algorithm in terms of average normalized entropy.

obtained by using the Euclidean distance and by drawing the set of nuclei uniformly from S_0. The n representatives are identified as weighted means with Gaussian isotropic weights; no dimensional reduction is performed. Finally, clustering of the representatives is obtained through k-means. The final classification map is generated via a majority vote on cluster assignment. Results are evaluated in two ways: by means of the misclassification error rate with respect to the true realization of the field, and by means of the average normalized entropy associated to the final classification. Note that only the latter is computable in real applications, where the true latent field of labels is unknown. Hence comparing the two indexes within the context of a simulation study is relevant for deciding whether the average normalized entropy could be interpreted as a proxy of the misclassification error. The final evaluation of the algorithm is obtained by repeating each simulation 30 times, and by calculating the mean values of

misclassification error and average spatial entropy.

Results are illustrated in Figure 4.3. Consider the top panels of the picture, showing the mean misclassification error for different values of β and n, and for the number of different labels in the latent field equal to $L = 2, 3, 4$ (left, center, and right panel respectively). Note that the limiting case $n = 50 \times 50 = 2500$ corresponds to the application of (non–spatial) standard k-mean clustering. First, consider the mean misclassification error with respect to n: we appreciate that, for all L and for large values of β, the value of n minimizing the misclassification error is strictly in between the limiting cases $n = 1$ (no clustering) and $n = 2500$ (non–spatial clustering). Indeed, we notice that misclassification error is uniformly smaller (with respect to n) for larger values of β: hence the improvement introduced by the Bagging Voronoi classifiers algorithm is stronger in the presence of a stronger spatial dependence in the latent field of labels, for any n. Moreover, for larger values of β, the procedure increasingly exploits the information carried by neighboring data and the optimal value of n decreases (implying on average bigger Voronoi elements). Conversely, when β decreases spatial dependence becomes weaker, and the procedure looks for less coarse tessellations, eventually leading to an optimal $n = 2500$. When there is no spatial dependence, there is no need to use the Bagging Voronoi classifiers algorithm.

Now, look at the bottom panels in Figure 4.3: in the same conditions as above, we here report the mean of the average normalized entropy. Recall that, contrary to the misclassification error, no information on the true latent field of labels is used when computing this second index. It is interesting to note that these pictures also point at a minimum value for n, that is close to the optimal n with respect to the misclassification error, provided that spatial dependence is sufficiently high. Hence, in real applications, the average normalized entropy can guide the selection of the tessellation dimension. Indeed, one might suspect a very weak spatial dependence in the latent field when the average normalized entropy function assumes consistently high values.

4.3.2 Simulation 2: on the choice of the number k of clusters

In most real applications, the number of labels in the underlying latent field is unknown, and one has to initialize the Bagging Voronoi classifiers algorithm specifying the number k of clusters the algorithm is going to detect. In this second simulation study, we fix the spatial dependence in the latent field by setting $\beta = 2$ (a reasonable value for the Bagging Voronoi classifiers algorithm to be effective) and we let $n \in \{1, 5, 10, 25, 50, 125, 250, 500, 1000, 2500\}$ and $k \in \{2, \dots, 7\}$. All other parameters controlling data generation and the algorithm specifications are set the same as in Simulation 1.

Figure 4.4 illustrates the results of this second simulation study. In the top panels of the picture, the values of the mean misclassification error are shown, for different choices of k, n, and L. The bottom panels represent the values of the mean average normalized entropy. Each curve refers to a different value of k. The curve relative to the correct choice $k = L$ is always drawn in black. When $k \neq L$ we need to specify a label matching criterion before computing the misclassification error. This is done by choosing the permutation of the cluster

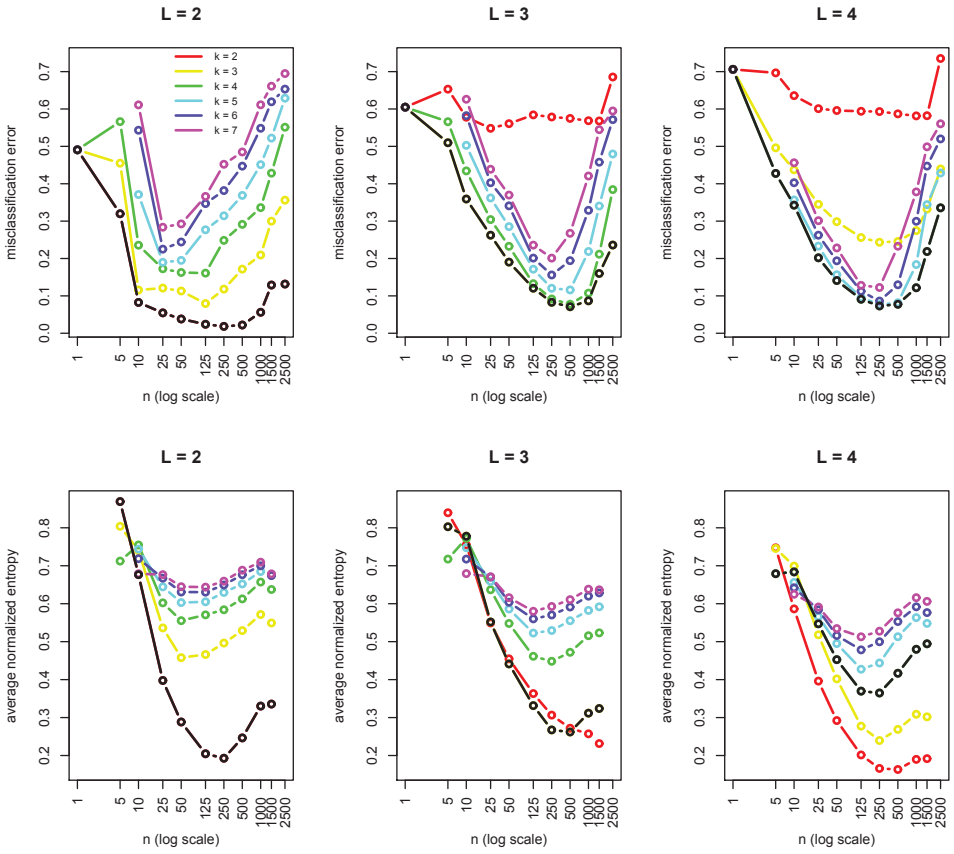

Figure 4.4: Results of the second simulation study: in the top panels, misclassification error obtained via Bagging Voronoi classifiers algorithm over different choices for n and k – average over 30 replicates of the procedure; each panel corresponds to a different number L of labels in the latent field. The black line in each panel corresponds to the correct choice $k = L$ for the number of clusters. The bottom panels represent the analogous curves for the mean average normalized entropy.

labels which minimizes the extra–diagonal elements in the contingency table of true and assigned label for each pixel.

We first inspect the black curves, corresponding to $k = L$. We notice that in the top panels they are uniformly lower than the curves obtained for values of $k \neq L$. Moreover, the optimal n in terms of mean misclassification error (top) corresponds to the n minimizing the average normalized entropy (bottom). However, in real applications L is unknown; for choosing n we thus need to consider the curves of the average normalized entropy obtained for different values of k. Indeed, the bottom panels in Figure 4.4 show that for almost all values of k, the n minimizing the average normalized entropy is in a close neighborhood of its optimal value as observed on the black curves. We also notice that this optimal value does not change much with L, thus supporting the conjecture that

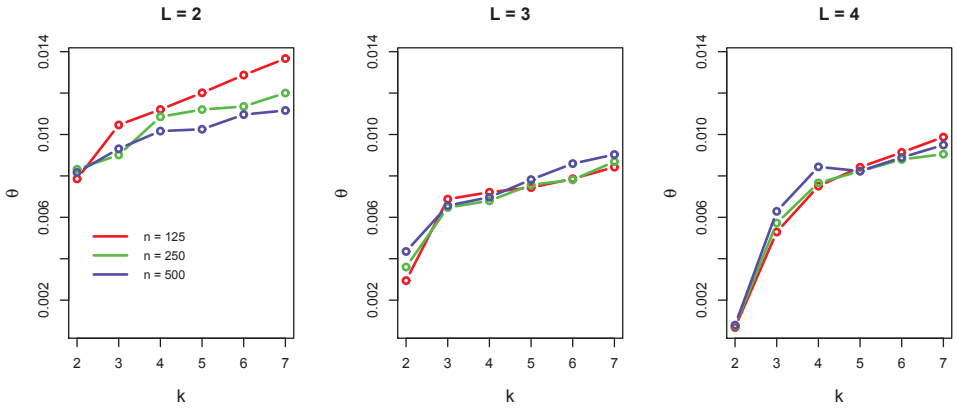

Figure 4.5: Results of the second simulation study: values of θ obtained via Bagging Voronoi classifiers algorithm over different choices for k, and when $n = 125, 250, 500$.

it depends uniquely on the spatial correlation structure of the latent field (which is constant in all the scenarios considered in Simulation 2).

A subtler observation is also worth mentioning. For $k \geq L$, the average normalized entropy curves increase with k uniformly for reasonable values of n. For values of k less than L, one might observe lower values of the average normalized entropy. Hence, if the choice of k is based on the inspection of the average normalized entropy curves, one might be induced to select a value for k that is smaller than the correct one, eliciting a parsimonious description of the latent field of labels[1].

We thus deepen the analysis for the choice of k by computing the index θ introduced in (4.3). Results are shown in Figure 4.5 for different values of k and $n = 125, 250, 500$. We appreciate an elbow corresponding to the correct choice of $k = L$ in the three scenarios $L = 2, 3, 4$. This is quite independent from the value chosen for n, provided this value belongs to a neighborhood of the optimal n as observed on the black curves in Figure 4.4. This supports the robustness of our algorithm with respect to the choice of n and the idea that classical methods for choosing k can suit also to this framework.

4.4 Case study: clustering irradiance data

We now illustrate an application of our classification algorithm to irradiance data, carried out to investigate the possible exploitation of solar energy in different areas of the planet. In particular, power production via collectors that are able to track the sun diurnal course is strongly influenced by solar irradiance and atmospheric conditions. In fact, solar thermal power employs only direct sunlight and it is therefore best positioned in areas, such as deserts, steppe or savannas, where large amounts of humidity, fumes or dust, that may deviate the sunbeams, do not occur (see Richter [63]).

[1] *Frustra fit per plura quod fieri potest per pauciora* (William of Ockham)

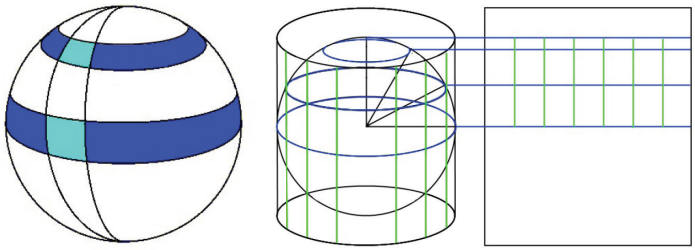

Figure 4.6: A scheme of the non–uniform lattice S_0 considered in the irradiance data application (left), and a synthetic scheme of the cylindrical projection with equal areas used to plot the classification maps for this application.

Insolation is a measure of solar radiation energy received on a given surface area in a given time. It is commonly expressed as average irradiance in kilowatt–hours per square meter per day (kWh/(m^{2}day)). We consider *direct insolation*, i.e., the solar irradiance measured at a given location on earth with a surface element perpendicular to the sunbeams, excluding diffuse insolation, i.e., the solar radiation that is scattered or reflected by atmospheric components in the sky. Direct insolation is equal to the solar constant minus the atmospheric losses due to absorption and scattering: while the solar constant varies with the earth–sun distance and solar cycles, the losses depend on the time of day (length of light path through the atmosphere depending on the solar elevation angle), cloud cover, moisture content, and other impurities.

We try to identify areas of the planet which are optimal with respect to the positioning of solar power collectors by considering parameters, which depend on direct insolation, suited for sizing batteries or other energy-storage systems. Typical solar insolation parameters are the *minimum available insolation over a consecutive–day period (%)*, the *solar radiation deficits below expected values incident on a horizontal surface over a consecutive–day period (kWh/m^2)* and the *equivalent number of NO-SUN or BLACK days that must be supplied by the storage backup system (days)*. They are fully described in the NASA website (NASA 2010). Knowledge on these parameters is desirable since unusually cloudy conditions occurring over a number of consecutive days continually draw reserve power from batteries or some other storage device for solar systems not connected to an electrical power grid. Storage devices must be designed to withstand continuous below–average conditions in various regions of the globe. More precisely, we analyze the maximum deficit below average value of solar radiation incident on a horizontal surface over a consecutive–day period (kWh/m^2), which is strictly related to the equivalent number of NO–SUN or BLACK days, and which is also increasing in the monthly average irradiance (see the NASA website [55] for details on available data sets). From an engineering point of view, this quantity is considered as a proxy of the buffer extra-capacity that is needed to be installed in order to fulfill the possible gaps in energy supply provided by solar power plants. These gaps, in a particular site at a particular time of the year, can be due to unfavorable environmental conditions. From now on, we will name this quantity *buffer capacity*.

Rough data consist of vectors in $\mathbb{R}^{12}$ indexed by the sites of a spatial lattice. In each site, the 12 measures correspond to the values of the monthly maximum energy deficit with respect to the monthly average. Both the maximum and the average values are computed over the 22 years time period from July 1983 to June 2005. Sites are located on a non–uniform lattice $S_0 = \bigcup_{\lambda \in Z_1; \theta \in Z_2} A_{\lambda\theta}$, where $Z_1 = \{-180, -179, ..., 178, 179\}$ and $Z_2 = \{-66, -65, ..., 65\}$: each element $A_{\lambda\theta}$ is the portion of the earth surface which is included between the meridians at longitude λ and $\lambda + 1$ in degrees, and between the parallels at latitude θ and $\theta + 1$ in degrees. This lattice is of course non–uniform, and includes 47520 worldwide non–polar districts (a naive scheme of the lattice is shown in Figure 4.6). In each site of the lattice, we observe the buffer capacity $Y^{\nu}_{\lambda,\theta}$ for each month $\nu \in \{July, \ldots, June\}$. More precisely, the observed quantity $Y^{\nu}_{\lambda,\theta}$, for $\lambda \in Z_1, \theta \in Z_2$ and $\nu \in \{July, \ldots, June\}$, is obtained, once fixed a month, using observations along 22 years from July 1983 to June 2005, as

$$Y^{\nu}_{\lambda,\theta} = \max_{\bar{n} \in \{1,3,7,14,21,30\}} \left[\bar{n} * \overline{ins}^{\,\nu} - \min_{j=1,\ldots,(n_\nu - \bar{n}+1)} \sum_{l=0}^{\bar{n}-1} ins^{\nu}_{j+l,min} \right] \tag{4.7}$$

where $\overline{ins}^{\,\nu}$ is the average monthly insolation over the 22 years,

$$\overline{ins}^{\,\nu} = \frac{\sum_{i=1983}^{2005} \sum_{j=1}^{n_\nu} ins^{\nu}_{ij}}{22 * n_\nu},$$

and the quantity ins^{ν}_{ij} is the site average insolation in the i-th year, $i = 1983, \ldots, 2005$, and in the j-th day of the ν-th month, for $j = 1, \ldots, n_\nu$, where n_ν is the total number of days in the considered month. On the other hand, $ins^{\nu}_{j,min}$ for $j = 1, \ldots, n_\nu$ is the minimal daily insolation over the same 22 years,

$$ins^{\nu}_{j,min} = \min_{i=1983,\ldots,2005} ins^{\nu}_{ij}.$$

The observation of the annual pattern of $Y^{\nu}_{\lambda,\theta}$ in each site at a given time of the year gives approximately the amount of energy that needs to be stored in a solar power plant in order to successfully cover for gaps in energy supply: the approximation is due to the fact that, while the number of considered consecutive days in the computation of the maximum monthly solar radiation deficit should vary in the discrete set $\mathbb{N} \cap [1; n_\nu]$, for $\nu = 1, \ldots, 12$, the buffer capacity we analyze, $Y^{\nu}_{\lambda,\theta}$, is computed considering a set of choices for the number of consecutive days $\bar{n}$ in the set $\{1, 3, 7, 14, 21, 30\}$, as outlined in equation (4.7). Moreover, all polar sites are excluded from the considered data set, since they are characterized by at least one month during the year ($\overline{\nu}$) during which insolation is such that $ins^{\overline{\nu}}_{ij} \equiv ins^{\overline{\nu}}_{j,min} \equiv \overline{ins}^{\,\overline{\nu}} \equiv 0$: thus, the buffer capacity obtained via (4.7) is not properly defined.

Moreover, for each site we obtain a functional datum $Y_{\lambda,\theta}(t)$ by smoothing $\{Y^1_{\lambda,\theta}, \ldots, Y^{12}_{\lambda,\theta}\}$ with a Gaussian kernel with bandwidth equal to 1.5: the collection of these functional data, indexed by the sites of S_0, is the input of the Bagging Voronoi classifiers algorithm.

For this application, we fix the number of bootstrap replicates to $B = 100$ and we test different values for the number n of elements of the Voronoi tessellation

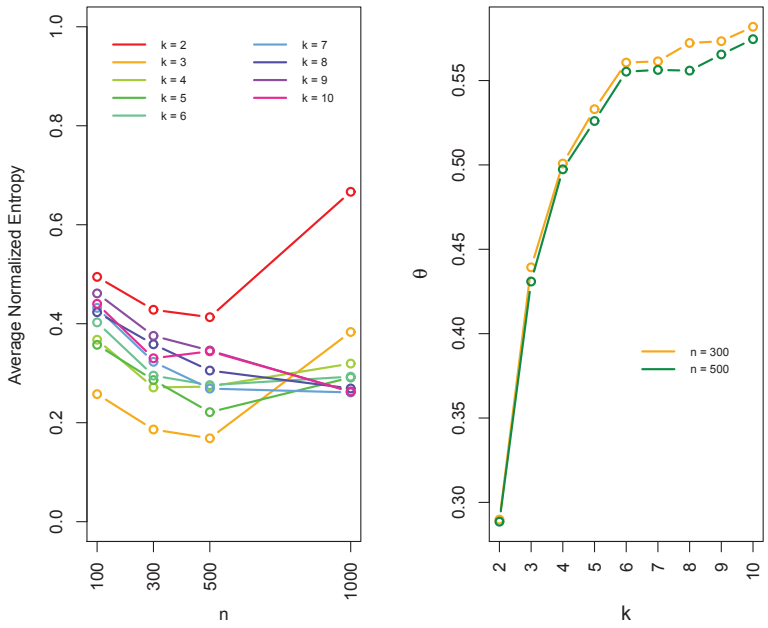

Figure 4.7: Results of Bagging Voronoi classifiers algorithm on buffer capacity data from the *Surface meteorology and Solar Energy database*: in the left panel, average normalized entropy obtained with different choices of k and for $n = 100, 300, 500, 1000$. In the right panel, values of the index θ introduced in (4.3) associated to the final classification with $k = 2, \ldots, 10$, and for $n = 300$ and $n = 500$.

and the number k of clusters initializing the clustering algorithm. The n elements are drawn from a uniform distribution on S, the surface of the sphere of diameter equal to the earth. The set of nuclei for the Voronoi tessellation is then chosen by selecting the n sites among those in S_0 nearest in terms of geodesic distance to each of the n generated elements. We then use a Gaussian isotropic kernel to calculate local representatives, and we choose the first $p = 3$ functional principal components to project data, since they explain a proportion of total variance that exceeds 95%. Finally, for clustering the n representatives we use k-means with the L^2 semi–metric induced by the principal components.

In Figure 4.7, for different values of n and k, the performance of the Bagging Voronoi classifiers algorithm is evaluated both in terms of average normalized entropy (i.e., sharpness of the image) and in terms of the index θ defined in (4.3) (i.e., differences among clusters). In the left panel of Figure 4.7 the values of the average normalized entropy are reported. The first fact to be noticed is that for most values of k, $n = 500$ provides a good choice to obtain a neat image. For small values of k, $n = 500$ is actually a minimum over the tested values, and it is hence chosen for setting the algorithm. Secondly, given $n = 500$, one can see that, in terms of classification sharpness, good values of k seem those between 3 and 7. In particular, two local minima are observed for $k = 3$ and $k = 5$.

To further investigate the choice of k, in the right panel of Figure 4.7 the

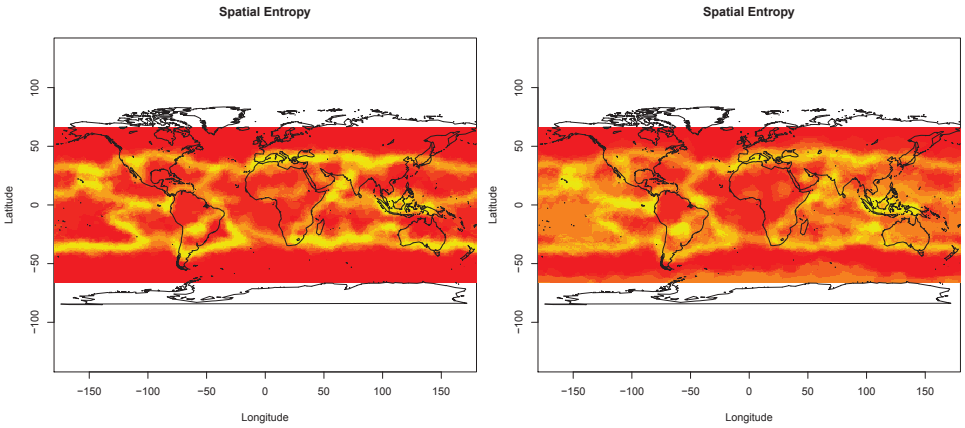

Figure 4.8: Results of Bagging Voronoi classifiers algorithm on buffer capacity data from the *Surface meteorology and Solar Energy database*: normalized spatial entropy maps associated to the classification with $k = 5$ (left) and $k = 6$ (right). Colors from red to white correspond to values from 0 to 1; higher values identify areas where classification is more uncertain.

values of the index θ defined in (4.3) are reported as a function of k for two values of n. Notice that the shape of the graph is robust with respect to the dimension of the Voronoi tessellation. The plot suggests $k = 6$ as the maximum reasonable number of clusters. Greater values of k are not paid off by a significant improvement in the description of data. Slightly smaller values for k ($k = 4, 5$) seem admissible as well, even though minor features are probably lost. Values $k = 2, 3$ are definitely not suggested. On the whole, $k = 5$ and $k = 6$ seem to be good choices for obtaining a spatially neat classification, with important differences among clusters. This is confirmed by inspection of Figure 4.8, where the two maps of spatial normalized entropy obtained for $k = 5$ and $k = 6$ are reported (left and right panel, respectively). Both plots show a neat classification, even though the one corresponding to $k = 5$ seems more reliable. In Figure 4.9 the final clusters obtained with $k = 5$ and $k = 6$ are reported on the Earth surface. The two classifications do not contradict each other; on the contrary, the latter is a refinement of the former supporting the robustness of the obtained classification.

In particular, for $k = 5$, the Bagging Voronoi classifiers algorithm identifies different homogeneous macro-areas which – prima facie – seem interpretable in terms of the observed phenomenon. A climatological analysis, which is beyond the scopes of this paper, could of course deepen their explanation. Indeed, the same macro-areas are not captured by customary unsupervised classification procedures, that do not take into proper account the spatial dependence among data. The final results for the choice of $k = 5$ are shown in Figure 4.9 and 4.10 (left panels). In Figure 4.10 (left panel) a sample of local representatives is shown, each representative colored with a label corresponding to the macro-area it belongs to (Figure 4.9, left panel). The red cluster is characterized by a non-seasonal pattern, and by intermediate average buffer capacity along the year. It

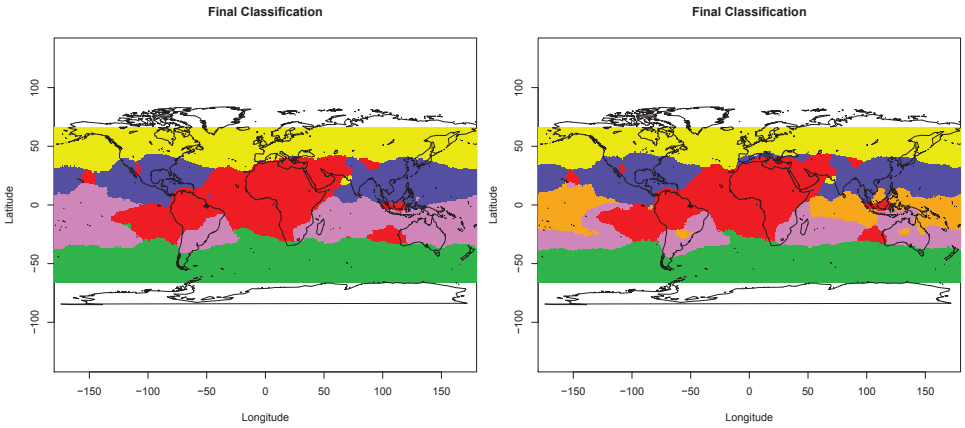

Figure 4.9: Results of Bagging Voronoi classifiers algorithm on buffer capacity data from the *Surface meteorology and Solar Energy database*: final classification map obtained via a majority vote on frequencies of assignment, and by setting $k = 5$ (left) or $k = 6$ (right).

covers Africa, Middle-East and equatorial America and its presence is not explained only in terms of latitude. From North to South we can then identify four clusters with seasonal patterns depending on the hemisphere and on the average buffer capacity along the year: north-low (yellow), north-high (blue), south-high (violet), south-low (green). It is interesting to note that, while in the Americas all five clusters are present, the north-high and south-high clusters are absent in Europe and Africa, and the red cluster is almost absent in Asia.

The main difference obtained by choosing $k = 6$ is the fact that a new cluster, which spurts from the former south-high (violet) cluster, appears along the equator (see Figure 4.9, right panel). This cluster, depicted in orange, is characterized by a very high seasonality of the buffer capacity (see Figure 4.10, right panel) that makes it strongly unsuited for electricity production by solar power. All other clusters remain unaffected while moving from $k = 5$ to $k = 6$, in particular the red one. Interestingly, from an engineering point of view, the red cluster is the one that shows an annual buffer capacity pattern which is optimal in terms of electricity production by solar power: it needs the minimal buffer capacity installation (the maximal annual need for energy is the lowest among the five detected patterns), associated to a constant reliability along the year.

4.5 Conclusion

In this Chapter a new method, the Bagging Voronoi classifiers algorithm, was proposed to deal with large data sets of geo-referenced functional data. This is a problem not largely treated so far in the literature that poses new methodological challenges. These concern for instance the handling and the statistical analysis of huge sets of complex data, or the intrinsic difficulty of modeling and estimating spatial dependence in a high-dimensional setting. Even though the

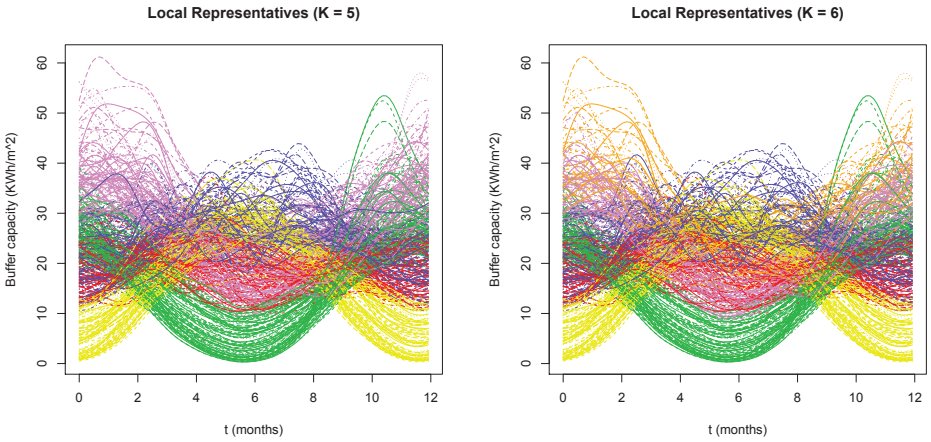

Figure 4.10: In the left panel, a set of functional local representatives obtained with $n = 500$ in one of the iterations of the algorithm and clustered with $k = 5$; the colors for the plot are chosen coherently with the final classification map in Figure 4.9 (left). In the right panel, the same set of functional local representatives is clustered with $k = 6$, and colored coherently with the final classification map in Figure 4.9 (right).

Bagging Voronoi classifiers algorithm is here presented as a clustering procedure, its rationale is much more general: the algorithm, if suitably set, is virtually able to manage many kinds of analysis, e.g., dimensional reduction, supervised classification, regression, etc. Indeed, the naive idea beyond the algorithm is simply to (*i*) replace the original data set with a reduced one, composed by local representatives of neighborhoods covering the entire investigated area; (*ii*) analyze the reduced data set according to the purpose of the analysis; (*iii*) repeat the analysis many times for different reduced data sets associated to different sets of neighborhoods, thus obtaining many different weak formulations of the analysis; (*iv*) finally, bag together the weak analyses to obtain a final strong formulation.

This approach does not require an explicit modeling of the spatial dependence among data, which is instead explored by means of random local aggregation of neighboring data. We showed that, when spatial dependence is not negligible, this approach is able to exploit the mutual information carried by neighboring data and thus generating a final boosted analysis. Indeed, the final analysis turns out to be not only more performing than the weaker analyses it is composed of, but also more effective than an analysis directly performed on the original data set. The proposed approach is also efficient from a computational perspective, since it allows the statistical analysis of data sets that would be untractable because of their large sample size. Indeed the Bagging Voronoi classifiers algorithm replaces a single analysis of a huge data set by many analyses of small data sets, that can possibly be carried on through parallel programming.

Probably, the major feature of the algorithm is its high flexibility. Indeed, the Bagging Voronoi approach can be modified according to the particular nature of

the problem under investigation. For an expressive illustration, in the present work we generated the random neighborhoods by means of Voronoi maps with random nuclei, we created the local representatives by spatial smoothing, we clustered them via a k-mean algorithm, and we aggregated the results through a majority vote. All these specifications can be modified for adapting the algorithm to the problem at hand, without changing the rationale motivating it.

A key point to obtain a successful implementation of the Bagging Voronoi classifiers algorithm appeared to be the right choice for the number n setting the dimension of the Voronoi tessellations. Indeed, simulations show that when spatial dependence is sufficiently strong, there exists a value of n optimizing the bias-variance trade-off in the local representatives sample. This optimized information-gathering maximizes the power of the weak formulations of the analysis. Interestingly, we found that a novel criterion based on a entropy index could be extremely effective in applications for choosing the value of n. A theoretical study on the connections between spatial dependence and data distribution on the one hand, and the optimal dimension n of the Voronoi tessellations on the other, is of paramount importance for a deeper understanding of the theoretical properties of the algorithm; it will be the object of future work.

Chapter 5

Exploration of Local Telephonic Time Patterns for City Planning

In this Chapter we describe a case study in city planning[1] that stimulated our research in the analysis of functional data spatially distributed on a lattice[2]. Data are measures along time (every 15 minutes for two weeks) of the use of the Telecom mobile phone network across a lattice covering the area of Milan (Italy); these measures, named *Erlang*, refer to the total number of calls dialled by mobile phones located in each particular site of the lattice, in each considered time interval. A subsample of randomly selected telephonic patterns is shown in Figure 5.1: the most relevant characteristic of these data, as it is evident from the picture, is the presence of strong localized features like peaks. Moreover, Erlag measurements show a great variability across locations: this fact seems natural, since in the same lattice are included both sites referred to the centre of Milan, and sites referred to the countryside. The strenght of our analysis stands in finding the most exhaustive, and at the same time synthetic, description of the signal components, and of their action both in time and space.

The great advantage in the exploitation of Erlang measures is that they are costless and freely available to any mobile phone network provider; nevertheless, the analysis of such a kind of data can give insight on different aspects of the urban area they are referred to, and can be developed with various scopes: the segmentation of the area into districts characterized by homogeneous telephonic patterns; the identification of a set of "reference signals" able to describe the different patterns of utilization of the mobile phone network; the description of the influence of each telephonic pattern in each site of the lattice, via projection of raw data on the selected reference signals.

We can expect that spatial dependence among functional data, when carefully exploited, could help both in the segmentation procedure, and in the identification of relevant functional features. We thus want to perform a data–driven dimensional reduction of functional data while accounting for spatial dependence.

[1] Data are courtesy of Convenzione di ricerca DiAP - Telecom Italia, Politecnico di Milano (Italy).

[2] The analysis has been conducted in collaboration with Fabio Manfredini and Paolo Tagliolato from Dipartimento di Architettura e Pianificazione (DiAP), Politecnico di Milano.

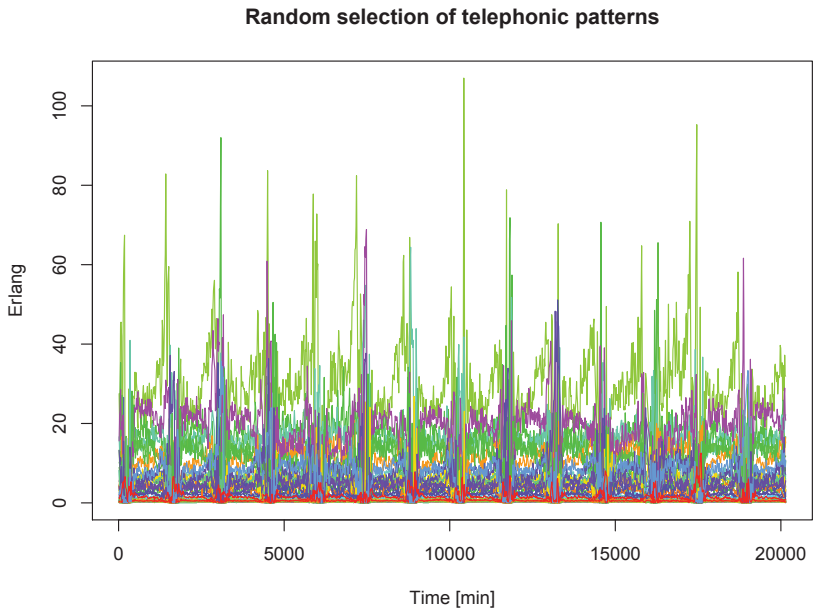

Figure 5.1: A sample of 50 randomly chosen telephonic time patterns.

To this aim, the Bagging Voronoi algorithm idea can be exploited: at each iteration, a random tessellation of the area is generated and local representatives of functional data are computed. Dimensional reduction is then achieved through the identification of relevant and interpretable basis functions (in the described application we will consider treelets, i.e., a data-driven wavelets), with respect to the considered local representatives sample. Finally, treelet bases provided by different iterations are pooled and reordered (via a proper bases matching procedure) to obtain a better estimate of the final reference basis.

The Chapter is structured as follows. In Section 5.1, the general paradigm introduced in Chapter 4 for the problem of clustering spatially dependent functional data is generalized to the scopes of the present analysis (dimensional reduction). In Section 5.2 the data–driven strategy for dimensional reduction, based on treelet decomposition, is briefly described. The basis matching procedure adopted to align the B final bases obtained via Bagging Voronoi is detailed in Section 5.3. Finally, Section 5.4 sums up the adopted strategy for the analysis of telephonic time patterns, and Section 5.5 is devoted to the detailed description of the case study.

5.1 From Bagging Voronoi classifiers to a general Bagging Voronoi strategy

The assumption of a latent field of labels $\Lambda_0 : S_0 \to \{1, \dots, L\}$ being associated to each site of the lattice S_0 is not proper for the considered application. Indeed, in the considered city planning application, the complexity of the considered area, and the peculiarities of some sites with respect to the others in terms of

the observed signal, are so high that a segmentation of the region fails to give interpretable results. We instead prefer to think to the observed functional signal as the result of the combination of different basis functions, whose contribution to the observed datum is different in each site; moreover, we immagine the contribution of each basis function to vary continuously in the lattice, due to the continuity of the observed phenomenon.

In such a situation the aim of the analysis, rather than being classification, is mainly *dimensional reduction*: we need to decouple the functional signal into its constitutive parts, to eventually analyze each of them in the light of urban planning considerations; however, the most appropriate basis has to be chosen taking into account also the spatial dependence which is intrinsic in the spatial localization of each function. Moreover, we aim at summarizing the spatial distribution of the influence of each functional pattern on the observed data.

Therefore, we expect the final result of the analysis to be a couple:

- a reference basis, describing some typical attitudes of the network user;
- a set of maps, one for each basis function, describing the relative influence of each basis component in each site of the lattice.

To this aim, the Bagging Voronoi strategy introduced in Chapter 4 could again be exploited. The driving idea is simple: once functional local representatives have been identified, we use this "less dependent" and "less noisy" data set to perform any kind of statistical analysis (not just classification). Indeed, at each replication of the first two steps, a local representatives sample is obtained, which exploits a specific structure of spatial dependence; this sample can be used to obtain a coarse estimate of the unknown reference basis which composes the functional signal. A more accurate final reference basis is obtained after B replications, by *bagging* together all single bases. Note that the same ideas already illustrated in Chapter 4 can be applied here: higher values of B, imply a higher accuracy of the final estimate, which includes all the estimates of the B single replicates.

The bagging strategy will be described in Section 5.3. In the next Section we will detail the chosen method to find a reference basis at each replication of the algorithm.

5.2 Dimensional reduction: treelet decomposition

A possible approach to p–dimensional reduction of functional data is using a *treelet basis*, introduced by Lee *et al.* in [44]. This data–driven basis seems the most suited to the application of our procedure in the analysis of telephonic time patterns, which presents extremely localized functional features interesting to the scopes of the analysis.

Indeed, as in multi-resolution analysis, treelets provide a set of "scaling functions" defined on the nested subspaces $V_0 \supset V_1 \supset \cdots \supset V_J$, and a set of orthogonal "detail functions" defined on residual spaces $\{W_j\}_{j=1}^{J}$, where $V_j \oplus W_j = V_{j-1}$ for all $j = 1, \ldots, J$. We remark that treleets are very close to wavelets, even though they are not a wavelet trasform; indeed, in treelets computation, the

wavelet approach is mixed with principal components analysis, which is hierarchically performed on the couple of closest variables at a given level.

We have also to remark that treelet algorithm has been designed and developed for the case of sparse unordered high dimensional data. However, since our aim is decoupling the signal into a functional basis as sparse as possible, we can nevertheless use this algorithm provided the sample of functional local representatives is described on a fine equispaced grid of common abscissa points $\mathbf{t} = \{t_j\}_{j=1}^J$ via a proper evaluation basis (see Ramsay and Silverman for details [61]).

5.2.1 Treelets algorithm

At each level of the tree, the most similar variables are grouped together and replaced by a coarse–grained *sum variable*, and by a residual *difference variable*. The new variables are computed by a local principal components analysis in two dimensions. Difference variables are then stored, and only sum variables are processed at higher levels of the tree.

More precisely, the algorithm is initialized with a Dirac basis $B_0 = [\phi_{0,1}|\phi_{0,2}|\cdots|\phi_{0,J}]$, where B_0 is the identity matrix in $\mathbb{R}^J$,and with the set of original variables $\mathbb{X}^{(0)} := \mathbb{X}$. We denote with $\mathbb{X}$ the design matrix of our functional sample $\chi_1, \ldots, \chi_n$ described with the evaluation basis; hence, $\mathbb{X}_{ij} = \chi_i(t_j)$, for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. We can also indicate the design matrix $\mathbb{X}$ with the following notation $\mathbb{X} = [\mathbf{X}_1|\cdots|\mathbf{X}_J]$, i.e. $\mathbf{X}_j \in \mathbb{R}^n$ for $j = 1, \ldots, J$ is the vector including the values assumed by the j-th variable in the sample, and we can indicate with $\mathbf{x}^{(0)} = (s_{0,1}, \ldots, s_{0,J})^T$, where $s_{0,j} = t_j$ for $j = 1, \ldots, J$ the set of original variables. We also denote with $\hat{\Sigma}^{(0)} \in \mathbb{R}^{J\times J}$ the sample covariance matrix. The initial set of sum variables is $\mathcal{J}^{(0)} = \{1, \ldots, J\}$. Finally, a similarity measure $\tilde{d} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ between variables is chosen to give the similarity matrix $\hat{M}^{(0)} \in \mathbb{R}^{J\times J}$, whose elements are such that

$$\{\hat{M}^{(0)}\}_{ij} = \tilde{d}(\mathbf{X}_i, \mathbf{X}_j), \quad i, j = 1, \ldots, J.$$

Then, the following steps are repeated for $l = 1, \ldots, L$:

1. **Find the two most similar sum variables according to the similarity matrix $\hat{\mathbf{M}}^{(\mathbf{l-1})}$**. Let

$$(\alpha^l, \beta^l) = \underset{i,j\in\mathcal{J}}{\operatorname{argmax}}\, \hat{M}_{ij}^{(l-1)},$$

   with $i < j$, and where maximization is performed only over pairs of sum variables belonging to the set $\mathcal{J}$. As in standard wavelet analysis, difference variables (defined in step 3.) are not processed.

2. **Perform a local PCA on the pair of selected variables**. This means finding a Jacobi rotation matrix $R_l \in \mathbb{R}^{J\times J}$ that decorrelates $\chi_i(t_{\alpha^l})$ and

$\chi_i(t_{\beta^l})$, for all $i = 1, \ldots, n$

$$R_l(\alpha^l, \beta^l, \theta^l) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & -s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix},$$

where $c = cos(\theta^l)$, $s = sin(\theta^l)$ and θ^l is a rotation angle such that $|\theta^l| \leq \pi/4$ and $\hat{\Sigma}^{(l)}_{\alpha^l\beta^l} = \hat{\Sigma}^{(l)}_{\beta^l\alpha^l} = 0$, being $\hat{\Sigma}^{(l)} = R_l^T \hat{\Sigma}^{(l-1)} R_l$. This transformation corresponds to a change of basis $B_l = B_{l-1}R_l$, to a new design matrix $\mathbb{X}^{(l)} = \mathbb{X}^{(l-1)}R_l$ and to a change of coordinates $\mathbf{x}^{(l)} = R_l^T \mathbf{x}^{(l-1)}$. Update the similarity matrix $\hat{M}^{(l)}$ to be the one associated to $\mathbb{X}^{(l)}$.

3. **Multi–resolution analysis**. Assume that the indices α^l and β^l respectively correspond to the first and second principal components, i.e. we assume $\hat{\Sigma}^{(l)}_{\alpha^l\alpha^l} \geq \hat{\Sigma}^{(l)}_{\beta^l\beta^l}$. Define the sum and difference variables at level l, s_l and d_l, as the α^l-th and β^l-th element of the vector $\mathbf{x}^{(l)}$, respectively; similarly define the scaling and detail functions ϕ_l and ψ_l to be the columns α^l and β^l of the basis matrix B_l. Remove the difference variable from the set of sum variables, $\mathcal{J}^{(l)} = \mathcal{J}^{(l-1)} \backslash \{\beta^l\}$.

Note that at each level l of the tree we have the following *treelet decomposition*

$$\mathbb{X} = \sum_{i=1}^{J-l} s_{l,i}\phi_{l,i} + \sum_{i=1}^{l} d_i \psi_i,$$

where the new set of scaling vectors $\{\phi_{l,i}\}_{i=1}^{J-l}$ is the union of the vector ϕ_l with the scaling vectors $\{\phi_{l-1,j}\}_{j \neq \alpha^l, \beta^l}$ from the previous levels, and the new coarse–grained sum variables $\{s_{l,i}\}_{i=1}^{J-l}$ are the projections of the original data onto these vectors. As in standard multi–resolution analysis, the first term is the coarse–grained representation of the signal, while the second sum captures the residuals at different scales.

The output of the algorithm can be summarized in terms of a hierarchical tree with a height $L \leq J-1$ and an ordered set of rotations and pairs of indices $\{(\theta_l, \alpha_l, \beta_l)\}_{l=1}^{L}$. The default choice of the treelet transform is a maximum height tree with $L = J-1$: this choice leads to a fully *parameter–free* decomposition of the data, and it is the one adopted in the application discussed in Section 5.5. One can alternatively also choose any of the orthonormal bases at levels $l < J-1$ of the tree: data are then represented by a certain number of coarse–grained sum variables for a set of clusters detected at that level of the tree, and by difference variables that describe the finer details. We will not give further details on treelet decomposition; for a more detailed discussion about theoretical issues see Lee *et al.* [44].

Considering the sample of local representatives $\{\mathbf{g}_1^b, \ldots, \mathbf{g}_n^b\}$, evaluated on a fine grid of common points of the abscissa $\mathbf{t} = \{t_j\}_{j=1}^J$, at the b-th iteration of the Bagging Voronoi algorithm, we can use the treelet algorithm to obtain the set of orthonormal basis functions $\{\psi_l^b\}_{l=1}^{J-1}$, also referred to as detail functions: they represent the "difference" between data projections at two consecutive levels in the tree, and are mutually orthogonal. The orthonormal basis of the detail functions obtained at each iteration, for $b = 1, \ldots, B$, is then used as a coarse estimate of the reference basis which will be finally used to project the data; the procedure to obtain the reference basis from the B coarse treelet bases will be detailed in the next section.

5.3 1-median alignment for bases matching

In many applications, clustering is not the final aim of the analysis, but we might for example be interested in outlier detection, or simply in achieving a better understanding of the observed phenomenon. In this section another possible way of summing up coarse results obtained along runs of the Bagging Voronoi algorithm is introduced, which has the aim of aggregating sets of basis functions rather than classifiers. Moreover, the bagging strategy described in the following can be interpreted as a "discrete version" of the Procrustes alignment procedure described in subsection 1.3.2.

We shall distinguish two possible situations. If the functional basis used for p-dimensional reduction of the sample of local representatives is fixed along runs (e.g. a wavelet basis, or a Fourier basis), then no matching is needed and we can simply average the B bases component by component. The bagging strategy is then simply an averaging strategy, and we can compute the reference basis by choosing among the possible notions of centrality for functional data described in subsection 1.1.1. If, instead, the functional basis is data–driven (treelet basis is data–driven and unordered, and thus changing at each repetition of the algorithm), then a matching of basis elements obtained along runs is needed while computing the final reference basis.

Different approaches to *bases matching* are possible. In the present work, we developed a procedure of *1-median bases alignment*, which jointly computes the reference basis from the B coarse bases, while also reordering them. This procedure is inspired to the joint clustering and alignment methods described in Chapter 2: each object here is a multivariate fuctional data (one of the coarse bases), and we look for the unique prototype (the reference basis) which best describes the set of functional objects, while also aligning their components (i.e., changing the order of functions in the bases) by permutation of the basis functions.

Consider all bases obtained along runs, $\left\{\psi_1^b, \ldots, \psi_{J-1}^b\right\}_{b=1}^B$, and choose a proper measure $d(\cdot, \cdot)$ of the distance (or dissimilarity, i.e. semi–metrics, as in the previous Chapters) between two functions, which in our application will be the $L^1(T; \mathbb{R}^J)$ distance. Aim of 1-median bases alignment procedure is to calculate, starting from these B bases, one reference (template) basis $\{\tilde{\psi}_1, \ldots, \tilde{\psi}_{J-1}\}$. A sketch of the proposed algorithm is given in the following subsection.

5.3.1 1-median bases alignment algorithm

Initialize the algorithm by randomly selecting a reference basis for the first iteration, $\{\tilde{\psi}_1^{[0]}, \ldots, \tilde{\psi}_{J-1}^{[0]}\}$, among the available coarse bases. Then, iterate until convergence the following two basic steps (consider the l-th iteration):

1. **Alignment step.** For each basis find the best matching, by permutation of the components, to the reference basis $\{\tilde{\psi}_1^{[l-1]}, \ldots, \tilde{\psi}_{J-1}^{[l-1]}\}$ according to the measure $\tilde{d}(\cdot, \cdot)$. Hence, $\forall\ b = 1, \ldots, B$ a permutation $\{k_1^{b,[l]}, \ldots, k_{J-1}^{b,[l]}\}$ in the order of basis functions $\{\psi^b_{k_1^{b,[l]}}, \ldots, \psi^b_{k_{J-1}^{b,[l]}}\}$ is found such that the distance to the reference basis at the l-th iteration is minimized;

2. **Estimation step.** Calculate the functional median of the B reordered bases according to the measure $\tilde{d}(\cdot, \cdot)$, thus obtaining a reference basis. Hence, the j-th basis function in the reference basis at the l-th iteration, $\tilde{\psi}_j^{[l]}$, for $j = 1, \ldots, J-1$, is identifyied as

$$\tilde{\psi}_j^{[l]} = \underset{\varphi \in E}{\operatorname{argmin}} \sum_{b=1}^{B} d(\psi^b_{k_j^{b,[l]}}, \varphi).$$

   In the specific case, $E \equiv L^1(T; \mathbb{R}^J)$, d is the natural norm in this space, and thus the functional median can be estimated as the point–wise median of the j-th elements of all the B bases, aligned at the current alignment step.

The algorithm is stopped when the permutation of the basis functions is constant in two subsequent iterations for all bases (no change in the matching). If the permutation in the order of basis elements in the alignment step remains stable in two subsequent runs for all bases, i.e., for some $\bar{l}$ we have $k_j^{b,[\bar{l}-1]} \equiv k_j^{b,[\bar{l}]}$ for all $b = 1, \ldots, B$ and $j = 1, \ldots, J-1$, the final reference basis is identifiyed as $\{\tilde{\psi}_1, \ldots, \tilde{\psi}_{J-1}\} \equiv \{\tilde{\psi}_1^{[\bar{l}]}, \ldots, \tilde{\psi}_{J-1}^{[\bar{l}]}\}$.

5.4 A general procedure for detecting relevant functional patterns in presence of a spatial dependence

We can now give the final procedure for detecting a proper dimensional reduction of functional data indexed by a lattice. The pseudocode scheme of the proposed procedure is sketched below.

Algorithm. Bagging Voronoi Functional Bases.

Bootstrap:
Initialize B, n. Choose a metric $d(\cdot,\cdot)$ for computing the tessellation.
`for` $b := 1$ `to` B `do`

`step 1.` generate a random Voronoi tessellation of the lattice, i.e., isolate neighbouring groups of data, to capture potential spatial dependence;

`step 2.` identify a local representative for each element of the tessellation to sum up local information: neighbouring data are most likely drawn from the same functional distribution;

`step 3.` perform functional dimensional reduction of local representatives to select relevant functional features in the data via a proper functional basis.

`end for`

Aggregation:

- `if` dimensional reduction is data–driven `do`
  perform `bases matching` via 1-median bases alignment algorithm, according to a proper semi–metric $\tilde{d}$: match the basis functions along bootstrap replications, to obtain a final reference basis.
  `else`
  compute a `reference basis` by averaging corresponding basis functions along bootstrap replications.
- project the whole functional dataset on the reference basis, to obtain `spatial scores maps`.

This procedure allows to gain two final results:

- a reference basis which explains relevant functional patterns present in the dataset, while also accounting for spatial dependence;
- a set of maps, one for each reference basis function, giving the influence of each functional pattern in each site of the lattice, by projection of raw data on basis functions.

Note that only $p < J - 1$ basis functions in the reference basis are finally selected: the ones which, according to a graphical inspection, are explaining interpretable features in terms of network user behaviour. Moreover, this graphical selection can be evaluated and strengthened by considering the scores variance: only treelets whose associated scores vector has variance higher than a given threshold will be considered. In our application, only treelets characterized by high variance scores vectors had a pretty neat interpretation, and where therefore included in the analysis.

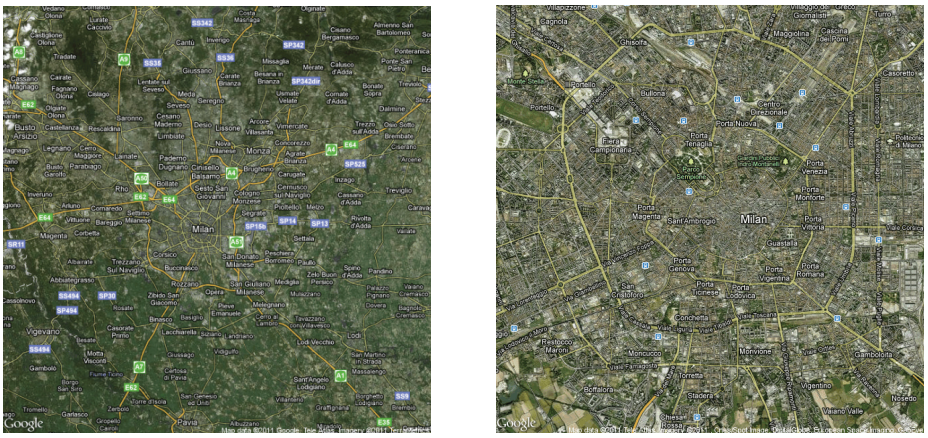

Figure 5.2: Map of the geographical region in which the lattice of Erlang measures used in the analysis of Milan metropolitan area is located (left panel), and a zoom on Milan urban area considered in subsection 5.5.2 (right panel), both obtained via `http://maps.google.com`

5.5 Case study: the analysis of local telephonic time patterns

In this last section, we describe the case study which motivated the generalization of the Bagging Voronoi strategy to dimensional reduction purposes. Aim of the present study is to gain insight on the urban structure of a city (Milan), and thus to perform a city planning analysis of the city itself, through the analysis of mobile phone traffic patterns along time. In particular, we consider measures along time of the total minutes of calls dialled in Telecom mobile phones network every quarter of hour, on each site of a lattice covering the whole area of Milan (Italy). The scopes of the analysis are various: spatial segmentation, functional dimensional reduction, and the detection of the effect of each functional pattern on the spatial distribution of the signal, to identify districts characterized by a homogeneous telephonic pattern.

5.5.1 Analysis of Milan Metropolitan Area

In this first part of the analysis, we consider data on a lattice S_0 composed by $N_0 = |S_0| = 65450$ sites, 275 in the W-E direction and 238 in the N-S direction; the site $\mathbf{s}_{NW}$, which is located in the extreme N-W position, has geographic coordinates $(45.7333^\circ\ N, 8.83455^\circ\ E)$, and the dimension of each site in S_0 is 308.96 metres W-E and 231.72 metres N-S (which means a total area covered by the lattice of 84.96×55.15 km – W-E $\times$ N-S). A map of the geographical region cosidered in the analysis, in which sites of the lattice S_0 are located, is shown in Figure 5.2 (left panel).

In each site of this lattice, every quarter of hour a quantity named *Erlang* is recorded from March 18^{th}, 2009, 00:15 am, till March 31^{st}, 2009, 23:45 pm (from now on, we will call T this time interval). Erlang is a measure of the total

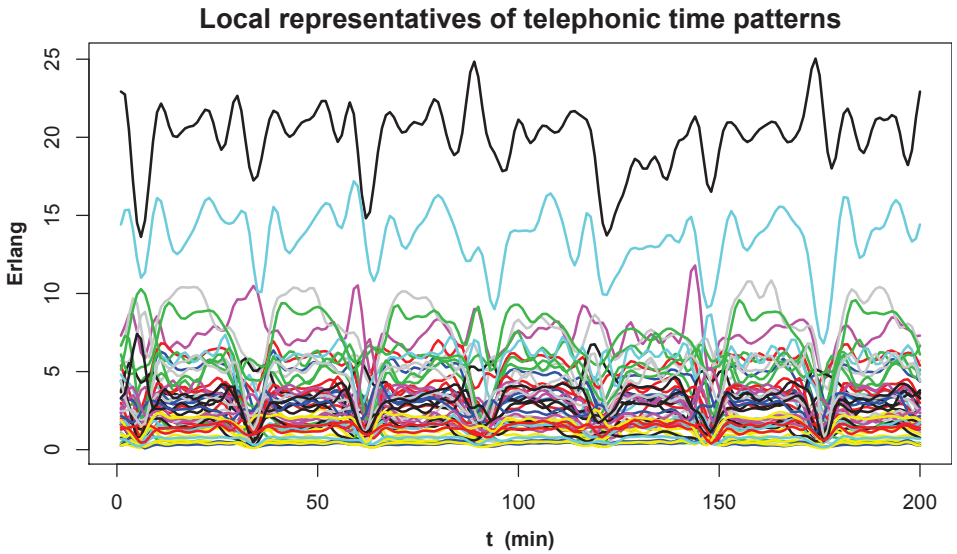

Figure 5.3: A sample of 50 randomly chosen local representatives of telephonic time patterns, $n = 400$.

minutes of calls dialled from numbers corresponding to mobile phones active in that site of the lattice, in a fixed time interval (a quarter of hour); thus, Erlang is not related to single users, but to the global intensity of the mobile phone network at that time/location. This measure is extremely low–cost, since it is easily provided by network antennas recordings: this fact motivates the interest in this analysis, which could deepen the understanding of the city structure using already existing information.

The Erlang measures for all $\mathbf{s} \in S_0$, and for all $t \in T$, are then normalized with respect to a reference value for Erlang, i.e. the Erlang average site intensity in the whole region (Lombardia): this is obtained, for each $\bar{t}$ fixed, dividing the total Erlang in Lombardia measured at $\bar{t} \in T$ by the number of sites in a lattice covering the whole region, in which Erlang was measured at $\bar{t}$. We should remark at this point that there are time intervals in which the total Erlang has not been evaluated, and hence data recorded in S_0 cannot be normalized; since these time intervals are treated as missing data, functional data are evaluated on an unequally spaced grid of abscissa points. The final length of the vector of abscissa values, after normalization, is 1308. Moreover, a normalization of Erlang measures along time in each site of the lattice is also needed: both negative measures and values of Erlang which exceed a reasonable threshold should be removed, since they are due to errors in the measurement system; under suggestion of network provider experts, we calculated the threshold using Chebicev's inequality with $k = 50$.

In the application of the Bagging Voronoi strategy to the analysis of local telephonic time patterns, we set $B = 50$ and $n = 400$; the choice of a tessellation of 400 elements has been carefully investigated in order to choose the value of n providing best results.

The nuclei of the Voronoi tessellation are repeatedly drawn from a uniform distribution on the regular lattice S_0, and the tessellation is based on euclidean distance in $\mathbb{R}^2$. We choose Gaussian isotropic weights to calculate local representatives, and we perform dimensional reduction using a treelet basis (see Section 5.2): this choice is due to the need for catching the very localized features of telephonic patterns through an adaptive data–driven basis; this localized behaviour, e.g. the rapid increase in the network using intensity during the first hours in the morning, is evident both from Figure 5.1, and from Figure 5.3, where a sample of 50 randomly selected local representatives, obtained in one of the iterations of the algorithm, is shown. Finally, being the chosen basis for dimensional reduction data–driven, bases matching is performed via 1-median alignment algorithm described in Section 5.3.

In Figures 5.4 and 5.5, the most relevant basis functions selected from the reference basis obtained via 1-median alignment are shown (the vertical lines are drawn at midnight, the dotted lines are drawn every two hours, and the origin is at 00:00 of Wednesday, March 18^{th}, 2009). The total length of the time interval T on which the basis elements are defined is one week, since the original functional data on the lattice have been reconstructed via a Fourier evaluation basis of 27 elements[3] with period of one week. The basis functions shown in the two pictures are relevant from a city planning point of view, since they provide interesting insight on the different patterns of mobile phone using detectable in the considered area. Moreover, the influence of each functional phone using profile on each site of the lattice has to be related to the most relevant urban typology of the area the site is referred to (e.g., a residential area, a shopping area, a metro or train station, a university, a hospital, ...).

For what concerns the interpretation of the shape of each basis function, we recognize in the first one (Figure 5.4) the description of the average mobile phone usage intensity in the whole network, with a decrease during the night and an overall positive contribution. The second basis function has to be related to a use of mobile phone during working hours (8:30 am till 8 pm), versus a non–working hours usage. The third one activates only during the late night / early morning hours (3 am till 6 am) in weekdays, and seems thus related to working activities that take place during the night (hospitals, big market areas, tram or buses deposits, ...); it will be extremely important to look at data projections with respect to this basis function, in order to evaluate if the proposed analysis strategy is able to effectively catch the exact locations of very localized night activities. The fourth and fifth basis functions are closely related, since they both concern morning commuters: the fourth one activates during the traffic peak in the morning of weekdays (6:30 am till 8:30 am), with a minor contribution on Saturday, while the fifth one catches the late morning commuters, or most likely the opening of shopping areas (8:30 am till 10 am). The sixth basis function is the less interpretable among the elements of the selected final reference basis, and might describe the late evening / night mobile phone usage, possibly associated to nightlife.

[3]The dimension of the Fourier basis was chosen by evaluation of the power spectrum corresponding to basis with different dimensions.

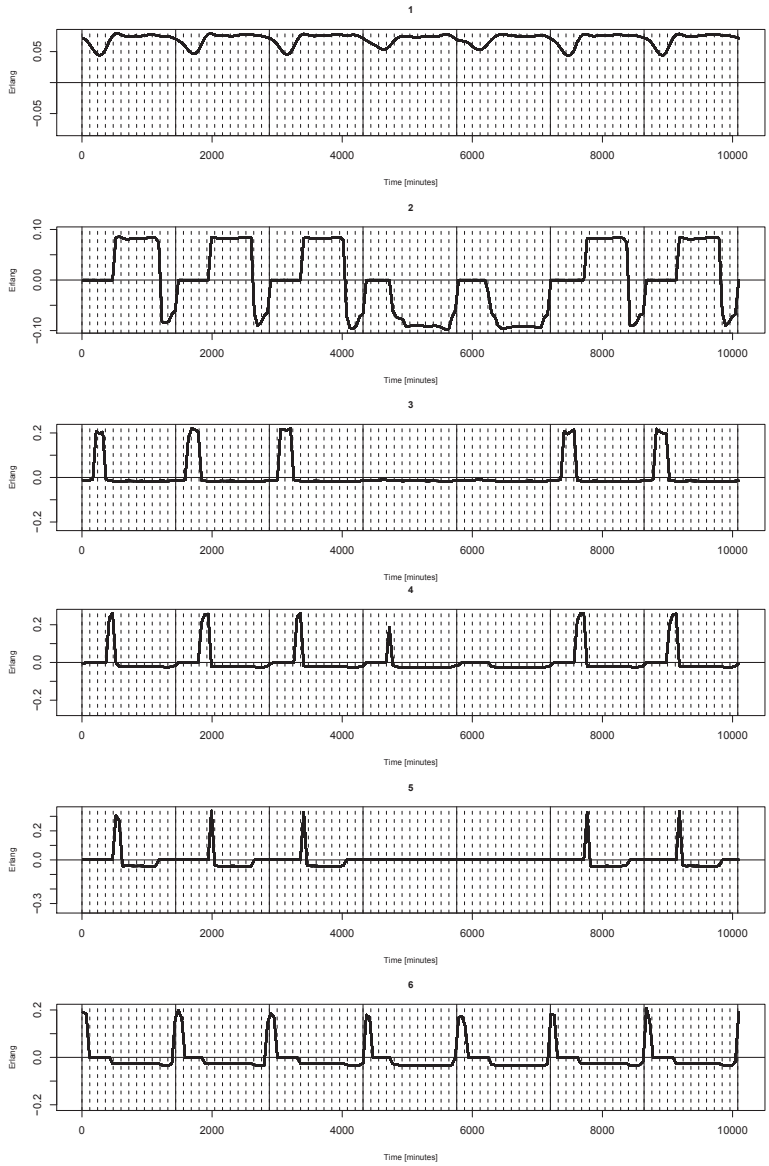

Figure 5.4: Selected basis functions (from top to bottom, first to sixth function) from the reference basis identified via 1-median basis alignment procedure. The vertical lines correspond to midnight, and the origin is at midnight between tuesday and wednesday. The dashed lines are drawn every two hours.

Looking at Figure 5.5, we notice that also basis functions number 12 and 26 (respectively the first and third from the bottom in the picture) are closely linked to each other, being both activated only during the week and both referred to evening commuters: the former refers to late evening commuters (7 pm till 9 pm), while the latter seems associated to earlier working out (5 pm till 7 pm) and is also contrasting with the daily hours. Then, the other three relevant elements of

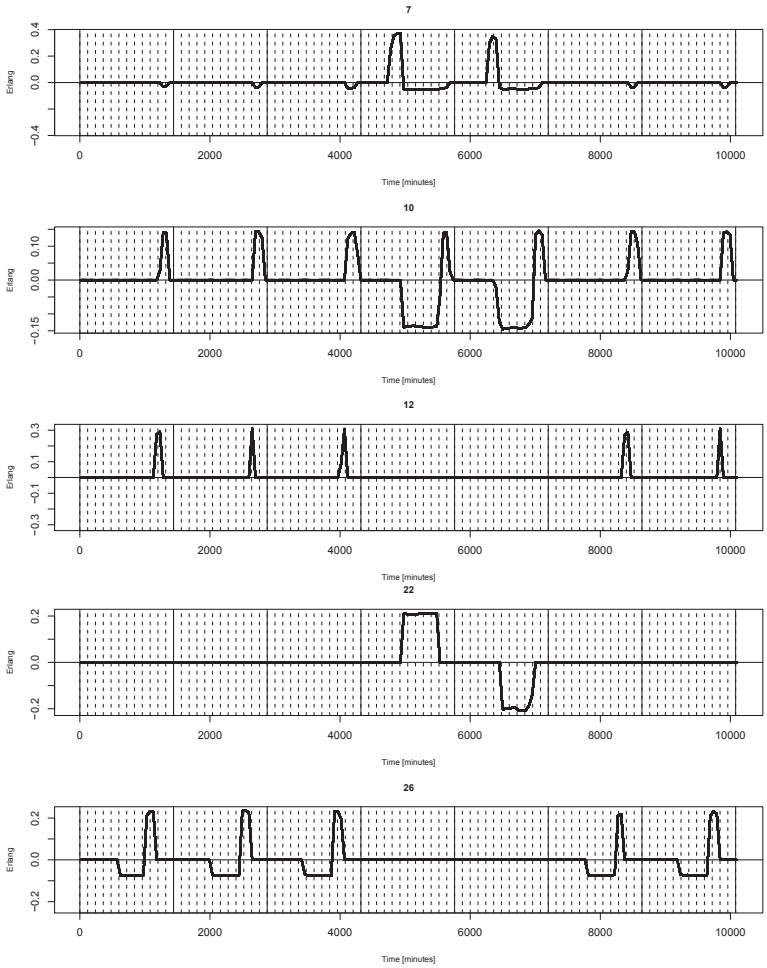

Figure 5.5: Selected basis functions (from top to bottom, functions 7, 10, 12 22 and 26) from the reference basis identified via 1-median basis alignment procedure. The vertical lines correspond to midnight, and the origin is at midnight between tuesday and wednesday. The dashed lines are drawn every two hours.

the final reference basis are the number 7, 10 and 22, which seem all devoted to catching the weekend mobile phone use patterns: while the seventh basis function (the first one from the top in Figure 5.5) seems to contrast weekend mornings versus afternoons, the tenth one (the second from the top in the picture) is a contrast between evenings time and weekend afternoons, thus indicating a mobile phone activation in residential areas with respect to its usage during leisure. Finally, the twentysecond basis function (second from the bottom in the picture) is instead a contrast between Saturday and Sunday afternoon. Other elements of the reference basis show a less clear functional pattern, and were thus not used to the scopes of an urban planning interpretation.

Indeed, each of the selected basis functions shown in Figures 5.4 and 5.5 has a clear interpretation in terms of a mobile phone using profile typical of a

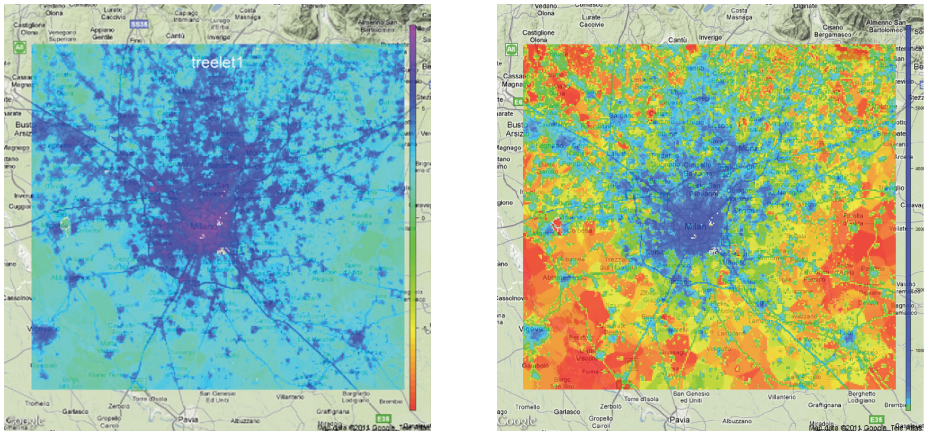

Figure 5.6: Projection of the Telecom dataset on the first reference basis function identified via 1-median basis alignment procedure, with superimposed map of the considered region [source: `http://maps.google.com`]: unprocessed scores map (left panel), and map obtained by applying a 200-mean clustering algorithm to the scores (right panel).

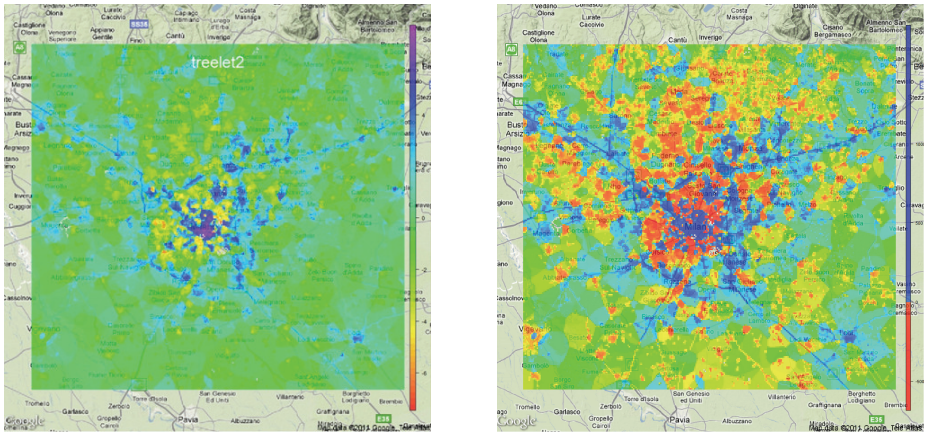

Figure 5.7: Projection of the Telecom dataset on the second reference basis function identified via 1-median basis alignment procedure, with superimposed map of the considered region [source: `http://maps.google.com`]: unprocessed scores map (left panel), and map obtained by applying a 200-mean clustering algorithm to the scores (right panel).

particular aspect of urban life. This fact induced us to analyze separately the scores distribution corresponding to each relevant basis function, i.e. to project raw data on the lattice on the detected reference basis: the result is a set of maps describing the contribution of each basis function to the overall telephonic pattern, in each site of the lattice. Each scores map has been superimposed to a street map of the urban area of Milan obtained via `http://maps.google.com`.

Scores maps are shown in Figures 5.6–5.11 (for brevity, we only report a

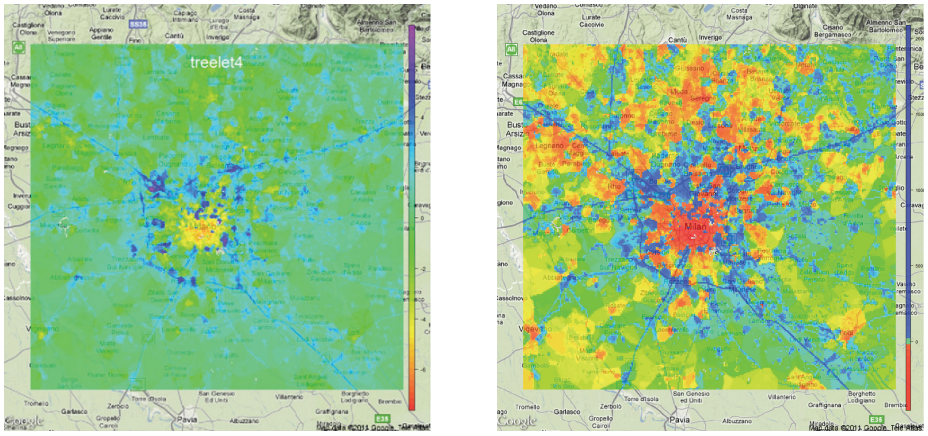

Figure 5.8: Projection of the Telecom dataset on the fourth reference basis function identified via 1-median basis alignment procedure, with superimposed map of the considered region [source: `http://maps.google.com`]: unprocessed scores map (left panel), and map obtained by applying a 200-mean clustering algorithm to the scores (right panel).

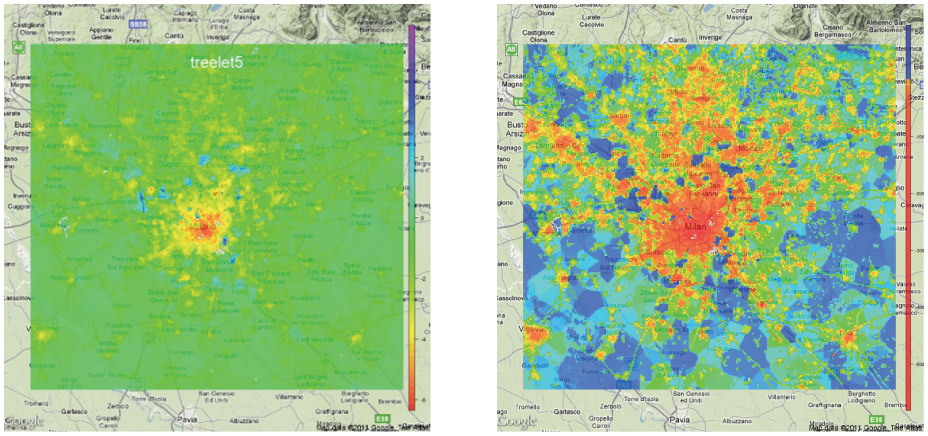

Figure 5.9: Projection of the Telecom dataset on the fifth reference basis function identified via 1-median basis alignment procedure, with superimposed map of the considered region [source: `http://maps.google.com`]: unprocessed scores map (left panel), and map obtained by applying a 200-mean clustering algorithm to the scores (right panel).

selection of the most relevant / interpretable projections). In the left panel of each figure, the scores map in a homogeneous color scale is reported; in the right panel, instead, the same scores map is drawn by selecting the color scale via k-mean with $k = 200$. From inspection of these pictures, we can do some interesting considerations. Looking at Figure 5.6, we note that the highest values in the map identifies the urban structure of the city of Milan (highly populated areas vs poorly populated areas), while in Figure 5.8 (the fourth basis function

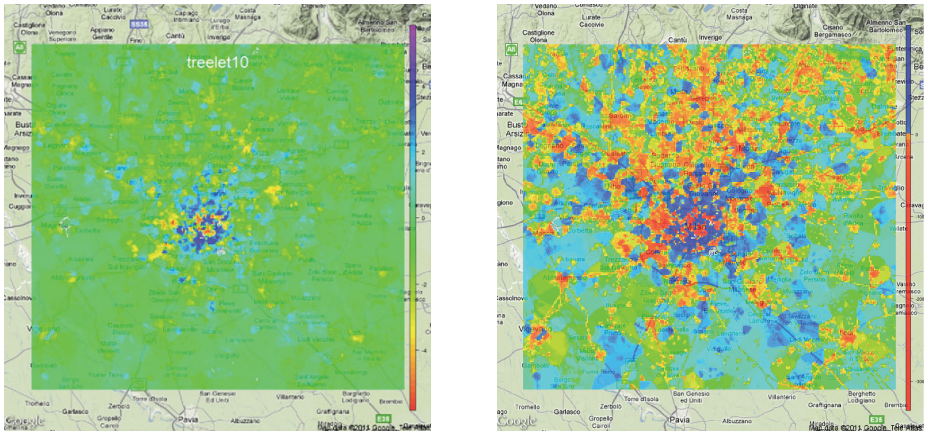

Figure 5.10: Projection of the Telecom dataset on the tenth reference basis function identified via 1-median basis alignment procedure, with superimposed map of the considered region [source: `http://maps.google.com`]: unprocessed scores map (left panel), and map obtained by applying a 200-mean clustering algorithm to the scores (right panel).

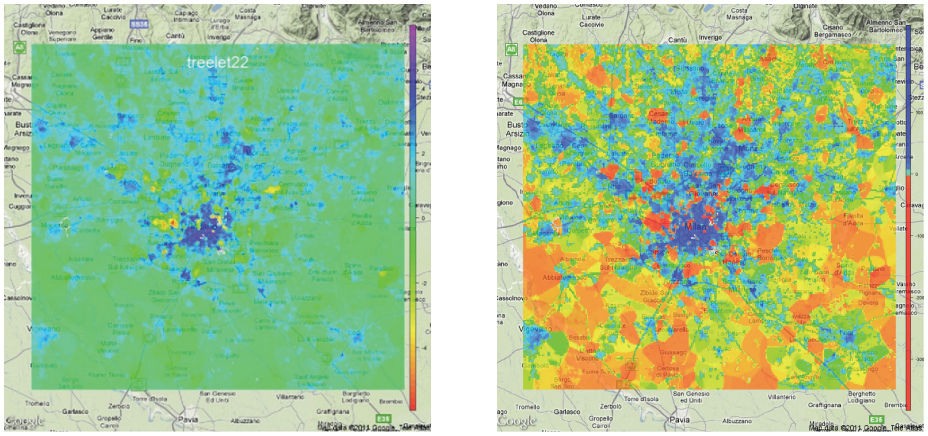

Figure 5.11: Projection of the Telecom dataset on the twenty–second reference basis function identified via 1-median basis alignment procedure, with superimposed map of the considered region [source: `http://maps.google.com`]: unprocessed scores map (left panel), and map obtained by applying a 200-mean clustering algorithm to the scores (right panel).

scores) we can clearly distinguish ring roads and motorways simply by looking at the highest values in the map. Also the projection on the second element of the reference basis (Figure 5.7) shows a neat indication of the motorways and ringroads positions; however, here roads are detected together with the most central zones of the city, almost completely devoted to shopping / business areas. Finally, the map corresponding to data projections on the twentysecond element of the reference basis (contrasting Saturday vs Sunday afternoon), clearly show

a high negative peak of the scores values in the exact localization of San Siro (Meazza) football stadium, and Cernusco sul Naviglio commercial area (opening on Sunday).

5.5.2 Analysis of Milan Urban Area

One of the interesting insights coming from the analysis of Erlang measurements in the metropolitan area of Milan is the description of the region in terms of flows of people. Indeed, four among the most relevant functional patterns describing mobile phone usage in the network are referred to commuters, and in particular the ones referred to morning commuters seem to explain a huge amount of displacements: once data are projected on the fourth or fifth element of the reference basis shown in Figure 5.6, thus obtaining the two scores maps shown in Figure 5.8 and 5.9, the road network of the whole region becomes evident, and the main roads in which the morning traffic is flowing (highways A1, A4 and A7, and the city East, West and North ringroads) are clearly depicted. This fact suggests that, observing the phenomenon on a macroscale (i.e., observing a huge region including Milan, its outskirts and the towns nearby), the most evident contribution to the overall signal is given by a dynamical use of mobile phone, and in particular its use during commuting.

However, one could wonder whether this conclusion is due to the scale of observation of the phenomenon, and whether passing to a microscale of observation would change the conclusion, or at least give different insights. Indeed, some of the reference basis elements shown in Figures 5.6 and 5.7 have a clearly interpretable functional pattern, but at the same time the corresponding scores maps do not show interesting links with the underlying geographical / urban structure.

Hence, we conducted the same analysis on a smaller region, including only the urban area of Milan (city centre and suburbia), and shown in Figure 5.2 (right panel). In this second part of the analysis, we consider the same Erlang measures on a smaller lattice $S_1 \subset S_0$, composed by $N_1 = |S_1| = 2112$ sites, 44 in the W-E direction and 48 in the N-S direction; the site which is now located in the extreme N-W position has geographical coordinates $(45.5094^\circ\ N, 9.1141^\circ\ E)$ (for a total area covered by the lattice of 14.504×9.888 km – W-E $\times$ N-S).

In the analysis of local telephonic time patterns on the smaller lattice S_1 (referred to the urban area of Milan), we leave unchanged all parameters defining the Bagging Voronoi strategy, except for n. The dimension of the Voronoi tessellation for this second application, in fact, has to be carefully chosen: the a–priori idea is keeping fixed (with respect to the previous analysis) the ratio between the tessellation dimension and the total number of sites in the lattice, a choice which would have led to setting $n = 13$; however, this choice is not necessarily the best in terms of interpretability of the final result. Thus, an a–posteriori evaluation of the estimated reference basis for different values of n lead to the final choice $n = 32$, which corresponds to an average number of sites per Voronoi element of 66.

The reference basis detected via Bagging Voronoi strategy with $n = 32$ is shown in Figures 5.12 and 5.13. While observing the two pictures, and com-

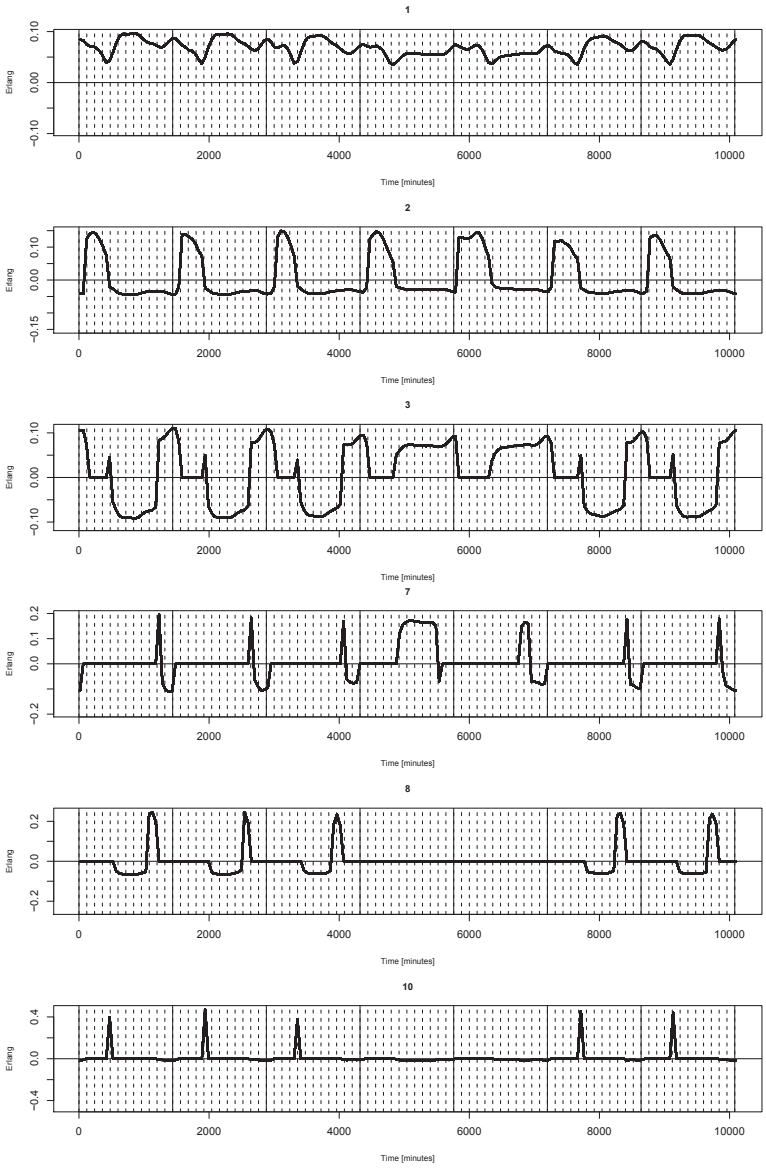

Figure 5.12: Selected basis functions (from top to bottom, first to third, seventh, eighth and tenth function) from the reference basis identified via 1-median basis alignment procedure. The vertical lines correspond to midnight, and the origin is at midnight between tuesday and wednesday. The dashed lines are drawn every two hours.

paring them with the plot of the reference basis detected in the analysis of the metropolitan area (Figures 5.4 and 5.5), we first note that all the elements of the new reference basis are somehow related to one (or more) of the elements of the previous one; nevertheless, there are some interesting differences, suggesting us that the analysis of the urban area makes us able to capture much more detailed

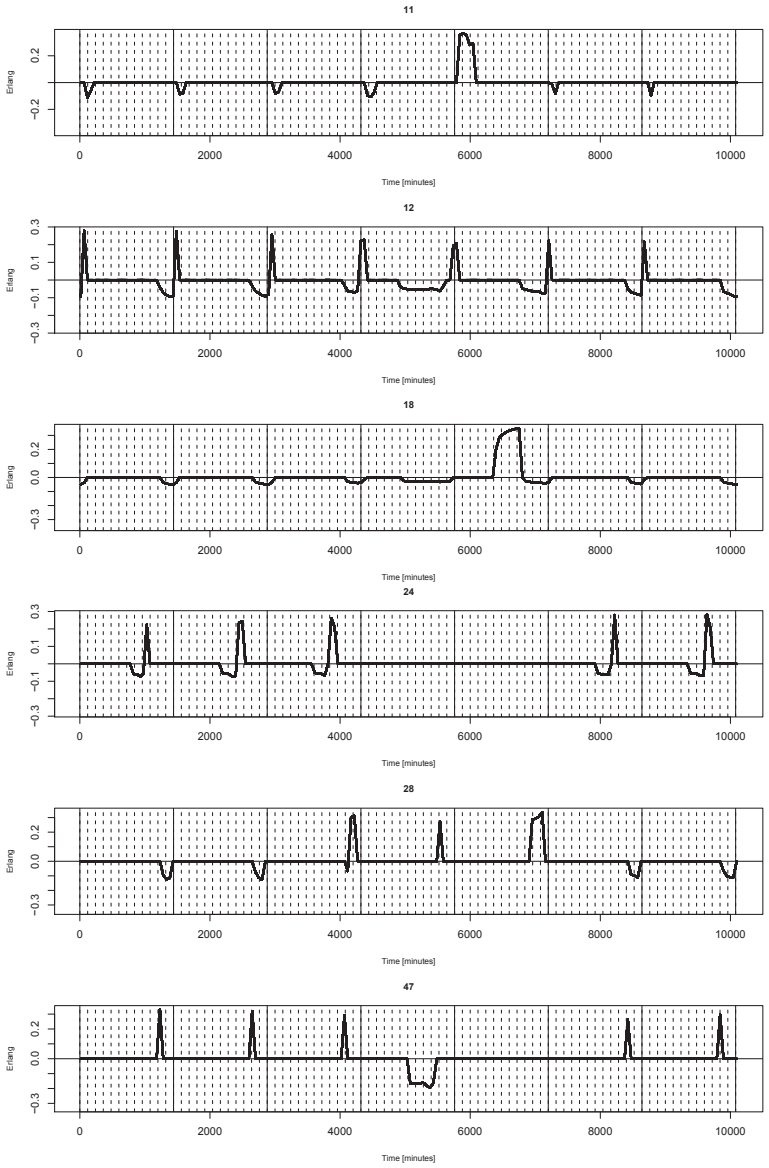

Figure 5.13: Selected basis functions (from top to bottom, functions 11, 12, 18, 24, 28 and 47) from the reference basis identified via 1-median basis alignment procedure. The vertical lines correspond to midnight, and the origin is at midnight between tuesday and wednesday. The dashed lines are drawn every two hours.

information. We will thus discuss the new reference basis in the light of the prevoiusly detected one, in order to point out both common aspects and relevant differences.

The first basis function of the new reference basis, shown in the top panel of Figure 5.12, is describing the average intensity of calls in the network, coherently

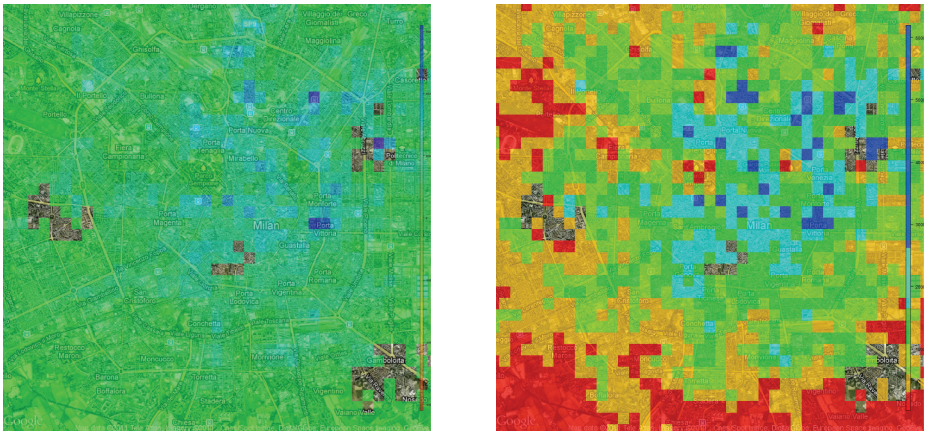

Figure 5.14: Projection of the Telecom reduced dataset on the first reference basis function identified via 1-median basis alignment procedure, with superimposed map of the considered region [source: `http://maps.google.com`]: unprocessed scores map (left panel), and map obtained by applying the Jenks optimization method ($k = 6$) to the scores (right panel).

with the first basis element in the previous analysis; however, the new pattern seems more precise, with higher intensity during the weekdays than in the weekend, and higher intensity in the night than during the day in the weekend. The second and third basis functions in Figure 5.12 are exchanged with respect to their corresponding functions in Figure 5.4, but they are indeed coherent; also differences seem quite interesting: the new second basis function (the old third one) activates during the night, both during the week and in the weekend, and the intensity/duration of the activity during Saturday night seems higher; hence, this basis function could merge both night working activities and nightlife. Concerning the new third basis function (the old second one), it coherently describes working hours (8:30 am till 8 pm) versus non–working hours usage of mobile phone, but within non–working hours activation we also observe the appearance of a small intensity peak in the early morning (7 am till 8:30 am).

The seventh element of the reference basis has little to share with basis functions detected in the previous analysis (it only vaguely resembles the old tenth element): this functional pattern contrasts the dinner time with the evening during weekdays, and the Saturday afternoon / Sunday late afternoon with Sunday evening. In the light of this interpretation, the lack of consistency with the elements of the reference basis for the metropolitan area finds a reason: this functional pattern, which emerges only in a detailed analysis, is describing a typical behaviour of people resident in Milan.

The eighth basis function in the new reference basis is quite coherent with the twentysixth of the old one, being both activated only during the week and both referred to evening commuters. More interestinlgy, morning commuters are described by a single functional pattern in the new reference basis, which is the tenth element (last panel in Figure 5.12), while in the old one two basis functions were devoted to the description of flows of people during the morning. Hence,

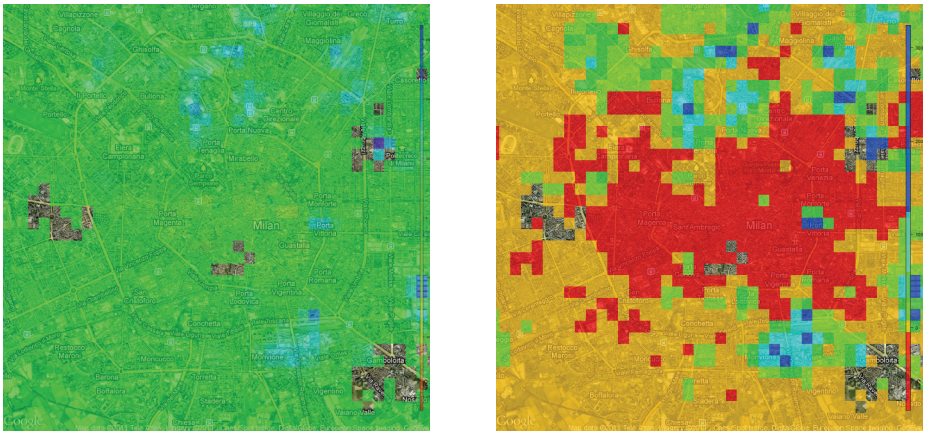

Figure 5.15: Projection of the Telecom reduced dataset on the second reference basis function identified via 1-median basis alignment procedure, with superimposed map of the considered region [source: `http://maps.google.com`]: unprocessed scores map (left panel), and map obtained by applying the Jenks optimization method ($k = 6$) to the scores (right panel).

the importance of the commuters on the overall network intensity seems to be reduced at the urban level.

Other comparisons can be made by looking at other basis elements, shown in Figure 5.13. The twelfth basis element is quite similar to the sixth element of the old reference basis, even if the new one is clearly a contrast between early evening and late evening, thus more clearly associated to nights out. The eighteenth basis function, shown in the third panel of Figure 5.13 from the top, is quite similar to the twentysecond, except for the fact that it is now describing the Sunday afternoon, only. Finally, the twentyfourth basis function and the fourtyseventh one (third and first panels from the bottom of Figure 5.13, respectively) are both related to the old twelfth basis function, but the former is contrasting late evening commuters only with the afternoon of weekdays, while the latter is including also Saturday afternoon together with weekdays.

There are still two basis functions belonging to the reference basis detected in the analysis of the urban area of Milan that have been left apart: we should carefully look at the first and fifth panel from the top of Figure 5.13, since these two basis functions had not been detected when analyzing the metropolitan area. Indeed, when interpreting the functional pattern associated to these two functions, it seems natural to motivate their appearence by the fact they are closely related to urban life, and in particular to nightlife. The eleventh basis function (top panel in Figure 5.13) is contrasting Saturday night with all other nights in the week, while the twentyeighth basis function (fifth panel from the top in Figure 5.13) seems to describe the happy hour time during the weekend (8 pm till 11 pm on Friday, 7 pm till 9 pm on Saturday, and 7 pm till 11 pm on Sunday).

The maps of the projections of raw data from the urban area on the elements of the new reference basis shown in Figures 5.12–5.13 (again for brevity we report

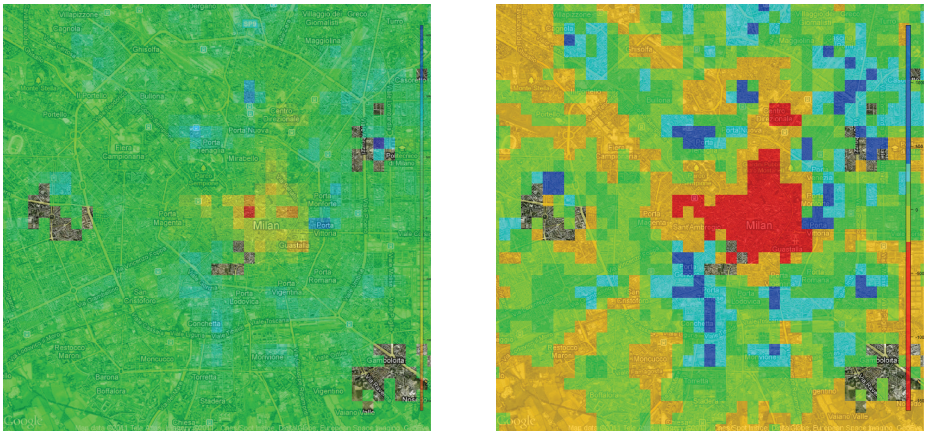

Figure 5.16: Projection of the Telecom reduced dataset on the third reference basis function identified via 1-median basis alignment procedure, with superimposed map of the considered region [source: `http://maps.google.com`]: unprocessed scores map (left panel), and map obtained by applying the Jenks optimization method ($k = 6$) to the scores (right panel).

a selection of the most relevant / interpretable projections), are depicted in Figures 5.14-5.21. In analogy with the scores maps obtained in the analysis of the metropolitan area, and reported in Figures 5.6–5.11, the scores map is reported in a homogeneous color scale in the left panel of each figure. In the right panel, instead, the same scores map is drawn by selecting the color scale via a method named *Jenks optimization method*.

The Jenks optimization method, also called the Jenks natural breaks classification method, is a data classification method designed to determine the best arrangement of values into different classes. This is done by seeking to minimize the average deviation from the class mean in each class, while also maximizing the deviation from the means of the other groups of each class mean. In other words, the method seeks to reduce the variance within classes and maximize the variance between classes. This method is applied to the values in each scores map (without considering their spatial distribution) in order to obtain the best partition of the data in classes; each of the classes determined in this way will be then associated to a single color in the representation of the scores. For the maps in Figures 5.14-5.21, the Jenks method has been applied to select a partition of the colorscale in 6 colors.

From the inspection of the maps in Figures 5.14 and 5.16 one can make the same considerations we have already made for the metropolitan area: the first treelet is detecting the most polutated areas of the city (on average), while the third one is describing the areas of the city mostly populated during the weekdays (working areas) vs residential ones. The inspection of the map in Figure 5.15 is instead interesting, because each area associated to high score values (green or blue coloured areas in the Jenks map) can be interpreted in term of a night activity: either a metro/tram/bus deposit, or a marketplace, or a discotheque. Moreover, from the analysis of this map, a zone in south-east area of Milan

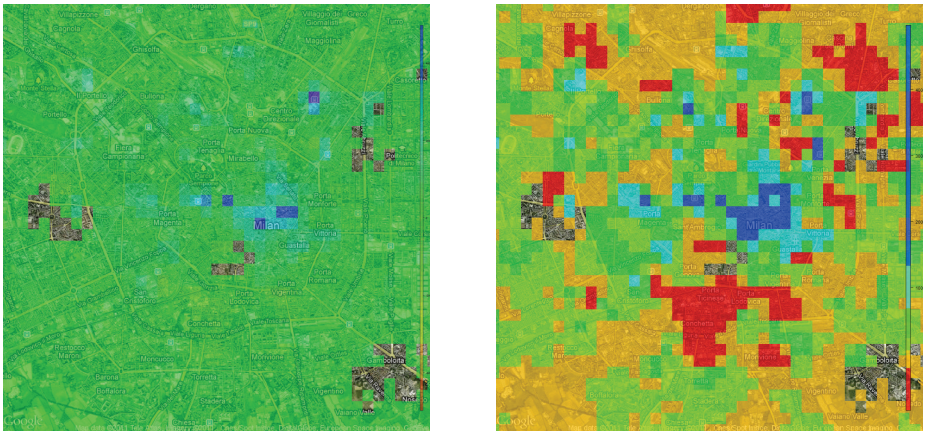

Figure 5.17: Projection of the Telecom reduced dataset on the seventh reference basis function identified via 1-median basis alignment procedure, with superimposed map of the considered region [source: `http://maps.google.com`]: unprocessed scores map (left panel), and map obtained by applying the Jenks optimization method ($k = 6$) to the scores (right panel).

(among viale Ortles, viale Toscana and on both sides of via Ripamonti) emerges for its high score values: this can be either associated to some nightclub, or be the evidence for some activity which is not known in terms of the planning scheme of the city.

The score map corresponding to the seventh treelet for the urban area, shown in Figure 5.17, distinguishes between two typical activities of people living in Milan, carried out during leisure in weekdays: the after-work drink with colleagues, and the after-dinner drink with friends. Many areas known for the presence of pubs or bars are evident in the map, such as Navigli, Porta Ticinese, Colonne, Corso Como, ...

The score maps corresponding to the eighth and tenth treelets are not clearly interpretable, confirming the fact that the movements of commuters are better detected on the metropolitan area with respect to the urban one. Some interpretable places involving commuters' movements at a urban level are indeed detectable: in both maps, the highest score value is in the site where the Central Railway Station is located; moreover, the highly positive values in both maps (represented in light blue) correspond to crossroads among main urban roads, to minor railway stations or to underground stations.

Finally, let us consider the scores maps corresponding to the eighteenth and twentyfourth treelets. The former describes the behaviour of mobile phone users during sunday afternoon, while the latter is again associated to afternoon commuters. Beside the San Siro Stadium, which is detectable a the left border of the eighteenth map for its high score value, the sunday afternoon activities of Telecom mobile phone users seem associated to big market places, parks, and again to the Central Railway Station. Instead, the scores map corresponding to the twentyfourth treelet again indicates the positions of underground and railway stations.

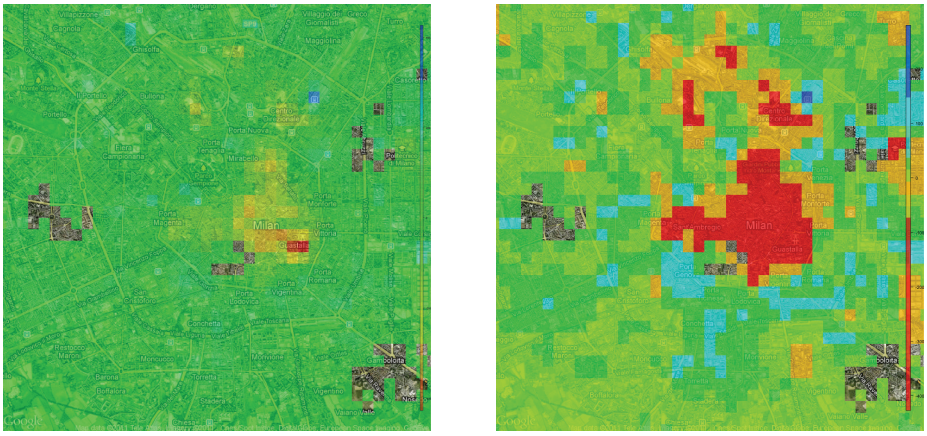

Figure 5.18: Projection of the Telecom reduced dataset on the eighth reference basis function identified via 1-median basis alignment procedure, with superimposed map of the considered region [source: `http://maps.google.com`]: unprocessed scores map (left panel), and map obtained by applying the Jenks optimization method ($k = 6$) to the scores (right panel).

5.5.3 Conclusions

We have conducted an analysis of the Erlang measures both at a metropolitan and at a urban level, and both the analyses seem to give interesting and innovative results. However, undoubtedly the most important aspect of the results is their interpretability in terms of the urban structure of the city, and the new insights some of the results can give, that go beyond the Milan planning scheme.

We clearly see some further developments, for instance given by a more accurate overlap of the scores map to the map of the city: indeed, Google Maps information can be very inaccurate at a urban level. Hence, many results would have been discussed more precisely by considering GIS associated to urban maps to a higher level of precision, given by urban planners.

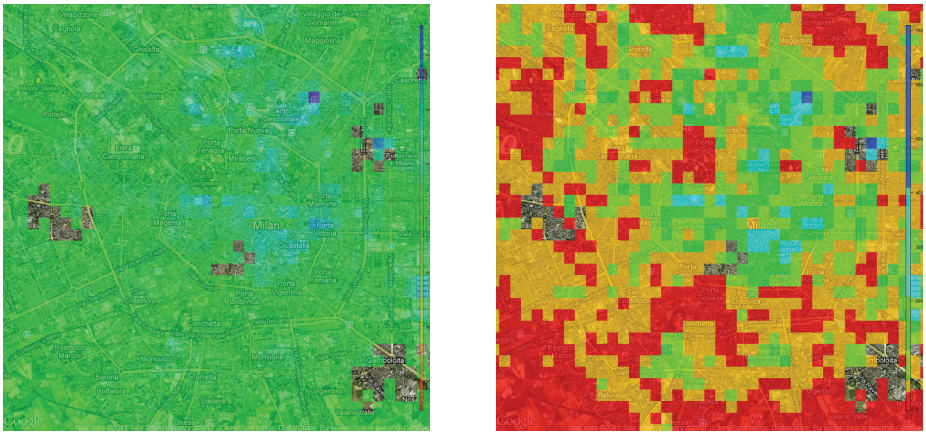

Figure 5.19: Projection of the Telecom reduced dataset on the tenth reference basis function identified via 1-median basis alignment procedure, with superimposed map of the considered region [source: `http://maps.google.com`]: unprocessed scores map (left panel), and map obtained by applying the Jenks optimization method ($k = 6$) to the scores (right panel).

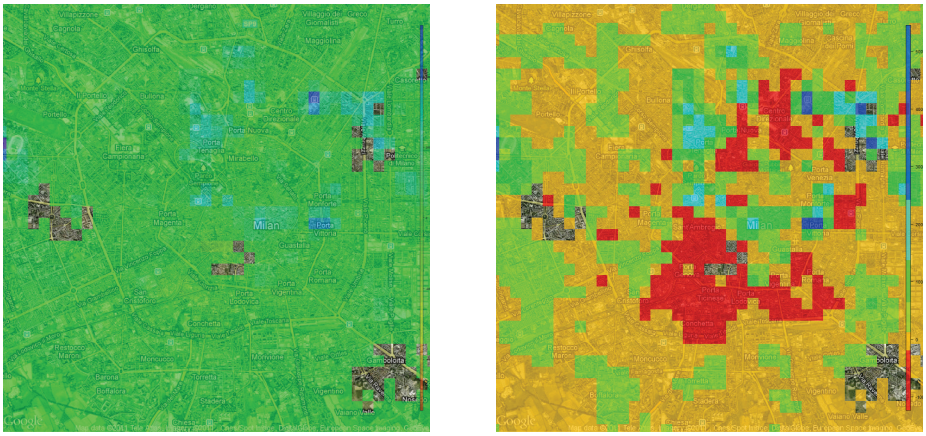

Figure 5.20: Projection of the Telecom reduced dataset on the eighteenth reference basis function identified via 1-median basis alignment procedure, with superimposed map of the considered region [source: `http://maps.google.com`]: unprocessed scores map (left panel), and map obtained by applying the Jenks optimization method ($k = 6$) to the scores (right panel).

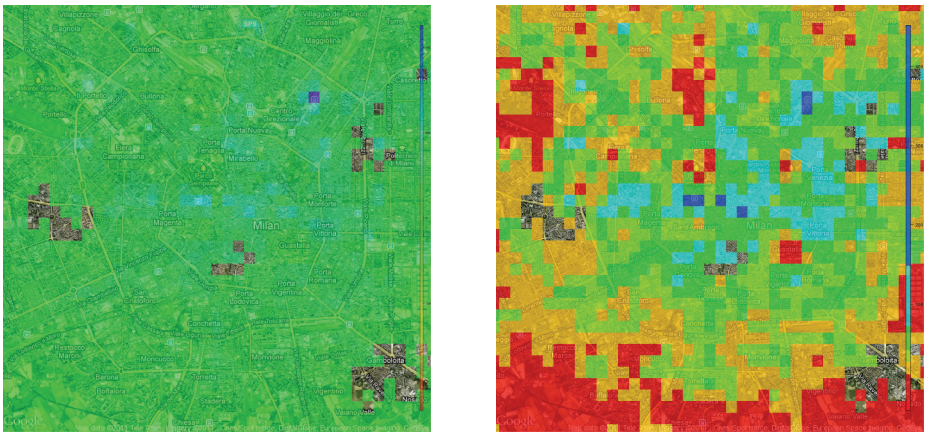

Figure 5.21: Projection of the Telecom reduced dataset on the twentyfourth reference basis function identified via 1-median basis alignment procedure, with superimposed map of the considered region [source: `http://maps.google.com`]: unprocessed scores map (left panel), and map obtained by applying the Jenks optimization method ($k = 6$) to the scores (right panel).

Bibliography

[1] N. S. Altman and J. C. Villarreal. Self-modelling regression for longitudinal data with time-invariant covariates. *Canad. J. Statist.*, 32:251–268, 2004.

[2] L. Anselin. Local indicators of spatial association – lisa. *Geographical Analysis*, 27:93–115, 1995.

[3] L. Anselin, I. Syabri, and O. Smirnov. Visualizing multivariate spatial correlation with dynamically linked windows.

[4] V. Baladandayuthapani, B.K. Mallick, M. Young Hong, J.R. Lupton, N.D. Turner, and R.J. Carroll. Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics*, 64:64–73, 2008.

[5] S. Banerjee, B. Carlin, and A. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, Monographs on Statistics and Applied Probability, 2004.

[6] F. Beltran, B. Sanso, R. T. Lemos, and R. Mendelssohn. Joint projections of north pacific sea surface temperature from different global climate models. Technical Report 03, UCSC, 2011.

[7] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B*, 48(3):259–302, 1986.

[8] B. Blviken. A method for spatially moving correlation analysis. *Norsk Epidemiologi*, 13(2):229–232, 2033.

[9] S. Boudaoud, H. Rix, and O. Meste. Core shape modelling of a set of curves. *Computational Statistics and Data Analysis*, 54:308–325, 2010.

[10] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[11] H. Cardot, F. Ferraty, and P. Sarda. Linear functional model. *Statistics and Probability Letters*, 45:11–22, 1999.

[12] J. M. Chiou and P. L. Li. Functional clustering and identifying substructures of longitudinal data. *J. R. Statist. Soc. Series B*, 69:679–699, 2007.

[13] N. Cressie. *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics, 1993.

[14] J. A. Cuesta-Albertos and R. Fraiman. Impartial trimmed k-means for functional data. *Computational Statistics and Data Analysis*, 51:4864–4877, 2007.

[15] I. Daubechies. Orthonormal basis of compactly supported wavelets. *Communictions on Pure and Applied Mathematics*, 41:909–996, 1988.

[16] P. Delicado, R. Giraldo, C. Comas, and J. Mateu. Statistics for spatial functional data: some recent contributions. *Environmetrics*, 21:224–239, 2010.

[17] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis*. Springer, 2006.

[18] D. Floriello, P. Secchi, and V. Vitelli. Sparse k-means classification of functional data. Technical Report in preparation, MOX - Dipartimento di Matematica, Politecnico di Milano, 2012.

[19] R. Fraiman and G. Muniz. Trimmed means for functional data. *Test*, 10(2):419–440, 2001.

[20] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[21] J.H. Friedman and J.J. Meulman. Clustering objects on a subset of attributes. *Journal of the Royal Statistical Society, Ser. B*, 66:815–849, 2004.

[22] L. A. Garcia-Escudero and A. Gordaliza. A proposal for robust curve clustering. *Journal of Classification*, 22:185–201, 2005.

[23] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[24] D. Gervini and T. Gasser. Self-modelling warping functions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66:959–971, 2004.

[25] A. D. Gordon. *Classification*. Chapman and Hall, 1999.

[26] D. J. Hand. *Construction and Assessment of Classification Rules*. John Wiley & Sons, 1997.

[27] M. S. Handcock and J. R. Wallis. An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *Journal of the American Statistical Association*, 89(426):368–378, 1994.

[28] J. A. Hartigan. Asymptotic distributions for clustering criteria. *The Annals of Statistics*, 6(1):117–131, 1978.

[29] T. Hastie, R. Tibshirani, and J.Friedman. *The Elements of Statistical Learning, first edition*. Springer, 2001.

[30] T. Hastie, R. Tibshirani, and J.Friedman. *The Elements of Statistical Learning, second edition*. Springer, 2008.

[31] S. Hormann and P. Kokoszka. Weakly dependent functional data. *The Annals of Statistics*, 38(3):1845–1884, 2010.

[32] S. Hormann and P. Kokoszka. *Functional Time Series*. Handbook of Statistics: Time Series Analysis-Methods and Applications. (Forthcoming.), 2011.

[33] L. Horvath and P. Kokoszka. *Inference for Functional Data*. Preprint., 2008.

[34] F. Ieva, A. M. Paganoni, D. Pigoli, and V. Vitelli. Multivariate functional clustering for the analysis of ecg curves morphology. Technical Report 04, MOX, Dipartimento di Matematica, Politecnico di Milano, 2011.

[35] G. M. James. Curve alignment by moments. *The Annals of Applied Statistics*, 1:480–501, 2007.

[36] L. Kaufman and P. Rousseeuw. *Finding Groups in Data*. Wiley Series in Probability and Mathematical Statistics, 1990.

[37] D. Kaziska and A. Srivastava. Gait-based human recognition by classification of cyclostationary processes on nonlinear shape manifolds. *Journal of the American Statistical Association*, 102:1114–1128, 2007.

[38] C. Ke and Y. Wang. Semiparametric nonlinear mixed-effects models and their applications. *J. Amer. Statist. Assoc.*, 96:1272–1298, 2001.

[39] R. Kindermann and J. L. Snell. *Markov Random Fields and Their Applications*. American Mathematical Society, 1980.

[40] A. Kneip and T. Gasser. Convergence and consistency results for self-modeling nonlinear regression. *The Annals of Statistics*, 16(1):82–112, 1988.

[41] A. Kneip, X. Li, K.B. MacGibbon, and J.O. Ramsay. Curve registration by local regression. *The Canadian Journal of Statistics*, 28(1):19–29, 2000.

[42] H. Kunsch, S. Geman, and A. Kehagias. Hidden markov random fields. *The Annals of Applied Probability*, 5(3):577–602, 1995.

[43] W. H. Lawton, E. A. Sylvestre, and M. S. Maggio. Self modeling nonlinear regression. *Technometrics*, 14:513–532, 1972.

[44] A. B. Lee, B. Nadler, and L. Wasserman. Treelets – an adaptive multi-scale basis for sparse unordered data. *The Annals of Applied Statistics*, 2(2):435–471, 2008.

[45] C. K. Leung and F. K. Lang. Maximum a posteriori spatial probability segmentation. *IEE Proceedings - Vision, Image and Signal Processing*, 144(3):161–167, 1997.

[46] M. J. Lindstrom and D. M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46:673–687, 1990.

[47] A. E. Linsday. *ECG learning centre*. University of Utah School of Medicine, Salt Lake City, Utah, 2006.

[48] X. Liu and H. G. Muller. Modes and clustering for time-warped gene expression profile data. *Bioinformatics*, 19(15):1937–1944, 2003.

[49] X. Liu and M. C. K. Yang. Simultaneous curve registration and clustering for functional data. *Computational Statistics and Data Analysis*, 53:1361–1376, 2009.

[50] S. Lopez-Pintado and J. Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734, 2009.

[51] R. Luss and A. d'Aspremont. Clustering and feature selection using sparse principal component analysis. *Optimization and Engineering*, 11(1):145–157, 2010.

[52] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics, 2000.

[53] J. Mller. *Lectures on Random Voronoi Tessellations*. Springer, 1994.

[54] D. C. Montgomery, G. C. Runger, and N. F. Hubele. *Engineering Statistics*. Wiley, 2007.

[55] NASA. Surface meteorology and solar energy. *A renewable energy resource web site (release 6.0)*, `http://eosweb.larc.nasa.gov/cgi-bin/sse/sse.cgi?#s01`, [accessed on the 25^{th} of November, 2010].

[56] M. D. Penrose. Laws of large numbers in stochastic geometry with statistical applications. *Bernoulli*, 13(4):1124–1150, 2007.

[57] D. Pigoli and L.M. Sangalli. Wavelets in functional data analysis: estimation of multidimensional curves and their derivatives. *Computational Statistics and Data Analysis (to appear)*, 2011.

[58] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[59] J. O. Ramsay and X. Li. Curve registration. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60:351–363, 1998.

[60] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis*. Springer, 2002.

[61] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, 2005.

[62] J. O. Ramsay and H. Wickham. fda: Functional data analysis. *r package version 1.1.8*, 2007.

[63] C. Richter, S. Teske, and J. A. Nebrera. Concentrating solar power global outlook 09. Technical report, Greenpeace International / European Solar Thermal Electricity Association (ESTELA) / IEA SolarPACES, 2009.

[64] L. M. Sangalli, P. Secchi, and S. Vantini. Explorative functional data analisys for 3d-geometries of the inner carotid artery. In S. Dabo-Niang and F. Ferraty, editors, *Functional and Operatorial Statistics*, pages 289–296. Springer, Contributions to Statistics, 2008.

[65] L. M. Sangalli, P. Secchi, S. Vantini, and A.Veneziani. A case study in exploratory functional data analysis: Geometrical features of the internal carotid artery. *J. Amer. Statist. Assoc.*, 104:37–48, 2009.

[66] L. M. Sangalli, P. Secchi, S. Vantini, and A.Veneziani. Efficient estimation of three-dimensional curves and their derivatives by free-knot regression splines, applied to the analysis of inner carotid artery centrelines. *J. R. Stat. Soc. Ser. C, Applied Statistics*, 58:285–306, 2009.

[67] L. M. Sangalli, P. Secchi, S. Vantini, and V. Vitelli. Functional clustering and alignment methods with applications. *Communications in Applied and Industrial Mathematics*, 1(1):205–224, 2010.

[68] L. M. Sangalli, P. Secchi, S. Vantini, and V. Vitelli. K-mean alignment for curve clustering. *Computational Statistics and Data Analysis*, 54:1219–1233, 2010.

[69] A. M. Scher and A. C. Young. Ventricular depolarization and the genesis of the qrs. *Annals of New York Academy of Science*, 65:768–778, 1957.

[70] P. Secchi, S. Vantini, and V. Vitelli. Algoritmi statistici di classificazione per dati funzionali spazialmente correlati. *Deliverable VIII (finale) del Progetto di Ricerca: "Analisi multidisciplinare di dati telerilevati con sensori ottici per l'individuazione di fenomeni di microseepage d'idrocarburi riconducibili alla presenza di giacimenti profondi", Cap. 5*, pages 331–414, 2010.

[71] P. Secchi, S. Vantini, and V. Vitelli. Bagging voronoi classifiers for clustering spatial functional data. Technical Report 26, MOX, Dipartimento di Matematica, Politecnico di Milano, 2011.

[72] P. Secchi, S. Vantini, and V. Vitelli. Spazializzazione di informazioni geografiche puntuali: ricostruzione di un campo spaziale a partire da misurazioni puntuali. *Deliverable Finale dell'Attivit addendum al Progetto di Ricerca: "Validazione di metodologie di analisi multidisciplinare di dati telerilevati con sensori ottici (MIVIS), finalizzate all'individuazione di fenomeni di microseepage d'idrocarburi riconducibili alla presenza di giacimenti profondi"*, pages 1–???, 2011.

[73] P. Secchi and V. Vitelli. Algoritmi statistici di classificazione per dati funzionali spazialmente correlati: sviluppo di metodologie, validazione su dati sintetici, applicazioni. *Deliverable II-III-V-VII del Progetto di*

Ricerca: "Analisi multidisciplinare di dati telerilevati con sensori ottici per l'individuazione di fenomeni di microseepage d'idrocarburi riconducibili alla presenza di giacimenti profondi", Cap. 6, pages 381–428, 2009.

[74] M. Shaked and J. G. Shanthikumar. *Stochastic Orders*. Springer Series in Statistics, 2010.

[75] J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer Series in Statistics, 1995.

[76] A. Sheehy, T. Gasser, L. Molinari, and R. H. Largo. An analysis of variance of the pubertal and midgrowth spurts for length and width. *Annals of Human Biology*, 26:309–331, 2000.

[77] A. Sheehy, T. Gasser, L. Molinari, and R. H. Largo. Contribution of growth phases to adult size. *Annals of Human Biology*, 27:281–298, 2000.

[78] N. Shimizu and M. Mizuta. Functional clustering and functional principal points. *LNAI*, 4693:501–508, 2007.

[79] A. Struyf, M. Hubert, and P. Rousseeuw. Clustering in an object-oriented environment. *Journal of Statistical Software*, 1(4):1–30, 1997.

[80] T. Tarpey and K. K. J. Kinateder. Clustering functional data. *Journal of Classification*, 20:93–114, 2003.

[81] D. Telesca and L. Inoue. Bayesian hierarchical curve registration. *J. Amer. Statist. Assoc.*, 103(481):328–339, 2008.

[82] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58:267–288, 1996.

[83] R. Tibshirani and D. Witten. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

[84] S. Tokushige, H. Yadohisa, and K. Inada. Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Computational Statistics*, 22:1–16, 2007.

[85] R. D. Tuddenham and M. M. Snyder. Physical growth of california boys and girls from birth to age 18. Technical Report 1, University of California Publications in Child Development, 1954.

[86] F. Y. Wu. The potts model. *Reviews of Modern Physics*, 54:235–268, 1982.

[87] Y. Zhang, S. Smith, and M. Brady. Hidden markov random field model and segmentation of brain mr images. *IEEE Transactions of Medical Imaging*, 20:45–57, 2001.