

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des Sciences
Institut de Recherche Multidisciplinaire pour la Modélisation et l'Analyse Quantitative
Institut de Statistique, Biostatistique et Sciences Actuarielles



Investigating functional data with sharp local features with applications to spectroscopy

Thèse présentée par CATHERINE TIMMERMANS
en vue de l'obtention du grade de
docteur en Sciences,
orientation Statistique, Biostatistique et Sciences Actuarielles.

Promoteur :

R. VON SACHS

Membres du jury :

V. DELOUILLE

P. FRYZLEWICZ

I. VAN KEILEGOM

M. VERLEYSEN

Louvain-la-Neuve

2012

Acknowledgments

I warmly thank everyone who contributed to the completion of this thesis!

The thesis benefited from fruitful collaborations with my supervisor Rainer von Sachs, as well as Laurent Delsol and Piotr Fryzlewicz. Thank you to each of them for their encouragements, feedbacks and useful discussions!

Véronique Delouille, Piotr Fryzlewicz, Ingrid Van Keilegom and Michel Verleysen have immediately accepted the role of members of the Jury for the thesis. I thank them for their interest in this work.

I prepared this thesis under the position of teaching assistant at Institute of Statistics, Biostatistics and Actuarial Sciences of Université catholique de Louvain. The Institute is a very friendly and supporting working environment. Thank you to each of its members!

Along the thesis, I had the opportunity to present my work at several national and international conferences and I benefited from several interesting stays in foreign research institutes: Sacramento Peak National Solar Observatory, Université d'Orléans & CNRS, International Space Science Institute, London School of Economics. Those stays were made possible respectively in the framework of the University of Arizona and National Solar Observatory joint Solar Physics Summer School and thanks to the invitations of Laurent Delsol, Matthieu Kretzschmar, Véronique Delouille and Piotr Fryzlewicz. Thank you to each of them, as well as to Rainer von Sachs who always supported those projects! More specifically, I would like to thank Piotr Fryzlewicz for his hospitality during my stay in London.

Those international activities were funded by the Fond National de la Recherche Scientifique, by University of Arizona and National Solar Observatory, by the International Space Science Institute and by the Institute of Statistics, Biostatistics and Actuarial Sciences. I gratefully acknowledge these financial supports. Moreover, financial support from the IAP research network grant P 06/03 of the Belgian Government (Belgian Science Policy) is also acknowledged.

Contents

Introduction	9
1 Bases Giving Distances	17
1.1 The Best Suited Unbalanced Haar Wavelet Basis	19
1.1.1 The BUUHWB Algorithm	20
1.1.2 Some Properties of the BSUHWB.	24
1.2 The BAGIDIS Semi-Distance	27
1.2.1 Definition	28
1.2.2 Properties	29
1.2.3 Discussion	34
1.3 Statistical Tests and Tools	48
1.3.1 Descriptive Statistics in the $b - d$ Plane	48
1.3.2 An ANOVA Test Within the BAGIDIS Framework	49
1.3.3 A Test for the Relative Positions of two Groups in Space.	50
1.4 Data Analyses Using Bagidis	50
1.4.1 A Simulated Example of Supervised Classification	51
1.4.2 A Simulated Example With Sparse Data	53
1.4.3 A Simulated Example of ANOVA F-test	54
1.4.4 Searching for Factors Effects on a Set of Spectrometric Curves	57

1.4.5	Clustering Time Series According to their Dynamics	63
1.5	Conclusion	66
2	Functional Prediction	69
2.1	Statistical framework	71
2.1.1	Functional models for prediction	71
2.1.2	Estimation in a nonparametric functional prediction model . .	74
2.1.3	BAGIDIS	76
2.2	The main results	79
2.2.1	Rate of convergence	79
2.2.2	Asymptotic optimality of a cross-validated choice	85
2.3	Applications and discussions	92
2.3.1	A systematic simulated study	93
2.3.2	Analysis of a real spectroscopic dataset	103
2.4	Conclusion	108
3	Spectroscopic case study	111
3.1	The AMD dataset	113
3.2	Statistical Analyses	115
3.2.1	Visual analysis	115
3.2.2	AMD detection	116
3.2.3	Search for biomarkers	117
3.3	Conclusion	122
4	BAGIDIS and images	123
4.1	The Transform	126
4.1.1	Core ideas	127
4.1.2	The G-BUHWHT Algorithm	129
4.1.3	Mathematical and Visual Properties	134

4.2	A BAGIDIS Semi-distance Between Images	143
4.2.1	Definition	144
4.2.2	Discussion	145
4.2.3	Methodology	147
4.3	Review of the Literature	149
4.4	Applications	154
4.4.1	The Multizone Noisy Image	154
4.4.2	The Moving Pixel	155
4.4.3	Handwritten Digits	160
4.5	Conclusion	164
	Conclusion	167
	Summary of main notations	177
	Bibliography	181
	Appendices	188
A	Technical details for Chapter 1	189
B	Proofs of the main results from Chapter 2	191
B.1	Proof of Theorem 1	191
B.2	Proof of Theorem 2	194
C	Some proofs and procedures used in Chapter 4	227
C.1	Proof of Property 1	227
C.2	Inversion of the G-BUHWI	230
C.3	Proof of Property 4.	231

Introduction

This thesis aims to tackle some common but challenging issues in ^1H NMR spectroscopic studies: investigating differences between groups of spectra in a dataset, determining a statistical model for the prediction of a measured quantity or a group membership associated to a spectrum, and identifying the zones of the spectra that carry significant information for the discrimination. Statistically, this requires the study of curves with sharp local features, those features being the peaks associated to given resonance frequencies in the spectra, and of which the intensity reflects the concentration of given chemical compounds.

As I see it, a challenge in this problem is to define an efficient measure of the dissimilarity between the spectra. Indeed, a given peak in a set of spectra might be affected by vertical amplification, horizontal shift or both simultaneously, the source of those variations possibly being a significant chemical difference or resulting from noise. However, commonly used dissimilarity measures do not return coherent results as soon as there is an horizontal component of variation from one spectrum to another. This thesis proposes therefore a new dissimilarity measure which has the ability to capture both horizontal and vertical variations of the peaks in datasets of spectra, in a unified framework. This dissimilarity measure has been called BAGIDIS for *BAses Giving DIStances*.

*

The above-mentioned approach of the statistical analysis of spectra deserves some comments. There are several possible methodological pathways to get a desired output from a dataset of curves. I think a very powerful approach is to decompose the problem into two steps. The first step consists in computing a dissimilarity matrix between all of the curves in the dataset. The second step is a processing of the dissimilarity matrix that leads to the output. From such a point of view, an analysis requires

two preliminary choices. First, one has to decide for the dissimilarity measure that is going to be used. This dissimilarity measure has to be suitably chosen in order to be able to extract the relevant features of the investigated curves. It thus depends on the nature of the data. Second, an algorithm is to be chosen for processing the dissimilarity matrix. This algorithm must be selected according to the nature of the desired output. Figure 1 summarizes this approach.

The present research concerns dissimilarity measures. It proposes a new semi-distance and investigates its properties. Thereby, it focuses on Step 1 of the methodology. As far as Step 2 is concerned, the thesis restricts itself to a small set of dissimilarity-based algorithms:

The Ward hierarchic clustering. It is an iterative clustering technique where N individuals are progressively grouped together, starting from N groups of 1 individual and leading to a unique cluster with N individuals. At each iteration, the grouping is determined so as to minimize the intern inertia of each group and to maximize the inertia between the groups. Intern inertia is defined as the mean of the square distances of the individuals to their group center of gravity. Inertia between groups is defined as the weighted mean of the square distances of the centers of gravity of each group to the global center of gravity of the dataset. A reference on this technique is Lebart et al. (2004). It results in a *dendrogram* allowing for a visual analysis of the dataset.

The multidimensional scaling. It is a linear projection technique that aims at preserving the distances between the individuals in a dataset. It results in a *map* in which the proximities of the projected individuals can be interpreted - up to a certain degree - as “real” proximities according to the chosen distance. A reference on this technique is Kaufman and Rousseeuw (1990). The well-known *principal component analysis* is a multidimensional scaling coupled with the Euclidean distance.

The nonparametric functional regression estimator. It is a distance-based extension of the Nadaraya-Watson kernel estimator. It provides with a *statistical model* for the *prediction* of a scalar value from a curve or for the *discrimination* of several groups of curves. A reference is Ferraty and Vieu (2006). The k -nearest neighbor estimator (Hastie et al. 2001) might be seen as a particular case of the kernel estimator, with a kernel valued 1 on its support $[0; 1]$ and a suitably defined local bandwidth.

Some distance-based tests for differences between groups of curves. Several tests are defined in Chapter 1.

Because of the decoupling of the definition of the dissimilarity measure and the choice of the processing algorithm, the BAGIDIS semi-distance can be used in a large range of purposes: visual analysis, statistical modeling for prediction, tests for differences between groups. From this point of view, the two steps approach has a very good “value-for-money” ratio as lots of analyses can be conducted on a dataset as soon as one has found a dissimilarity measure that is adapted to the nature of the data. On the other hand, the restriction to this two-steps approach marks the limits of this research.

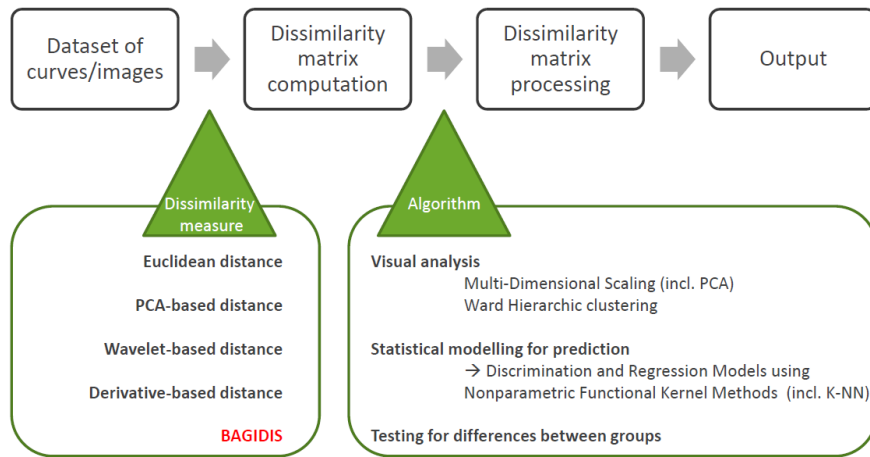


Figure 1: An overview of the methodological approach adopted in this thesis. Some examples of dissimilarity measures are provided as well as some examples of dissimilarity based algorithms.

*

The thesis is organized as follows.

Chapter 1 introduces the BAGIDIS dissimilarity measure and shows it to be a semi-distance. It explains how it originates from a latest generation wavelet-based algorithm by Fryzlewicz (2007). It discusses the interpretation, the use and the limits of the method. It also develops some statistical tests and tools that might be used together with BAGIDIS. It finally assesses the performances of the semi-distance with respect to competing ones, on several simulated and real datasets.

This chapter is really the methodological core of the thesis.

Chapter 2 is then concerned with the use of the BAGIDIS semi-distance in the framework of nonparametric functional data analyses. This field of statistical analyses is currently centered on a nonparametric functional regression estimator proposed by Ferraty and Vieu (2006). The estimator has a very general scope and a sound theoretical support. However, it has to be used in association with a semi-distance, and the predictive qualities of the estimated regression model strongly depend on the feature extraction capacities of the latter. This chapter therefore investigates the association of BAGIDIS with this estimator when predicting from curves with sharp local features, and shows that the association is powerful. It gives rates of convergence of the regression estimator in this setting, as a function of the parameters of the semi-distance. It proposes to optimize those parameters using a cross-validation procedure and shows the asymptotic optimality of the selected vector, this last result having got a larger scope than the BAGIDIS framework. It finally assesses the performances of the method against commonly used competitors on several examples.

Chapter 3 is a case study. It concerns the spectrometric study of blood serum samples in the framework of a pharmaceutical study about an eye disease, which is currently conducted at *Université de Liège*. A visual analysis of the data and a statistical model for detecting the illness are proposed, as well as a search for significantly discriminative compounds in the blood of ill and healthy patients. The advantages of BAGIDIS are discussed and compared with a recent statistical analysis of the same database by Rousseau (2011). This chapter shows how BAGIDIS can be used as a useful complement to the statistical tools that are commonly used in this field.

Chapter 4 provides an extension of the BAGIDIS methodology to image data, and more generally to multidimensional fields of data. The main requirement therefore is to propose an extension of the wavelet algorithm of Fryzlewicz (2007) so as to deal with image data. This is thus the central part of this chapter, which then discusses how BAGIDIS itself might be extended to image datasets.

Those four chapters are research papers by themselves, that have been submitted or are about to be submitted to peer-reviewed journals, although possibly in a slightly different version. Therefore they have been written so as to be self-contained, although several cross-references encourage the reader to switch from one chapter to another.

This paper-based organization might lead to some redundancies in the thesis, as Chapter 2 provides with a survey of the method detailed in Chapter 1. Conversely, it has the advantage to clearly draw the attention to the scope of the successive research projects.

Those papers share co-authorship with several researchers: my supervisor Rainer von Sachs (Chapter 1, 2 and 3), Laurent Delsol from Université d'Orléans and CNRS (Chapter 3), Piotr Fryzlewicz from London School of Economics (Chapter 4), Réjane Rousseau who is an alumni researcher from Université catholique de Louvain (Chapter 3) as well as Pascal de Tullio, Michel Frédérich and Vincent Lambert from Université de Liège (Chapter 3).

*

Although the Conclusion of the thesis includes a summary of the original contributions of this research, I already highlight some major points.

First, this thesis proposes a practical tool that is ready to be used in research projects involving the analysis of curves with sharp local patterns. Thereby, the thesis really has **an applied scope**. It is worth emphasizing the following elements.

BAGIDIS tackles a current practical need. The BAGIDIS methodology fills a gap amongst the available statistical tools for spectroscopic data analyses, as common dissimilarity measures often fail to efficiently deal with curves with possibly misaligned sharp local patterns.

BAGIDIS has a large scope of application. Although motivated by the study of spectroscopic data, BAGIDIS is not specifically designed for this field. It can be applied for analyzing any dataset of curves of which some sharp local patterns might have an horizontal component of variation. Time series analysis is an example outside the spectroscopic field. Moreover, Chapter 4 extends the methodology towards image datasets analyses, which also opens large perspectives of possible applications. For spectroscopic purposes, this extension to image data leads to the prospect of analyzing 2D-spectra. This encounters a currently leading research trend.

BAGIDIS contributes to current researches in pharmaceutical sciences. Chapter 3 consists in the analysis of a spectroscopic database that has been collected in the

framework of a large research project about age-related macular degeneration, an ocular disease that is a leading cause of vision loss in western countries. This project is conducted by a team at Université de Liège. BAGIDIS proves being a useful complement to recent statistical analyses of the same database. It identifies some biochemical pathways that could be implied in the anthology of age-related macular degeneration, and contributes to determining further research steps. It also opens perspectives for an automated tool for diagnosis assistance.

Second, the thesis has a **sound theoretical background**, as it fully takes into account three cutting-edge statistical concepts.

The nature of functional data. A functional data has the form of a vectorial quantity encoding the discretized measurements of a curve or a surface. This is the case of time series and spectra, for instance. In this framework, the suitable information unit is the underlying curve and statistical analyses are worth taking this into account. This point has driven lots of recent statistical research. One may mention Ramsay and Silverman (2005), Cai and Hall (2006), Crambes et al. (2008) and Ferraty (2011) for instance. Statistically, the nature of functional data implies the need to explicitly consider the notion of neighborhood between the discretized points of measurement. BAGIDIS does it.

The nonparametric functional regression techniques. Nonparametric regression modeling overcomes the need to make an hypothesis on the form of the regression operator. Only smoothness hypotheses are required. As far as functional data are concerned, a leading breakthrough has been provided by Ferraty and Vieu (2006). The theoretical work of Chapter 2 places itself in the statistical framework they developed.

Unbalanced Haar wavelets. Unbalanced Haar wavelet bases have been introduced by Girardi and Sweldens (1997) as a way to circumvent the dyadic restriction of classical wavelets. They were used in the context of nonparametric regression with stochastic design by Delouille et al. (2001). They have then been brought forward again by Fryzlewicz (2007) in the context of nonparametric piecewise constant estimation with equispaced design. The *Bottom-Up Unbalanced Haar Wavelet Transform* he defined in the latter context is the starting point of this work.

Last but not least, a major contribution of this thesis is to provide **a new paradigm for curves comparison**. Most dissimilarity measures compare curves through their expansion in a given basis. The basis might be chosen according to the features of the investigated dataset. On the contrary, BAGIDIS proposes to expand each curve in a *different* basis, which is chosen for each curve separately according to its particular features. The thesis shows the potential of comparing curves through their expansions in bases that are different from one curve to another but individually more efficient. This point of view is a real originality. It is therefore the origin of the name of the semi-distance, as BAGIDIS is an acronym for *BAses GIving DIStances*.

Chapter 1

Bases Giving Distances

A functional wavelet-based semi-distance is defined for comparing curves with misaligned sharp local patterns. It is data-driven and highly adaptive to the curves. A main originality is that each curve is expanded in its own wavelet basis, which hierarchically encodes the patterns of the curve. The key to success is that variations of the patterns along the abscissa and ordinate axes are taken into account in a unified framework. Associated statistical tools are proposed for detecting and localizing differences between groups of curves. This methodology is applied to ^1H NMR spectrometric curves and solar irradiance time series.

Keywords: distance, functional data, wavelet, misalignment, spectrometry, time series.

*

BAGIDIS: A New Method for Statistical Analysis of Differences Between Curves with Sharp Local Patterns

C. Timmermans and R.von Sachs

The BAGIDIS methodology is a three-steps procedure for investigating datasets of curves with sharp local features. In this work, we assume that each of the curves of the dataset is initially encoded as a series of N discrete regularly-spaced measurements of a real quantity. The method is centered on the definition of a new functional, data-driven and highly adaptive way for measuring dissimilarities between curves. It

is based upon the expansion of each curve in a different wavelet basis, one that is particularly suited to the curve. This is a major originality. We take into account the differences between the projections of the series onto the bases, as usual, but also the differences between the bases. Therefore, the name of the method stands for *BAses Giving DIStances*. Here is an outline.

- As a first step of the procedure, each series is expanded in an individual, series-adapted wavelet basis: the *Best Suited Unbalanced Haar Wavelet Basis*. We highlight the ability of this basis for a hierarchical encoding of the sharp patterns of the curves. The explicit identification of this notion of hierarchy is one of the contributions of this work, and the observation that the hierarchy makes the resulting bases comparable to each other, although different, is the starting point of the method. Section 1.1 presents the *Best Suited Unbalanced Wavelet Basis*, along with its properties.
- As a second step of the procedure, a dissimilarity measure is proposed for comparing the series through their expansions in their *Best Suited Unbalanced Haar Wavelet Bases*, which are different from one series to another but comparable due to the hierarchy. This dissimilarity measure is shown to be a *semi-distance*. The BAGIDIS semi-distance has the specificity to take into account variations of the sharp patterns observed in the series along the vertical and the horizontal axis in a unified framework. This makes the method particularly powerful for datasets of curves with sharp local features (peaks, abrupt level changes) that can be affected by horizontal shifts, changes in their magnitudes and shape variations (enlargement of a peak, change in its symmetry, or even modification or suppression of a pattern). This property is a major strength of the BAGIDIS methodology, as classical ways for comparing curves usually fail as soon as the sharp patterns in a dataset have a horizontal component of variation. An additional advantage of our methodology is that the BAGIDIS semi-distance itself can be used to get an insight on whether the series are affected by horizontal or vertical variations, or both. Section 1.2 defines the BAGIDIS semi-distance and discusses its features, range of application, parametrization and extensions.
- As a third step of the procedure, the BAGIDIS semi-distance is used for statistically investigating datasets. To this aim, we first define new *descriptive statistics*. They embed the capacity of BAGIDIS to take into account horizontally- and vertically varying sharp local features. We also propose *statistical tests* for diagnosing whether groups of curves do actually differ and how. An ANOVA-like F-test is proposed, which relies on the particular features of the BAGIDIS semi-distance. Another test for differences between groups of curves is applicable for any measure of dissimilarity. Finally, as far as *dataset visualisation* or *prediction models* are concerned, the BAGIDIS semi-distance can easily fit within

any dissimilarity-based method. For instance, Ward’s hierarchic agglomerative algorithm (Kaufman and Rousseeuw 1990; Lebart et al. 2004, e.g.) and multidimensional scaling representations (Cailliez 1983; Cox and Cox 2008, e.g.) are used in this study, while nonparametric functional kernel regression is investigated in Chapter 2. Section 1.3 is devoted to the new statistical tests and tools we propose to use within the BAGIDIS framework.

The performances of our methodology are assessed against classical competitors on several simulated and real data examples. This is the purpose of Section 1.4. It shows the superiority of our method as soon as variations of the significant patterns in the curves have an horizontal component. Moreover, when the patterns are well aligned and no horizontal difference has to be taken into account, the BAGIDIS semi-distance adapts itself and, although not better, the results we obtain are consistent compared to classical competitors, which is desirable. The real data examples consists in ^1H NMR spectrometric curves and in solar irradiance time series.

The original contribution of this work is threefold:

- First, we identify a new, promising, notion of hierarchy in the description of curves using wavelets. This notion emerges from the *Bottom-Up Unbalanced Haar Wavelet Transform* (Fryzlewicz 2007).
- Second, we highlight a new paradigm for comparing curves. Usual methods compare curves expanded in a common basis, which might be chosen according to the features of the given dataset. On the contrary, we expand each curve in a different basis, which is chosen for each curve according to its particular features. Our work shows the potential of comparing curves through their expansions in bases that are different from one curve to another but individually more efficient.
- Third, we provide for a way to efficiently quantify the closeness of curves with significant sharp local features that might be simultaneously vertically amplified and horizontally shifted. This kind of situation is not satisfactorily dealt with by classical methods.

1.1 The Best Suited Unbalanced Haar Wavelet Basis

As a first step of the procedure, each series of a dataset is expanded in a particular wavelet basis, which is referred to as the *Best Suited Unbalanced Haar Wavelet Basis*

(BSUHWB). We denote the expansion of a series x in this basis as

$$x = \sum_{k=0}^{N-1} d_k \psi_k, \quad (1.1.1)$$

where the coefficients d_k (hereafter the *detail* coefficients) are the projections of the series x on the corresponding basis vectors ψ_k . The BSUHWB basis is obtained using the *Bottom-Up Unbalanced Haar Wavelet Transform* (BUUHWB) proposed by Fryzlewicz (2007). This BUUHWB algorithm provides for a piecewise constant approximation of the curve underlying the series, in a wavelet basis which efficiently captures sharp patterns. In this Chapter, we explicitly highlight a particular notion of hierarchy in the resulting expansion, this notion of hierarchy being at the core of our methodology.

1.1.1 The BUUHWB Algorithm

The BUUHWB algorithm is a transform that associates a given series $x \in \mathbb{R}^N$ to its BSUHWB expansion (eq. (1.1.1)). It is described in Fryzlewicz (2007), p. 1325, and we refer the reader to this paper. Our contribution here is to reformulate this algorithm in a geometrical framework which emphasizes its opportunity for our purpose. We note that, in contrast to Fryzlewicz (2007) who focuses on obtaining the expansion coefficients $\{d_k\}_{k=0 \dots N-1}$ in the BSUHWB, we also explicitly need to build the basis vectors $\{\psi_k\}_{k=0 \dots N-1}$.

A series of length N encodes the expansion coefficients of a vector x in the canonical basis $\{e_j\}_{j=1 \dots N}$ of $\mathbb{R}^N$. We note that this basis is made of generating vectors of $\mathbb{R}^N$ that are ranked according to the horizontal location of the single measurement they support (ordering in time, wavelength, e.g.). Each basis vector is associated with a weight ω_j that we give to the j^{th} point of measurement. Most commonly it makes sense to assume identical weight $\omega_j = 1$ for all indexes $j, j = 1 \dots N$.

The BUUHWB algorithm can be seen as a functional data-driven procedure for dimension reduction. We reduce step-by-step the number of components available for describing x . This reduction is achieved by projecting one well-chosen plane, defined by two consecutive basis vectors, onto the single normed axis which support their weighted average. Here is how it works at the first iteration of the algorithm.

- At the first iteration of the algorithm, there are $N - 1$ candidate planes for dimension reduction, being the planes generated by the pairs $\{e_j, e_{j+1}\}$, with

$j = 1 \dots N - 1$. We note that we do not consider reducing the plane generated by $\{e_N, e_1\}$ or any plane generated by $\{e_j, e_l\}$ with $l \neq j + 1$ as this would not make sense for an ordered series of measurements. This is the *functional constraint* we impose on the dimension reduction.

- For each of the candidate planes, the associated candidate axis for the projection is defined as

$$\Psi_{j,j+1}^p = \frac{\omega_j}{\sqrt{\omega_j^2 + \omega_{j+1}^2}} e_j + \frac{\omega_{j+1}}{\sqrt{\omega_j^2 + \omega_{j+1}^2}} e_{j+1}.$$

It will be referred to as the *projection-axis*. Conform to the intuition, if the axes e_j and e_{j+1} have been given the same weight, the projection-axis is oriented along their bisector and support the mean level of the measurements at locations j and $j + 1$.

- Amongst the so-defined candidate planes, we have to select the one that we will actually reduce. This selection is a *data-driven* choice, with the aim to lose as little information as possible about the shape of the series in the reduction process. The loss associated to a reduction of a plane generated by $\{e_j, e_{j+1}\}$ is defined as

$$L_{j,j+1} := (\mathbf{x} \cdot \Psi_{j,j+1}^d)^2,$$

with $(\cdot)$ indicating the scalar product, and with

$$\Psi_{j,j+1}^d := \frac{\omega_{j+1}}{\sqrt{\omega_j^2 + \omega_{j+1}^2}} e_j - \frac{\omega_j}{\sqrt{\omega_j^2 + \omega_{j+1}^2}} e_{j+1},$$

being perpendicular to $\Psi_{j,j+1}^p$ and referred to as the *detail-axis*. Consequently, the reduction takes place in the plane generated by $\{e_{j^*}, e_{j^*+1}\}$ with

$$\{j^*, j^* + 1\} = \arg \min_{\{j,j+1\}} L_{j,j+1}.$$

Note that if there are several pairs $\{j, j + 1\}$ such that $L_{j,j+1}$ is identical and minimum, we select j^* minimum. This is an arbitrary choice that can not be avoided.

- Given this reduction, the vectorial subspace $V^{[N-1]}$ in which we describe the series $\mathbf{x}$ is now generated by the $N - 1$ following vectors:

$$e_1, \dots, e_{j^*-1}, \Psi_{j^*,j^*+1}^p, e_{j^*+2}, \dots, e_N$$

and $\mathbf{x}$ is approximated by

$$\mathbf{x}^{[N-1]} := \sum_{\substack{j=1 \\ j \neq j^*, j^*+1}}^N (\mathbf{x} \cdot \mathbf{e}_j) \mathbf{e}_j + (\mathbf{x} \cdot \Psi_{j^*, j^*+1}^p) \Psi_{j^*, j^*+1}^p.$$

The weight associated with Ψ_{j^*, j^*+1}^p is defined as

$$((\omega_1, \dots, \omega_{j^*-1}, \omega_{j^*}, \omega_{j^*+1}, \omega_{j^*+2}, \dots, \omega_N) \cdot \Psi_{j^*, j^*+1}^p),$$

while the weights ω_j associated to $\mathbf{e}_j$ with $j \neq j^*, j^*+1$ remain unchanged.

- The discarded vector Ψ_{j^*, j^*+1}^d is encoded as the basis vector ψ_{N-1} of the BSUHWB, and $(\mathbf{x} \cdot \Psi_{j^*, j^*+1}^d)$ is encoded as the detail coefficient d_{N-1} .

We then reduce by one additional unit the dimension of the vectorial subspace in which $\mathbf{x}$ is projected, and we do so until the series approximation reduces to the projection of $\mathbf{x}$ onto one single axis. Thereby, we successively build the BSUHWB vectors $\psi_{N-2}, \psi_{N-3} \dots \psi_1$ and the associated detail coefficients $d_{N-2}, d_{N-3} \dots d_1$, in exactly the same manner. The remaining basis vector ψ_0 is a constant normed vector and $d_0 = (\mathbf{x} \cdot \psi_0)$. We emphasize here that the basis vectors of the BSUHWB are ranked in the reverse order of their construction.

The algorithmic counterpart of this geometrical description is as follows.

The BUHWT algorithm

Input

$\mathbf{x} \in \mathbb{R}^N$ is encoded in the canonical basis $\{\mathbf{e}_j\}_{j=1 \dots N}$: $\mathbf{x} = \sum_{j=1}^N x_j \mathbf{e}_j$.

Output:

$\mathbf{x} \in \mathbb{R}^N$ is encoded in its BSUHWB $\{\psi_k\}_{k=0 \dots N-1}$: $\mathbf{x} = \sum_{k=0}^{N-1} d_k \psi_k$.

Notations:

Index i designates the iteration of the algorithm;

$\mathbf{x}^{[N-i]}$ is the approximation of x in dimension $N-i$ at iteration i ;

$\{\mathbf{f}_j\}_{j=1 \dots N-i}$ is a complete set of generating vectors for the vectorial subspace $V^{[N-i]}$ in which $\mathbf{x}^{[N-i]}$ is defined;

$\{\omega_j\}_{j=1 \dots N-i}$ is a set of weights associated to the $N-i$ components of $\mathbf{x}^{[N-i]}$;

$\Psi_{j, j+1}^p$ is the candidate projection-axis in the plane generated by $\{\mathbf{f}_j, \mathbf{f}_{j+1}\}$;

$\Psi_{j,j+1}^d$ is the associated detail-axis;

$L_{j,j+1}$ is the information loss associated to a projection of the plane $\{\mathbf{f}_j, \mathbf{f}_{j+1}\}$ onto the axis

$\Psi_{j,j+1}^p$.

Initialization:

$i := 1$;
 $\mathbf{x}^{[N]} := \mathbf{x} = \sum_{j=1}^N x_j \mathbf{e}_j$;
 $\{\mathbf{f}_j\}_{j=1 \dots N} := \{\mathbf{e}_j\}_{j=1 \dots N}$;
 $V^{[N]} := R^N$;
 for $j = 1 \dots N$, $\omega_j := 1$.

Iteration #i:

1. Definition of candidate pairs of generating vectors for dimension reduction:
 $\{\mathbf{f}_j, \mathbf{f}_{j+1}\}$, $j = 1 \dots N - i$
2. Definition of associated optimal candidate projection-vector, detail-vector and information loss:
 $\Psi_{j,j+1}^p := \sin \phi \mathbf{f}_j + \cos \phi \mathbf{f}_{j+1}$;
 $\Psi_{j,j+1}^d := \cos \phi \mathbf{f}_j - \sin \phi \mathbf{f}_{j+1}$;
 with $\phi := \arctan(\frac{\omega_j}{\omega_{j+1}})$, $\phi \in [0; \frac{\pi}{2}]$;
 $L_{j,j+1} := (\mathbf{x}^{[N-i+1]}, \Psi_{j,j+1}^d)^2$.
3. Selection of candidate pairs of generating vectors for dimension reduction:
 $\{\mathbf{f}_{j^*}, \mathbf{f}_{j^*+1}\}$ such that $\{j^*, j^* + 1\} = \arg_{j,j+1} \min L_{j,j+1}$.
4. Dimension reduction:
 $\mathbf{x}^{[N-i]} := \sum_{\substack{j=1 \\ j \neq j^*, j^*+1}}^{N-i+1} \mathbf{x} \cdot \mathbf{f}_j \mathbf{f}_j + \mathbf{x} \cdot \Psi_{j^*,j^*+1}^p \Psi_{j^*,j^*+1}^p$
5. Storage of BSUHWB vector:
 $\psi_{N-i} := \Psi_{j^*,j^*+1}^d$;
 $d_{N-i} = \sqrt{L_{j^*,j^*+1}}$
6. Preparation for the next iteration:
 $\{\omega_j\}_{j=1 \dots N-i} := \{\omega_1, \dots, \omega_{j^*-1}, \sqrt{\omega_{j^*}^2 + \omega_{j^*+1}^2}, \omega_{j^*+2}, \dots, \omega_{N-i+1}\}$;
 $\{\mathbf{f}_j\}_{j=1 \dots N-i} := \{\mathbf{f}_1, \dots, \mathbf{f}_{j^*-1}, \Psi_{j^*,j^*+1}^p, \mathbf{f}_{j^*+2}, \dots, \mathbf{f}_{N-i+1}\}$;
 $V^{[N-i]}$ is the vectorial subspace generated by $\{\mathbf{f}_j\}_{j=1 \dots N-i}$;
 $i := i + 1$.
7. Back to Step 1, until $\dim(V^{[N-i]}) = 1$.

Final-step:

$\psi_0 := \mathbf{f}_1$;
 $d_0 := \mathbf{x}^{[1]} \cdot \mathbf{f}_1$.

In the above implementation, the computational complexity of the BUHWT is $\mathcal{O}(N^2)$ because there are $N - 1$ iterations, plus the final step and $N - i$ candidate details are examined at each iteration i . However, at most two candidate details are modified at each iteration so that the computational complexity might be reduced by storing the ranked values of candidate details from one iteration to the next. Theoretically, this should lead to a computational complexity $\mathcal{O}(N \log N)$.

1.1.2 Some Properties of the BSUHWB.

We outline here some properties of the BSUHWB.

The Shape of the Basis Vectors. Going back to the construction of the basis vector ψ_{N-1} (at iteration 1 of the algorithm), imagine we reduce the plane generated by e_j and e_{j+1} , which supports the measurements at abscissas t_j and t_{j+1} . Assuming $\omega_j = \omega_{j+1} = 1$, the detail-vector $\Psi_{j,j+1}^d$ is obtained as $\frac{1}{\sqrt{2}}e_j - \frac{1}{\sqrt{2}}e_{j+1}$. When plotted as a function of time, $\Psi_{j,j+1}^d$ is thus characterized by an *up-and-down* pattern at abscissas j and $j+1$, while zero-valued otherwise. That *up-and-down* shape is actually the basic shape for all of our basis vectors (except for the constant vector), which are built as translations and asymmetric dilations of this *up-and-down* pattern. For instance, if we imagine reducing a plane generated by e_j and by one axis that has already been obtained as the average of two basis vectors e_{j+1} and e_{j+2} , the resulting detail-axis is defined as

$$\begin{aligned}\Psi_{j,j+1,j+2}^d &= \frac{\omega_{j+1,j+2}}{\sqrt{\omega_j^2 + \omega_{j+1,j+2}^2}}e_j - \frac{\omega_j}{\sqrt{\omega_j^2 + \omega_{j+1,j+2}^2}}\left(\frac{1}{\sqrt{2}}(e_{j+1} + e_{j+2})\right) \\ &= \frac{2}{\sqrt{6}}e_j - \frac{1}{\sqrt{6}}(e_{j+1} + e_{j+2}).\end{aligned}$$

It has thus a positive component at abscissa j and negative components at abscissas $j+1$ and $j+2$. This is again an *up-and-down* shape, but *unbalanced*, and it is easy to verify that this remains true for any detail-axis built by the BUUHWB algorithm. Furthermore, the so-built basis vectors of the BSUHWB are orthonormal, as the vectors Ψ^p and Ψ^d are always defined through a plane rotation in an orthonormal basis, so that orthonormality is preserved. For instance, we have

$$\begin{pmatrix} \Psi_{j,j+1}^p \\ \Psi_{j,j+1}^d \end{pmatrix} = \begin{pmatrix} \sin \phi & \cos \phi \\ \cos \phi & -\sin \phi \end{pmatrix} \begin{pmatrix} e_j \\ e_{j+1} \end{pmatrix}$$

with $\phi = \arctan \frac{\omega_j}{\omega_{j+1}}$; $\phi \in [0; \frac{\pi}{2}]$.

An Unbalanced Haar Wavelet Basis. The above properties of the BSUHWB make it a member of the general family of *Unbalanced Haar wavelets bases*, which was introduced in Girardi and Sweldens (1997). The bases of this family are made up of one constant normalized vector and a set of Haar-like wavelets of which a general

mathematical expression is given by

$$\psi_k(t) = \left(\frac{1}{b_k - s_k + 1} - \frac{1}{e_k - s_k + 1} \right)^{1/2} \cdot \mathbf{1}_{s_k \leq t \leq b_k} \quad (1.1.2)$$

$$- \left(\frac{1}{e_k - b_k} - \frac{1}{e_k - s_k + 1} \right)^{1/2} \cdot \mathbf{1}_{b_k + 1 \leq t \leq e_k}, \quad (1.1.3)$$

where $t = 1 \dots N$ is a discrete index along the abscissa axis, and where $s_k < b_k < e_k$ are values along this axis that determine the particular shape of one wavelet ψ_k (s_k , b_k and e_k stand for *start*, *breakpoint* and *end* respectively). Figure 1.1 is an illustration. The wavelets are thus defined as translations and asymmetric dilations of the *up-and-down* pattern which characterizes Haar wavelets. In such a way, each wavelet ψ_k is associated with a level change from one observation (or group of observations) to the consecutive one, this level change being located around the breakpoint b_k . The projection of the series x on the wavelet ψ_k encodes the importance of the related level change in the series x . This makes unbalanced Haar wavelets efficient at encoding peaks and level changes, as usual Haar wavelets do. Moreover, the possibility of having asymmetrical dilations of the *mother up-and-down* wavelet leads to wavelet bases that overcome the dyadic restriction, this feature precisely defining *unbalanced* wavelets. Actually, any choice of $N - 1$ sets of values (s_k, b_k, e_k) that leads to orthonormal wavelets defines a basis of $\mathbb{R}^N$. There is thus a whole family of unbalanced Haar wavelet bases. The BUHWT algorithm can thus be seen as a *data-driven* way for selecting one particular basis in this family. An example of the BSUHWB expansion obtained for one particular series is provided in Figure 1.2, *left*.

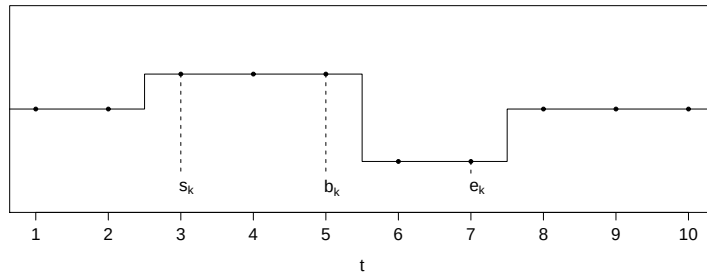


Figure 1.1: An unbalanced Haar wavelet. The points s_k , b_k and e_k (eq. 1.1.2) are identified.

The Signature of the Series in the b - d Plane. An interesting property of the Unbalanced Haar wavelet bases, is the possibility of a sparse encoding of the basis. This follows from the following property (Fryzlewicz 2007):

Property: The BSUHWB $\{\psi_k\}_{k=0\dots N-1}$ is uniquely determined by the ordered set of breakpoints $\{b_k\}_{k=1\dots N-1}$ of its wavelet basis vectors.

As a consequence, and because expansion (1.1.1) is invertible and hence unique, the set of points $\{y_k\}_{k=1\dots N-1} = \{(b_k, d_k)\}_{k=1\dots N-1}$ determines the *shape* of the series x uniquely, the complete determination of x requiring the additional coefficient d_0 that encodes the mean level of the series. Given this interesting property, we propose a new nearly-unique representation of a series, in a plane whose generating axes support breakpoints and details coefficients respectively. We refer to this representation as to *the signature of the series x in the b - d plane*. An example of such a representation is presented in Figure 1.2, *right*.

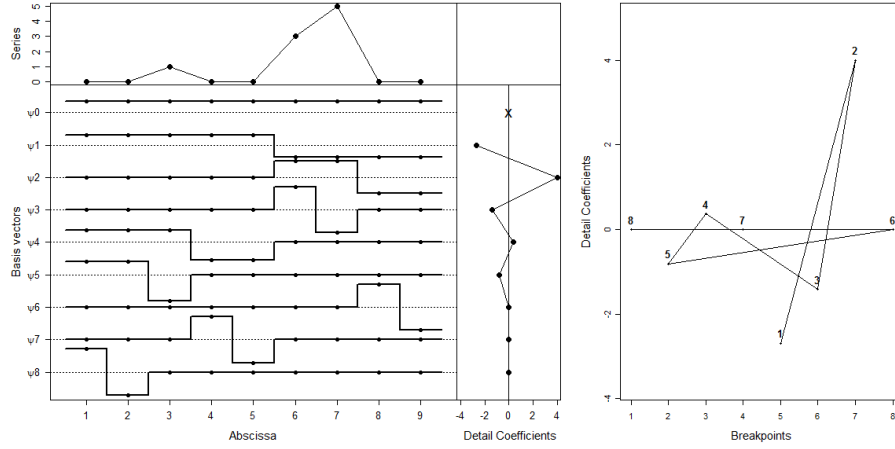


Figure 1.2: *Left: Illustration of the BSUHWB expansion of a series.* The series is plotted in the upper part of the Figure. The corresponding abscissa axis is common for that graph and for the main graph, so that it is located at the bottom of that one. The main part of the Figure shows the basis vectors of the Unbalanced Haar wavelet basis that is best suited to the series (BSUHWB basis). These vectors are represented rank by rank, as a function of an index along the abscissa axis. Dotted horizontal lines indicate the level zero for each rank. Vertically, from top to bottom on the right hand side, we find the detail coefficients associated with the wavelet expansion. Each coefficient is located next to the corresponding basis vector. For graphical convenience, the value of the coefficient d_0 associated with the constant vector ψ_0 is not indicated. *Right: Representation of the same series in the b - d plane.* Points $(b_k, d_k)_{k=1\dots N-1}$ are numbered according to their rank k in the hierarchical BSUHWB and linked to each other.

The Hierarchy in the BUUHW T Expansion. Finally, the BUUHW T induces an interesting property of hierarchy in the resulting BSUHW B expansion. The explicit identification of this hierarchy is the starting point of our work. The idea is that the ranking of the basis vectors of the BSUHW B reflects the decreasing importance of the patterns they encode, for the description of the global shape of the series. The notion of *hierarchy* that we refer to is the hierarchy induced by the BUUHW T algorithm itself : by construction, our series x is encoded in its BSUHW B as the sum of a constant mean level (rank $k = 0$) and a linear combination of level changes, the few first (small rank indexes) encoding the most striking features of the series, while the last ones (large rank indexes) are less important. This idea of “importance” in the hierarchy could be interpreted as *an ordering according to the power of the supported up-and-down pattern in the signal, under a functional constraint*. Indeed, the BUUHW T is an attempt to concentrate as little power as possible at large rank indexes, and as much power as possible at small rank indexes, while keeping in mind the functional nature of the series - this constraint implies that only neighbor measurements can be averaged in the data reduction process leading to the basis construction, and implies thus that the *detail* coefficients are not necessarily strictly decreasing as the rank index increases. We emphasize the fact that this ranking of the unbalanced Haar wavelets directly depends on the data, and not on the length of their support as for usual wavelets. It does not meet the notion of scale in usual wavelet analysis.

If we consider again the example provided in Figure 1.2, *left*, we observe that the first non-constant basis vectors support the highest peak of the series. This can be observed by looking at the location of the breakpoints b_k between positive and negative parts of the wavelets. Subsequent basis vectors point to smaller level changes while the few last basis vectors correspond to zones where there is no level change - as indicated by the associated zero coefficient. In this expansion, we note that detail coefficients are not ordered according to their magnitude. As mentioned above, this makes sense as it directly results from the *functional* nature of the algorithm, which only allows for reducing dimension by reducing a plane generated by *neighbor* vectors in $V^{[N-i+1]}$. Consequently, the set of projections on the candidate detail-axes (i.e. the candidate detail coefficients) over which we minimize is redefined at each step leading to a possible non-ordering of their successive minima.

1.2 The BAGIDIS Semi-Distance

As discussed here-above, the *Bottom-Up Unbalanced Haar Wavelet Transform* allows for an automatic and unique hierarchical description of each of the series into a series-adapted orthonormal basis. The hierarchy makes the resulting bases comparable to each other, although different. Consequently, we propose to compare the series rank

by rank through their signatures in the $b - d$ plane, i.e. through their mapping into the location-amplitude space of their breakpoints and details coefficients.

1.2.1 Definition

The BAGIDIS dissimilarity between two series $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ is built as a hierarchically weighted sum of partial distances in the $b - d$ plane:

$$d_p^{\text{BAGIDIS}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \sum_{k=1}^{N-1} w_k \delta_p^k(\mathbf{y}_k^{(1)}, \mathbf{y}_k^{(2)}), \quad (1.2.1)$$

with

$$\delta_p^k(\mathbf{y}_k^{(1)}, \mathbf{y}_k^{(2)}) = \left\| \mathbf{y}_k^{(1)} - \mathbf{y}_k^{(2)} \right\|_p, \quad \text{and} \quad \mathbf{y}_k^{(i)} = (b_k^{(i)}, d_k^{(i)}), \quad k = 1 \dots N - 1,$$

with w_k , $k = 1 \dots N - 1$ a non-negative weight function, and with p defining any kind of norm.

Interpretation. The BAGIDIS dissimilarity takes advantage of the hierarchy of the BSUHWB: breakpoints and details of similar rank k in the hierarchical description of each series are compared to each other, and the resulting differences can be weighted according to that rank. As the breakpoints point to level changes in the series, the term $|b_k^{(1)} - b_k^{(2)}|$ can be interpreted as a measure of the difference of location of the features, along the horizontal axis. Being a difference of the projections of the series onto wavelets that encode level changes, the term $|d_k^{(1)} - d_k^{(2)}|$ can be interpreted as a measure of the differences of the amplitudes of the features, along the vertical axis. Consistently with the definition of the $b - d$ plane, we do not consider rank $k = 0$ in expression (1.2.1), which means that we compare the structures of the series rather than their mean levels.

The Weight Function. The weight function w_k is a crucial parameter in equation (1.2.1). It should ideally be strictly positive at rank k if that rank carries significant information about the structure of the series and 0 otherwise. In such a way, the weights can act as a filter that extracts the part of the curves that carries relevant features. In the present study, we do not enter into the discussion of the optimization of the weight function, as we are mainly interested here in presenting our semi-distance as a new paradigm for defining dissimilarities, leaving technical details for subsequent Chapters (e.g. Chapter 2). Some perspectives are given in Subsection 1.2.3, however. In the examples of this Chapter, we make use of two different approaches.

- The following weight function

$$w_k = \frac{\log(N + 1 - k)}{\log(N + 1)}, \quad k = 1 \dots N - 1, \quad (1.2.2)$$

is used as a *prior*. It decreases slowly until k is close to $N - 1$ and drops rapidly afterward. This behavior is empirically expected for noisy series with several features of decreasing importance, as it takes into account the hierarchical nature of the BSUHWB: the partial distance $\delta_p^k = \left\| \mathbf{y}_k^{(1)} - \mathbf{y}_k^{(2)} \right\|_p$ at rank k should affect the global dissimilarity measure as much as the compared patterns carry major features of the series, and significant discriminative patterns are more likely to appear in the first ranks of the BSUHWB hierarchy. This empirical idea is in accordance with the observations of Fryzlewicz (2007) in his study of curve denoising using unbalanced Haar wavelets. This approach is used in most of the subsequent examples.

- The previous approach might not be the most suitable when the curves of the dataset (or the signal underlying the noisy curves) have a sparse encoding in the BSUHWB: the above-mentioned prior might give too much weight to the noise in that case. In case of sparse curves, of which the BSUHWB encoding reduces to a small number of non-zero coefficient, the weight function is defined hereafter as $w_k = 1$ for $k = 1 \dots K$ and 0 otherwise, with K the number of ranks needed to encode the signal.

1.2.2 Properties

We outline here some features and properties of d_p^{BAGIDIS} .

A Visual Motivation. The motivation for equation (1.2.1) is that the visual comparison of curves implicitly relies on hierarchy. Indeed, when we visually evaluate dissimilarities between series, we intuitively investigate first the global shapes of the series for estimating their resemblance. Then we refine the analysis by comparing their smaller features. It reveals a hierarchical comprehension of the curves. This visual approach inspired us to define our semi-distance: we expand each series in a (different, series-adapted) basis that describes its features hierarchically, i.e. the first few basis vectors carry the main features of the series while subsequent basis vectors support less significant patterns; afterward, we compare both the bases and the expansions of the series onto those bases, rank by rank, according to the hierarchy.

Theorem: The BAGIDIS Dissimilarity is a Semi-Distance. Mathematically, d is called a semi-distance on some space $\mathcal{F}$ if

- $\forall x \in \mathcal{F}, \quad d(x, x) = 0$
- $\forall x^i, x^j, x^k \in \mathcal{F}, \quad d(x^i, x^j) \leq d(x^i, x^k) + d(x^k, x^j).$

It has thus the same properties as a distance, except that $d(x^i, x^j) = 0 \not\Rightarrow x^i = x^j$. A semi-distance can be used instead of a distance when the interest lays in comparing the shapes of groups of curves, but not their mean level.

Proof: By definition of the norm $\|\cdot\|_p$, and because the weights w_k are non negative, we have immediately:

$$\begin{aligned} d_p^{\text{BAGIDIS}}(x^{(1)}, x^{(2)}) &\geq 0, \\ d_p^{\text{BAGIDIS}}(x^{(1)}, x^{(2)}) &= d_p^{\text{BAGIDIS}}(x^{(2)}, x^{(1)}), \end{aligned}$$

and

$$\begin{aligned} d_p^{\text{BAGIDIS}}(x^{(1)}, x^{(2)}) &= \sum_{k=1}^{N-1} w_k \left\| y_k^{(1)} - y_k^{(2)} \right\|_p \\ &\leq \sum_{k=1}^{N-1} w_k \left\| y_k^{(1)} - y_k^{(3)} \right\|_p + \sum_{k=1}^{N-1} w_k \left\| y_k^{(3)} - y_k^{(2)} \right\|_p \\ &= d_p^{\text{BAGIDIS}}(x^{(1)}, x^{(3)}) + d_p^{\text{BAGIDIS}}(x^{(3)}, x^{(2)}). \end{aligned} \tag{1.2.3}$$

Hence, the dissimilarity d_p^{BAGIDIS} is a semi-distance.

The reason why our dissimilarity is a semi-distance instead of a distance is twofold. First, we do not compare the coefficients $d_0^{(1)}$ and $d_0^{(2)}$, which means that two series having exactly the same shape but a different mean level will be measured at a distance zero. Second, nothing prevents some weights w_j to be zero, which means that the comparison of two series might be reduced to a comparison of only a few ranks in the hierarchy through a suitable choice of the weight function.

Comments on the Dimensionality. Although encoded in $\mathbb{R}^N$, the curves underlying the measured series are infinite dimensional objects. Reducing those curves to their vectorial series in $\mathbb{R}^N$ is too drastic, as it misses the notion of neighborhood which is crucial for efficiently dealing with functional data, as will be discussed in Subsection 1.2.3. Additional information is thus needed. We provide this by introducing the breakpoints coefficients. In such a way, we *reduce* an infinite dimensional problem to

a $2N - 2$ dimensional one, without losing its functional nature. Furthermore, dimension reduction can be significantly improved by an efficient sparse definition of the weight function w_k in expression (1.2.1).

A Unified Framework for Quantifying Horizontal Shifts and Vertical Amplifications. Being defined in the $b - d$ plane, the BAGIDIS semi-distance accounts jointly for differences in the locations and amplitudes of the features of given rank in the series. In such a way, BAGIDIS provides for a way to consistently quantify differences between curves characterized by vertically- and horizontally-varying sharp local features. This means that series having a similar feature, that might be affected either by a vertical amplification or by horizontal shifts or both, are measured close to each other, and far from a series having different features. This behavior is empirically expected, but rarely achieved with usual methods, as will be discussed in Subsection 1.2.3. It is thus a major strength of the BAGIDIS semi-distance.

An example is provided with the curves in Figure 1.3 *top left*. A series x is defined. It is compared with 5 different series. A series is defined as $1.125 * x$, this amplification factor being chosen so that the upper plate of the series rises by 1 unit. It is also compared with the series x shifted by one unit to the left. Those three series are expected to be measured close to each other. The series defined as $1.125 * x$ shifted by one unit to the left is expected to be close of them too, although slightly more distant of x as it is affected by the vertical and horizontal variations simultaneously. The fourth series is less similar, but still has the same gross shape as x . It is expected to be more distant as compared with the other series. The last series is completely different and is expected to be even more distant.

The signatures in the $b - d$ plane of the 5 above-defined series are presented in Figure 1.3 *top right, middle left, middle right, bottom left and bottom right*, respectively, as compared with the signature of x . All those series have a sparse encoding in the BSUHWB: less than 3 ranks are non-zero in the BSUHWB expansions. Consequently, we make use of the BAGIDIS semi-distance with $w_k = 1$ for $k = 1, 2, 3$ and $w_k = 0$ otherwise. The resulting dissimilarity matrix is provided in Table 1.2.2. It has the empirically expected behavior, which illustrates the consistency of the method. On the opposite, dissimilarity matrices obtained using common competitor distances or semi-distances (also provided in Table 1.2.2) fail to detect the similarity of the series that are horizontally shifted with respect to each other (in case of the Euclidean distance, the Euclidean distance between derivatives and the Haar distance) or fail to detect that those series are not exactly the same (in case of the dynamic time warping distance). The reason for these behaviors is discussed in Subsection 1.2.3.

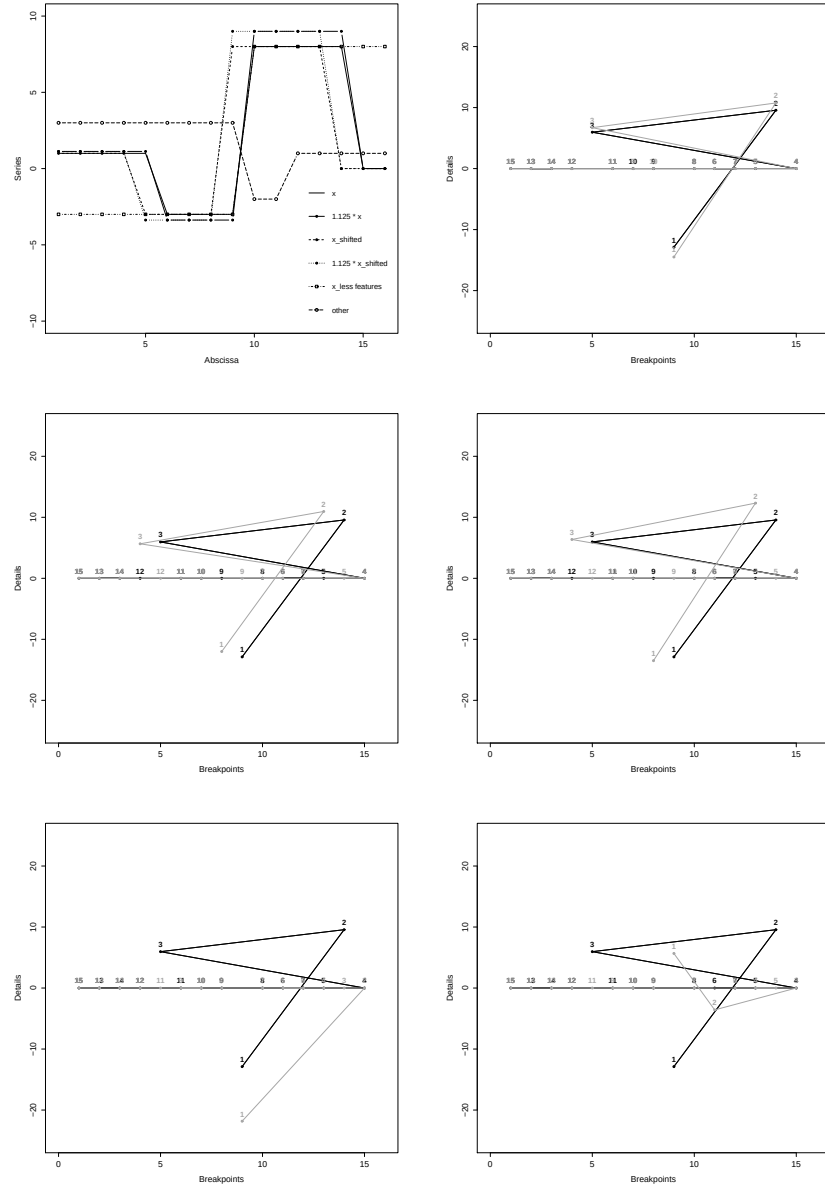


Figure 1.3: **Comparison of 6 series in the $b - d$ plane.** *Top left:* Representation of the 6 series. *From top right to bottom left:* Representation in the $b-d$ plane of the series x (in black) and $1.125 * x$, x shifted to the left, $1.125 * x$ shifted to the left, a simplified version of x and another series, successively (in grey).

Dissimilarity matrix						
BAGIDIS						
	x	1.125*x	x left shifted	1.125*x left shifted	x less features	other series
x	0.00	2.51	2.89	3.67	20.76	30.89
1.125*x	2.51	0.00	3.65	3.06	20.76	33.13
x - left shifted	2.89	3.65	0.00	2.53	22.98	31.64
1.125*x - left shifted	3.67	3.06	2.53	0.00	23.14	33.90
x - less features	20.76	20.76	22.98	23.14	0.00	23.95
other series	30.89	33.13	31.64	33.90	23.95	0.00
Euclidean Distance						
	x	1.125*x	x left shifted	1.125*x left shifted	x less features	other series
x	0.00	5.64	201.00	231.62	208.00	513.00
1.125*x	5.64	0.00	231.89	254.39	218.64	616.14
x - left shifted	201.00	231.89	0.00	5.62	377.00	486.00
1.125*x - left shifted	231.62	254.39	5.62	0.00	408.62	585.62
x - less features	208.00	218.64	377.00	408.62	0.00	769.00
other series	513.00	616.14	486.00	585.62	769.00	0.00
Euclidean Distance between Derivatives						
	x	1.125*x	x left shifted	1.125*x left shifted	x less features	other series
x	0.00	3.14	402.00	455.39	80.00	345.00
1.125*x	3.14	0.00	455.39	508.78	103.14	412.14
x - left shifted	402.00	455.39	0.00	3.14	322.00	235.00
1.125*x - left shifted	455.39	508.78	3.14	0.00	375.39	288.39
x - less features	80.00	103.14	322.00	375.39	0.00	265.00
other series	345.00	412.14	235.00	288.39	265.00	0.00
Haar distance						
	x	1.125*x	x left shifted	1.125*x left shifted	x less features	other series
x	0.00	4.58	200.94	231.06	207.00	511.44
1.125*x	4.58	0.00	230.25	254.31	214.51	610.94
x - left shifted	200.94	230.25	0.00	4.62	376.44	485.00
1.125*x - left shifted	231.06	254.31	4.62	0.00	405.56	581.62
x - less features	207.00	214.51	376.44	405.56	0.00	768.94
other series	511.44	610.94	485.00	581.62	768.94	0.00
Dynamic time warping distance						
	x	1.125*x	x left shifted	1.125*x left shifted	x less features	other series
x	0.00	14.125	0.00	14.00	36.00	77.00
1.125*x	14.125	0.00	14.00	0.00	53.125	83.00
x - left shifted	0.00	14.00	0.00	13.874	40.00	76.00
1.125*x - left shifted	14.00	0.00	13.875	0.00	57.00	82.125
x - less features	36.00	53.125	40.00	57.00	0.00	133.00
other series	77.00	83.00	76.00	82.175	133.00	0.00

Table 1.1: **Dissimilarity matrices between the 6 series defined in Figure 1.3.** The dissimilarity matrices are computed with the BAGIDIS semi-distance d_2^{BAGIDIS} , the Euclidean distance, the Euclidean distance between the first derivatives of the curves (estimated as the lag-1 differences) the Haar distance, and the dynamic time warping distance successively. The BAGIDIS semi-distance results in a dissimilarity matrix that matches the expected results, which is not true for competitor distances.

1.2.3 Discussion

The BAGIDIS semi-distance is *functional*, as the ordering of the measurement of each discretized curve is explicitly taken into account. It is also *wavelet-based* as it relies on a basis function expansion in which both the locations and the amplitudes of the patterns are encoded. The semi-distance *overcomes the dyadic restriction* that is attached with classical wavelet expansions. Moreover, *it does not require any preliminary smoothing or realignment* of the data. Vertical amplifications and horizontal shifts of a pattern are measured in a *unified framework*. Finally, a major originality of the BAGIDIS methodology is that our semi-distance relies on projections of the curves onto *basis functions that are different from one series to another*, however remaining comparable since built according to a common generating principle.

Comparison With Competitors

We describe here the key principles of three frequent methods to measure distances or semi-distances between discretized curves, ranging from the most intuitive to the most sophisticated. We illustrate their drawbacks for consistently quantifying differences between curves characterized by vertically- and horizontally-varying sharp local patterns. The problem is that curves similar except for a given vertical variation are measured relatively close to each other while curves similar apart from the same amount of variation along the horizontal axis are measured distant. This is not consistent. It indicates that the specific abilities of BAGIDIS to capture shifts along the horizontal axis makes it really valuable.

Classical l_p Distances and their Principal Components-based Extension. Classical l_p distances compare curves separately at each point of measurement (i.e. at each abscissa), along the ordinate axis. The *principal components-based distance* (Jolliffe 2002, e.g.) acts similarly except that the differences at each given abscissa are suitably weighted, according to the amount of variability in the dataset at that value of the abscissa. Those point-to-point distances are not able to consistently quantify the similarity of shifted curves. It is illustrated in Figure 1.4 as follows. Two modifications (dotted curves) of a similar curve (solid curve) are presented. In the left panel, the modified curve is a vertical amplification of the original pattern. In the right panel, the modified curve is an horizontal shift of the original pattern. The amount of vertical increase of the peak (on the left) is the same as the amount of horizontal shift (on the right). The sums of the vertical differences between the curves are given by the black areas on the second row of the Figure. Classical l_p distances and principal component-based distances use these areas to compare curves. We see that the area is

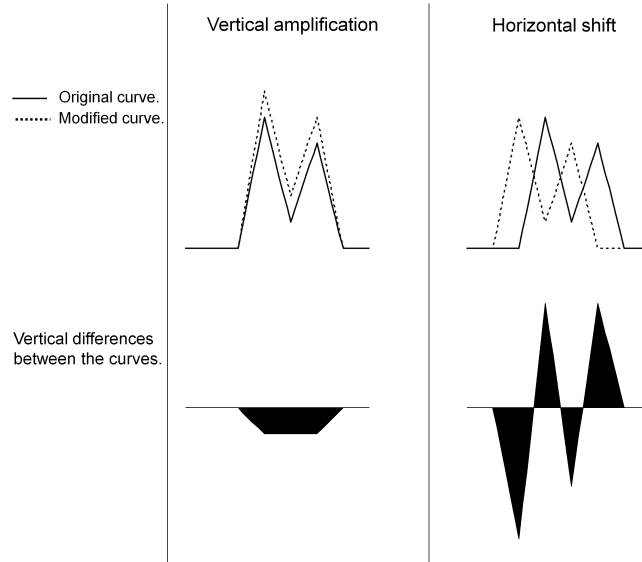


Figure 1.4: Schematic illustration of the difficulty for point-to-point distances to take into account horizontal variations of curves.

significantly larger for an horizontal shift than for a vertical increase. It is not satisfying. The weakness is that the ordering of the measurements in the series is not taken into account so that the evolutions of two series cannot be compared.

Functional Semi-Distances. Functional methods have been proposed in a view to be able to compare the evolution of curves. The general idea of those methods is to explicitly consider the fact that a series of measurements actually derives from a continuous process (i.e. a curve). In other words, they take into account the notion of neighborhood in the series. Commonly used *functional semi-distances* rely on the point-to-point comparison of the derivatives of the curves (Ferraty and Vieu 2006). Successful applications include the study of chemometric data (Ferraty and Vieu 2006) and the analysis of ozone concentration (Aneiros-Pérez et al. 2004), for instance. However, those methods happen to fail when dealing with curves with local sharp features that might not be well aligned from one curve to another one. Three cases are illustrated in Figure 1.5. The curves we compare are shown in the first row of the Figure (solid and dotted curves). Their derivatives are represented in the second row. As previously, the amount of vertical increase of the peak (on the left panel) is the same as the amount of horizontal shift (in the middle and right panels). The sums of the vertical differences between the derivatives are given by the black area in the third row of the Figure. Derivative-based functional semi-distances use those areas to compare

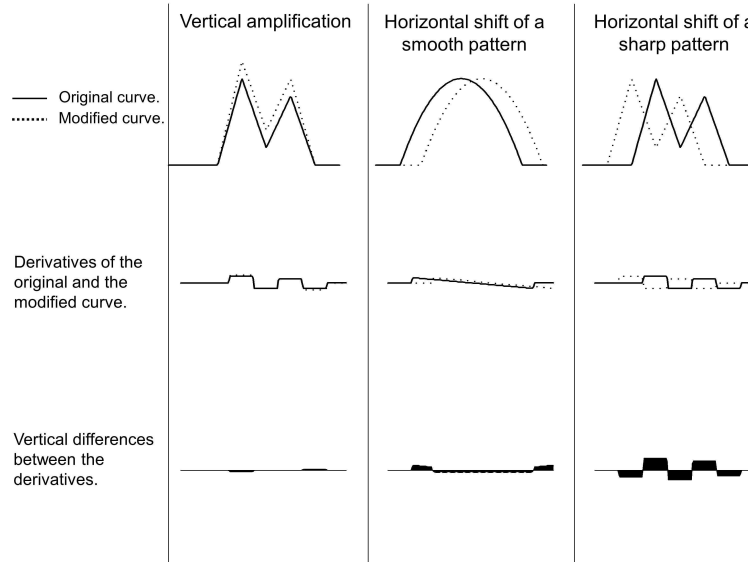


Figure 1.5: Schematic illustration of the difficulty for derivative-based functional semi-distances to take into account horizontal variations of curves with sharp features.

curves. The left panel illustrates that those semi-distances capture well the similarities of curves with sharp patterns when one curve is a vertically amplified version of the other. The middle panel shows that they capture well the similarity of curves with smooth patterns when one curve is a shifted version of the other. However, the right panel indicates that they do not capture the similarity of curves with sharp patterns when one curve is a shifted version of the other. Moreover, derivative-based semi-distances, as well as the functional extension of the principal components-based distance (Ferraty and Vieu 2006), rely on a smoothing of the data, which is problematic for curves with sharp local features.

Wavelet-Based Distances. In contrast to the point-to-point definitions of distances, *wavelet-based distances* rely on the expansion of the series in a well-designed multi-scale basis. The projections of the series onto this basis, i.e their *detail* coefficients, are then compared. For instance, Morris and Carroll (2006) have used such an approach in the context of random or mixed effects models, while Antoniadis et al. (2009) used wavelets for discriminating curves or reducing their dimension by separating “significant” from uninformative coefficients across curves with respect to some class membership. A wavelet basis is made of a fixed pattern present in the basis vectors at different scales and different locations along the abscissa axis. Because of this

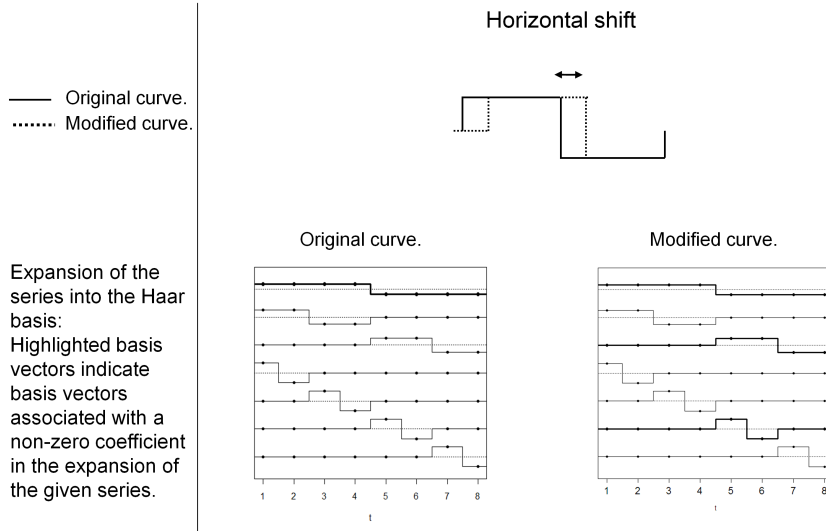


Figure 1.6: Schematic illustration of the difficulty for wavelet-based methods to detect the similarity of curves shifted relative to each other.

pattern-based character, distances based on wavelets can certainly be called functional. Furthermore, as the pattern repeated over different basis vectors has not necessarily to be smooth, wavelet bases are very well suited to highlight the position and amplitude of sharp patterns in the series. However, the description of a given feature is highly sensitive to its location in the series. This is related to the dyadic restriction of classical wavelet bases. The Haar basis is the simplest example: a step function of length $N = 2^m$, for some given m , whose discontinuity point is located at $\frac{N}{2}$ requires one single non-zero coefficient so as to be encoded into the Haar basis; the same step function shifted by one unit (so that its discontinuity point is located at $\frac{N}{2} + 1$) requires m non-zero coefficients when expanded in the Haar basis. This is illustrated in Figure 1.6 as follows. Two shifted step series are considered (solid and dotted curves). They are expanded in the Haar wavelet basis. This basis is represented twice in the second row of the Figure. In the left panel, the only basis vector that is associated with a non-zero coefficient in the expansion of the solid series is highlighted in bold. On the right panel, the three basis vectors that are associated with a non-zero coefficient in the expansion of the dotted series are highlighted in bold. We see that the encoding in the Haar wavelet-basis expansion differs highly for two similar series when the series is shifted a bit. As a consequence of this feature, it becomes difficult to detect their closeness.

Dynamic Time Warping Distance. The *Dynamic Time Warping* (DTW) distance is a method that allows for “speed variations” along the abscissa axis so that the distance between the values of two curves (measured along the ordinate axis) is minimum (see Giorgino (2009), e.g.). DTW is widely used for dealing with misaligned data as it provides for a distance measure which ignores horizontal shifts. Using a DTW distance is thus a way for removing horizontal noise in curve comparison. However, this is not satisfying when the signal consists in a horizontal shift or shape variation, as curves with shifted features are measured identical by DTW. This is illustrated in Figure 1.7. The amount of warping might possibly be encoded and serves as a measure of horizontal difference, but it then misses the vertical differences between the curves. Vertical and horizontal variations can thus not be accounted in a unified framework with DTW-based methods.

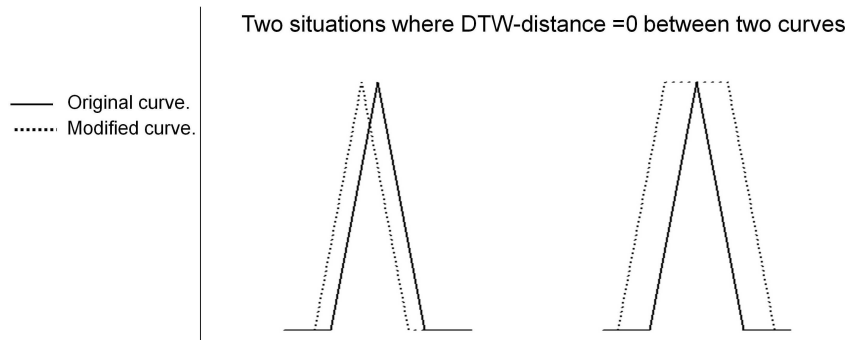


Figure 1.7: Schematic illustration of situations where the *dynamic time warping* distance returns a distance zero between curves of which a pattern is affected by an horizontal shift or shape variation.

Parametrization and Extensions

Scaling. Unavoidably, the dissimilarities we measure using p -norm partial distances (as in expression (1.2.1)) are directly dependent on the scales of measurements along the breakpoints and details axes, which are themselves determined by the original units of the problem. A balance parameter $\lambda \in [0, 1]$ might thus be introduced so as to take into account that sensitivity. This leads to the following variant of the BAGIDIS

dissimilarity between $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$:

$$\begin{aligned} d_{p,\lambda}^{\text{BAGIDIS}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) &= \sum_{k=1}^{N-1} w_k \left\| \mathbf{y}_k^{(1)} - \mathbf{y}_k^{(2)} \right\|_{p,\lambda} \\ &= \sum_{k=1}^{N-1} w_k \left(\lambda \left| b_k^{(1)} - b_k^{(2)} \right|^p + (1 - \lambda) \left| d_k^{(1)} - d_k^{(2)} \right|^p \right)^{1/p}. \end{aligned} \quad (1.2.4)$$

This dissimilarity measure is shown to be a semi-distance in exactly the same manner as d_p^{BAGIDIS} , by observing that $\left\| \mathbf{y}_k^{(1)} - \mathbf{y}_k^{(2)} \right\|_{p,\lambda} = \left\| \tilde{\mathbf{y}}_k^{(1)} - \tilde{\mathbf{y}}_k^{(2)} \right\|_p$, with $\tilde{\mathbf{y}}_k^{(i)} = (\lambda^{1/p} b_k^{(i)}, \lambda^{1/p} d_k^{(i)})$, $i = 1, 2$. This parameter λ corresponds to a scaling of the b - d plane, and hence also of the original units of the problem. In such a way, it makes sense to consider the same value for the scaling for each rank. Although slightly less intuitive, more flexibility might be gained by allowing λ to vary from one rank k to another. This generalization is not considered here, however. Optimizing λ over the semi-distances of series in a dataset according to any criterion will lead to the same relative dissimilarities, whatever the original units of measurements. This translates into a certain robustness of the method. Moreover, investigating how the semi-distances between series of a dataset evolve when λ varies can also be used as a diagnostic tool. In particular, looking at the semi-distances that one obtains by setting λ at its extreme values 0 and 1 allows to diagnose the separate effects of the two components of our measure: differences in the details and in the breakpoints.

As a preprocessing step of the analysis, the series might also be scaled before the BSUHWB is obtained. The opportunity of such a preliminary scaling is problem-dependent, as it often is in multivariate data analyses.

The Weight Function. The weights w_k can be optimized to make our semi-distance most efficient. Although this point is not investigated in the current Chapter, we introduce here some perspective of subsequent work.

- In case of a *supervised prediction problem*, w_k should ideally be 1 if rank k carries significant information for discriminating the series, and 0 otherwise. In such a problem, w_k can easily be optimized using a mean square error minimization criterion within a cross-validation procedure. If the curves in the dataset are sparse enough, this leads to major dimension reduction. That point is further investigated in Chapter 2 in the framework of kernel nonparametric functional regression as defined by Ferraty and Vieu (2006). In this context, Chapter 2 optimizes the weight function using a cross-validation procedure, and shows the optimality of the selected weights. Moreover, it shows that, when w_k can

be chosen sparse enough, the nonparametric functional regression estimator can reach competing rates of convergence provided that some regularity conditions on the regression are satisfied.

- When the BAGIDIS semi-distance is used in an *unsupervised problem*, the optimization of w_k is a bit more elaborated. The following data-driven procedure is currently being investigated. First, we obtain the statistical distribution of a Gaussian white noise in the $b - d$ plane, separately for each rank k . Then, at each rank k , we compare the distribution of the $\mathbf{y}_k = (b_k, d_k)$ obtained for Gaussian noises and the empirical distribution of the $\tilde{\mathbf{y}}_k = (\tilde{b}_k, \tilde{d}_k)$ obtained for the curves of interest, at the same rank k . The idea is that the importance (i.e. the weight) we should associate to the rank k for discriminating the curves is related to the quantity of “structure” that is present in the data at that rank.

An intuition on how to define a *prior* for the parameters. Defining a *prior* for the parameterization of the BAGIDIS semi-distance should take into account the following elements.

- If the curves are piecewise-constant, they are encoded in a sparse way in their BSUHWB: for a curve $\mathbf{x}^{(i)}$, only the first $K^{(i)}$ ranks of the signature carry information about the shape of curve, where $K^{(i)}$ is the number of level changes in the curve. Therefore, one should have $w_k = 0$ for all $k > K^* = \max_i(K^{(i)})$. For instance, one should have $K^* = 1$ if we compare step functions, $K^* = 2$ if we compare curves with one single peak, etc.
- The above observations also serve as a guideline for defining the weight function in case of non piecewise-constant curves (such as piecewise-constant curves with noise): one should have $w_k = 0$ for all $k > K^* = \max_i(K^{(i)})$, where $K^{(i)}$ is the number of ranks in the signature that support for *significant* information about the curves, i.e. $K^{(i)}$ is the number of *significant* level changes in $\mathbf{x}^{(i)}$. For instance, one should have $K^* = 1$ if we compare step functions with noise, $K^* = 2$ if we compare curves with one single peak plus noise, etc.
- The above value K^* is really an upper bound for the number of non-zero coefficients. It might be that all of the ranks of the signatures that are needed for describing the curves are not actually useful for discrimination purposes. The first reason is that there might be some redundancy in the information carried by different ranks. The second reason is that only some of the patterns of the curves might be of interest for discrimination purposes. Here is an example: we have a dataset of curves, each of them containing two peaks of which amplitude and location might vary; one peak remains always higher than the other; in this

case, $K^* = 4$ ranks are needed to encode the significant features of the curves; however, it might be that only the small peak is of interest for some application purpose; then, only ranks $k = 3$ and $k = 4$ should have non-zero weight, as they are sufficient to account for the variations of the small peak.

- The ratio of the values of the non-zero weights determines the ratio of the importances given to the related differences in the distance measure. In case we know that all selected ranks are equally important for the discrimination, setting $w_k = 1$ for all of them is the simplest choice. In case curves are noisy and it is unclear which K^* to choose, using a weight function which decreases when k increases is a way to account for the increasing uncertainty about the fact that rank k carries significant information (this fact comes from the hierarchic construction of the BSUHWB, as discussed in Subsection 1.1.2). This observation was used for defining the *prior* weight function used in this Chapter (eq. 1.2.2).
- If only the amplitude of the significant patterns changes, one should set $\lambda = 0$. For instance, this is the case when we compare a set of curves with a peak at one given location and of which the amplitude varies. If only the location of the significant patterns changes, one should have $\lambda = 1$. For instance, this is the case when a pattern of given amplitude is shifted from one curve to another in the dataset. If both the amplitude and location of a significant pattern change, then the value of λ should be chosen in $[0; 1]$ depending on the scales of measurement along the horizontal and vertical axes, as well as on the respective importances one wants to give to the breakpoints differences and to the detail differences in the measure of distance. For instance, if a dataset of curves consists in a set of step functions of which both the location and the amplitude vary, setting $\lambda = 1$ leads to a distance measure that ignores the amplitude differences: all curves of which the steps are located at the same position will be seen identical, whatever the height of the step.

Extension to Other Wavelet Bases and Basis Selection Procedures. An extension of the BAGIDIS semi-distance to other wavelet bases seems to require three conditions:

- the family of wavelet bases is made of adaptive unbalanced wavelets.
- there exists an algorithm for selecting a particular basis in the given wavelet basis family, that is suited to the series according to a principle of hierarchy.
- there exists a sparse encoding of the selected wavelet basis, so that unique and reversible signatures of the series might be defined.

We believe that the BAGIDIS semi-distance might be extended to piecewise linear unbalanced wavelet, by using the extension of the BUUHW algorithm introduced in Fryzlewicz (2007). Those “triangular” wavelets might lead to better adapted decompositions in case of curves with not-so-sharp peaks. Besides, an extension to multi-dimensional unbalanced Haar wavelets gives promising results for image processing (Chapter 4).

Regarding the selection of an unbalanced Haar basis that is well-suited to a series, an alternative to BUUHW is the top-down unbalanced Haar transform of Fryzlewicz (2007). However, this transform does not come with an associated property of hierarchy which makes it unsuitable for the BAGIDIS methodology. Moreover, the BUUHW transform has an extension to images, as defined in Chapter 4, which allows to generalize the BAGIDIS method. As far as we know, the top-down algorithm has no such multidimensional extension.

We note that the bottom-up character of the BUUHW might remind the reader of lifting methods (Sweldens 1996; Jansen et al. 2009) which also involve bottom-up algorithms for computing wavelet transforms, and have been a starting point, amongst other uses, for defining wavelet transforms that are adaptive to an irregular design (Delouille 2002). On the contrary, the BUUHW assumes the design to be regular. Moreover, it computes the detail coefficients in an order that depends on the *values* of the data. This is a key difference with lifting methods.

Limits and Range of Applicability

For the BAGIDIS semi-distance between two curves to be meaningful, it is needed that no ambiguity occur in the hierarchic encoding of the patterns in the BSUHWB. This requires that $\arg \min_{\{j, j+1\}} L_{j, j+1}$ is a unique pair $\{j^*, j^* + 1\}$, at Step 3 of each iteration in the BUUHW algorithm. This constraint essentially translates into the idea that the description of the series should not require more than one level change of any given amplitude. This might however be relaxed in some sense, as discussed hereafter.

Applicability for Even Patterns. A series consisting in an exactly symmetric pattern (such as a peak or through) centered on the middle point of the series (i.e. an even pattern with respect to $\frac{N}{2}$) is a typical case where the minimum value of $L_{j, j+1}$ occurs more than once: two level changes of exactly the same amplitude, but opposite signs, encode the left and right part of the pattern.

This ambiguity is solved by defining an orientation for the BUUHW algorithm. For instance, a left orientation is proposed in Section 1.1, meaning that a level change

of given amplitude that is located to the left of another one with the same amplitude is always encoded first. In such a way series with multiple minimum equal values of $L_{j,j+1}$ at certain iterations in the BUUHW algorithm can still be compared using BAGIDIS without ambiguity.

Applicability for Long Series With Numerous Patterns. In principle, the BAGIDIS semi-distance is able to deal with series of any length. However, increasing the length of the series we want to compare often implies also increasing the number of features that appear in the series. If we aim at comparing the fine structure of the series, this can be problematic as feature confusion could occur in the BSUHWB expansions of the different series. This might appear in case some series contain less multiple level changes of a given amplitude than others (absence of a peak in some series of spectrometric dataset, for instance). In that case, it appears useful to make use of a localized version of our semi-distance, when dealing with long series. The principle is as follows. We consider a sliding window of length Δ . For each localization of the window, we obtain the BUUHW expansion of the so-defined subseries and we compute the BAGIDIS semi-distance defined by expression (2.1.9). One can then take the mean of those local semi-distances so as to obtain a global measure of dissimilarity. Depending on the overlap of the successive windows, using this localized version of the algorithm might highly decrease the computational complexity of the algorithm. In the examples of Subsections 1.4.4 and 1.4.5, the overlap has length $\Delta - 1$.

For sure, the choice of the length Δ of the windows should be problem-dependent: Δ defines the range in which a given feature could possibly move along the abscissa axis while remaining identified as a unique feature from one series to another. Experience shows little sensitivity to small variations of that parameter, so that only an order of magnitude of Δ should actually be provided. A possible way to implement an automatic way to estimate this order of magnitude might be to compute the average distance between two consecutive sharp features of the same importance in the curves.

When dealing with long series, one can also be interested to point the abscissas where the series do significantly differ. Using a sliding semi-distance allows us to consider the curve of semi-distances between two series, so that we can automatically identify the abscissas where two series are close to each other (in the sense of the BAGIDIS semi-distance) and the ones where there are more distant. This could be a helpful diagnostic tool for comparing observational curves with a target curve for instance, or for investigating what makes groups of curves actually different.

Two examples involving long series with numerous patterns are provided in Subsections 1.4.4 and 1.4.5.

Applicability for Noisy Series. The method is applicable for noisy series, provided we make use of a suitable weight function that efficiently filters the noise. Several examples in Section 1.4 account for a certain robustness of the method with respect to the presence an additive noise. Nevertheless, an artifact of the method might occur in case of multiple minimum or quasi-minimum values of $L_{j,j+1}$ at certain iterations in the BUUHW algorithm. In case of a symmetric (or quasi-symmetric) peak, for instance, there are two detail coefficients being close in absolute value but of opposite sign. A permutation of the basis vectors might occur in the BSUHWB from one series to the other when the series are noisy, as the noise results in a possible reordering of the amplitudes (in absolute values) of the mentioned coefficients.

Given this, it is clear that a clustering of noisy series where a permutation problem occurs will lead to a spurious distinction into two groups, even when all the curves to be clustered are noisy versions of the same “model curve”. However, this does not make the BAGIDIS semi-distance unable to distinguish if there is or not a systematic effect of a factor even in case such permutations occur. In case there are two groups A and B of such “problematic series” and we want to know if there is a difference between them, this artifact will not invalidate the analysis, in general. If there are no differences between the groups, a clustering will give two groups (due to the permutations occurring in both series A and series B) but each group will contain a mix of series A and B, so that we will conclude that there are no differences in the distributions of groups A and B. On the opposite, if there are differences between groups A and B, a clustering will give four groups, amongst which two are made of series A (the distinction between the groups being an artifact) and two are made of series B. We will thus conclude to an effect of the A-B factor.

This point is illustrated by the following test-scenarios.

- Scenario 1: we consider 2 groups of noisy series (noise $N(0, \sigma = 0.5)$) derived from model A and model B, with A different from B (in this example, we take $A=(0,0,1,0,0,3,5,0,0)$ and $B=\text{rev}(A)$). The BUUHW transforms of those series do suffer from permutations. We compute the dissimilarity matrix between all pairs of series and try to cluster them blindly and provide a multidimensional scaling (MDS) representation of the dataset (Figure 1.8, *top left*). The groups are clearly linearly discriminable, although a spurious difference occur.
- Scenario 2: we perform exactly the same test with $A=B$, so that we should not detect any effect of the model. Results are shown in Figure 1.8, *top right*. Although two groups are distinguished, they contain both series from group A and group B so that we cannot conclude to a difference between the two models.
- Scenario 3: we perform the first test again, with A and B different, with half of the A-curves being shifted by 1 to the left and denoted by a , and with half of

the B-curves being shifted by 1 to the left and denoted by b . We add a Gaussian noise with $\sigma = 0.5$ to all the curves. Results are shown in Figure 1.8, *bottom left*. Again, a distinction between the groups (A+a) and (B+b) is visible, despite the spurious difference due to permutation.

In case of a regression or a classification problem, the key for a successful prediction model is that a given permutation should occur with a sufficiently high probability so that each spurious group contains enough representatives. Using a local approach for prediction (such as nonparametric functional regression (Ferraty and Vieu 2006) or k -nearest neighbors (Hastie et al. 2001), for instance) is then possible. This claim is theoretically supported by the rate of convergence of the nonparametric functional regression estimate based on BAGIDIS (see Chapter 2), obtained under mild conditions. Conform to the above mentioned intuition, the key for the estimator to converge relatively fast is that the probability to find curves in a small ball around the point of prediction is high enough. Chapter 2 provides more details.

Although we strongly believe that the above elements might lead to a proof of the consistency of lots of algorithms used with BAGIDIS, under relatively general conditions, we acknowledge that it might be reassuring to make the BUUHWT algorithm avoid permutations. A possible way would be to render the algorithm insensitive to the noise affecting the patterns, as we know that no permutation occurs for noiseless series provided that we decide for an orientation of the transform. This might possibly be done by testing for the equality of the quasi minimum candidate details, at each step of the algorithm. In case of non-rejection of the equality, the selected candidate-detail is chosen according to the orientation of the transform. Although implementing this approach would limit the risk of permutation, the price to pay would be a possible increase of the risk of not detecting an existing difference. This risk can not be controlled as it is related to the Type II error of the test for equality.

Applicability for Sparse Series. The method is *a priori* applicable for sparse data, as the BUUHWT transform can be obtained for any series, as is the mapping in the $b-d$ plane. What we mean here by sparse data is a signal that can be encoded with a small number of non-zero wavelet coefficients. In the case of unbalanced Haar wavelets, this is a series with a small number of level changes. The well-known “blocks” signal is an example (Fryzlewicz 2007). In such a case, the mapping into the $b-d$ plane and hence the dimensionality of the BAGIDIS semi-distance can be sharply reduced. This is obtained via the definition of a sparse weight function w_k . What we mean here is a weight function of which few of the elements are non-zero. In a prediction or discrimination context, it is believed that a sparse weight function often means a better rate of convergence of the prediction estimator, under some regularity condi-

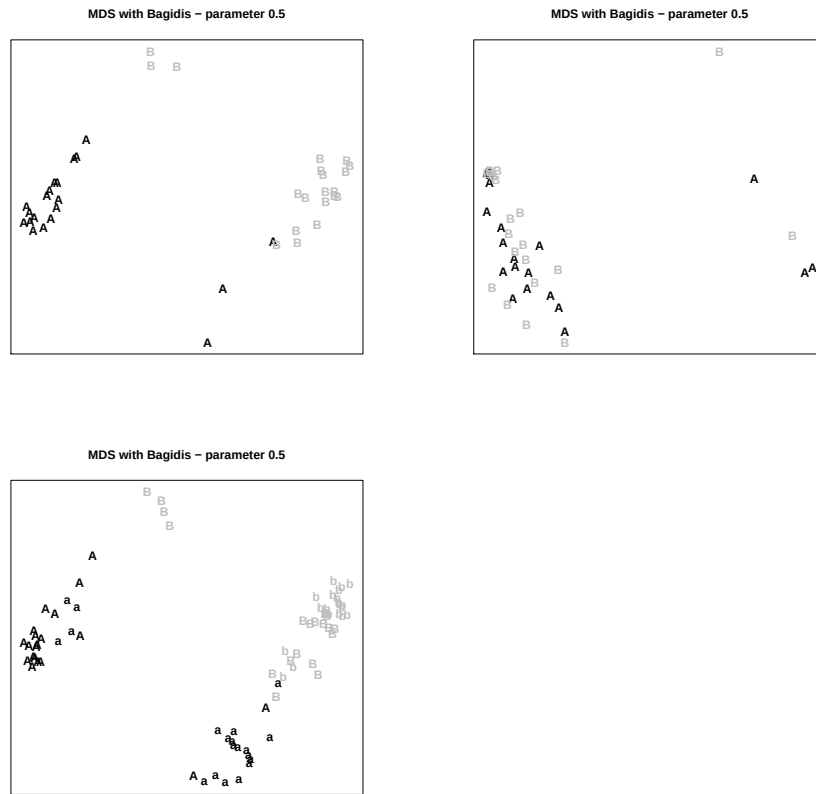


Figure 1.8: Study of the applicability of BAGIDIS for noisy series in case of a possible reordering in the hierarchy of the patterns. *From left to right:* MDS representation of the test dataset resulting from Scenario 1 to Scenario3.

tions. Theoretical support for this claim can be found in Chapter 2 in the context of nonparametric functional regression. An example involving sparse series is provided in Subsection 1.4.2.

Haar wavelet methods for horizontally shifted series

Traditional Haar wavelet decomposition is translation sensitive. This was discussed and illustrated in Figure 1.6, in the context of curve comparison. We briefly review here the implications of translation sensitivity of Haar wavelets from other statistical perspectives, and the existing solutions to overcome this limitation. We start by emphasizing that if a series is expanded in the Haar basis, then the same series arbitrarily shifted also does. Moreover, the energy of the observed curves (i.e. the sum of the squared detail coefficients) is preserved whatever the shift of the series. This is due to the Parseval equality. However, at any given scale, the Haar wavelet expansion of the series only provides information about its structure around dyadic locations. Thereby, certain differences between neighboring zones in the series are never computed, thereby missing potentially interesting information. As a consequence, the Haar wavelet expansion is more efficient at highlighting the structure of a curve in the case where its significant level changes occur at dyadic locations. This translates into a sparser Haar wavelet expansion in this case, for piecewise-constant curves. This feature also impacts the performance of curve estimation via thresholding: the wavelet thresholding estimator (Donoho and Johnstone 1994) does not perform equally well for all shifted versions of a given series. Some spurious artifacts might then occur in the reconstruction of some shifted series (Coifman and Donoho 1995). This is a pseudo-Gibbs phenomenon, resembling the Gibbs oscillations encountered in Fourier analysis.

As far as we know, ways to overcome the problem of translation sensitivity consist into two main proposals, developed in the context of curve estimation. The first proposal is the *unbalanced Haar wavelet* expansion that we use in this work. Fryzlewicz (2007) uses such wavelets in the context of nonparametric estimation of piecewise-constant functions and his proposed estimation procedure is shown mean squared consistent for a wide range of functions. He also uses unbalanced Haar bases as a change point detection tool (Fryzlewicz 2012). The second proposal for rendering Haar wavelet estimation insensitive to shifts is *translation-invariant (TI) wavelet smoothing* (Coifman and Donoho 1995; Berkner and Wells 2002). The idea is as follows: for all possible shifts of a given series, one decomposes the shifted series in the Haar basis, applies a wavelet thresholding, reconstructs the estimated shifted series, and inverts the shift so as to get an estimate of the initial series. All the inverse-shifted estimates are then averaged to define the final estimate of the curve. The resulting estimate is less good than with the traditional Haar thresholding procedure in

case the level changes occur at dyadic locations, but the result highly improves otherwise. The TI estimator leads to relatively smooth estimates of curves. Therefore (Berkner and Wells 2002), the TI algorithm provides nearly exact reconstruction of piecewise linear function, while traditional Haar wavelet thresholding provides this for piecewise constant functions. We are not aware that TI procedures have lead to any curve comparison methodology.

1.3 Statistical Tests and Tools

Dealing with the signatures of the curves in the $b-d$ plane raises new opportunities for testing and investigating datasets. Descriptive statistics and tests proposed hereafter rely on the signatures of the curves in the $b-d$ plane and on the BAGIDIS semi-distance. In such a way, they share with our semi-distance the ability to efficiently apply to datasets of curves with horizontally- and vertically-varying sharp local patterns. All the tests and tools described in this Subsection are illustrated on real and simulated examples in Section 1.4.

1.3.1 Descriptive Statistics in the $b-d$ Plane

Given a set of n curves, we propose to define the average curve in the $b-d$ plane as

$$\bar{\mathbf{x}}^{(b-d)} \equiv (\bar{\mathbf{b}}, \bar{\mathbf{d}}) \quad \text{where} \quad \bar{b}_k = \frac{1}{n} \sum_{i=1}^n b_k^{(i)} \quad \text{and} \quad \bar{d}_k = \frac{1}{n} \sum_{i=1}^n d_k^{(i)}. \quad (1.3.1)$$

At each rank k , $\bar{\mathbf{x}}_k^{(b-d)}$ is the center of gravity of the points $\{(b_k^{(i)}, d_k^{(i)})\}_{i=1 \dots n}$ characterizing the averaged curves at rank k . The signature $\bar{\mathbf{x}}^{(b-d)}$ of the average curve can then be easily displayed and interpreted in the $b-d$ plane. By construction, it allows for keeping track of possibly misaligned level changes defining sharp patterns, which usual *point-to-point* estimates of an average curve $\bar{\mathbf{x}}(t) = (1/n) \sum_{i=1}^n \mathbf{x}^{(i)}(t)$ would miss, for the reasons discussed in Subsection 1.2.3.

We have to notice that there is no obvious way to get back from that average signature in the $b-d$ plane to the average curve in the initial units of measurements, as the average values of the breakpoints are not necessarily falling on a point of discretization, and because nothing prevents a given value of the breakpoint to appear twice in the series $\bar{\mathbf{b}}$. Nevertheless, definition (1.3.1) is useful for interpretative purposes in the $b-d$ plane and, most importantly, it allows for quantifying the variability of a dataset

in line with the BAGIDIS semi-distance:

$$V^{2(b-d)} = \frac{1}{n} \sum_{i=1}^n d_{p\lambda}^{\text{BAGIDIS}}(\mathbf{x}^{(i)}, \bar{\mathbf{x}}^{(b-d)}) = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^{n-1} w_k \left\| \mathbf{y}_k^{(i)} - \bar{\mathbf{y}}_k \right\|_{p\lambda},$$

with $\bar{\mathbf{y}}_k = (\bar{b}_k, \bar{d}_k)$ and $\mathbf{y}_k^{(i)} = (b_k^{(i)}, d_k^{(i)})$. This measure of the variability is the key for defining an ANOVA-like test within the BAGIDIS framework.

1.3.2 An ANOVA Test Within the BAGIDIS Framework

Suppose we have n curves coming from G groups, whose empirical group-mean curves have been estimated as $\bar{\mathbf{x}}^{(d-d,g)} \equiv (\mathbf{b}^{(g)}, \mathbf{d}^{(g)})$, $g = 1 \dots G$, and suppose the global empirical mean curve has been estimated as $\bar{\mathbf{x}}^{(b-d,0)} \equiv (\mathbf{b}^{(0)}, \mathbf{d}^{(0)})$. Then if the distances of each group of curves around their group mean are independent and normally distributed with equal variance in each group, we can compute the ANOVA F-Ratio as

$$F = \frac{\frac{1}{G-1} \sum_{g=1}^G n_g d_{p\lambda}^{\text{BAGIDIS}^2}(\bar{\mathbf{x}}^{(b-d,g)}, \bar{\mathbf{x}}^{(b-d,0)})}{\frac{1}{n-G} \sum_{g=1}^G \sum_{i=1}^{n_g} d_{p\lambda}^{\text{BAGIDIS}^2}(\mathbf{x}^{(g,i)}, \bar{\mathbf{x}}^{(b-d,g)})}, \quad (1.3.2)$$

where n_g is the number of curves in group g , $g = 1 \dots G$, and $\mathbf{x}^{(g,i)}$ is a curve indexed i in the g^{th} group. By construction, this ANOVA F-ratio follows an $F_{G-1, M-G}$ distribution under the null hypothesis that all the group-means are equal.

Note that we do not want to claim here that the BAGIDIS semi-distances of each group of curves around their mean are necessarily normally distributed with equal variance in each group. This hypothesis has actually to be tested so as to check if we can consider satisfying the conditions for the test, in the same way as it has to be checked before performing a usual ANOVA. It is clear that the assumption of normality and homoskedasticity are quite strong. As for the usual ANOVA, Box-Cox transforms of the semi-distances might be used to approach those conditions. Moreover, nonparametric extensions of the usual ANOVA F-test (see Kruskal and Wallis (1952), e.g.) might probably be transposed into our BAGIDIS framework. However, this is beyond the scope of the present Chapter. The point here is to emphasize the fact that such a test for differences between groups is possible in the BAGIDIS framework.

1.3.3 A Geometrical Test for the Relative Positions of two Groups in Space.

Another possibility to diagnose the way groups of curves do differ is to directly look at the relative positions of the clouds of points they form in $\mathbb{R}^N$, according to a given distance or semi-distance d , the BAGIDIS semi-distance for instance. The idea is as follows. Suppose we aim at comparing curves of group 1 ($G1$) and group 2 ($G2$). First, compute pairwise cross-distances $d(G1, G2)$ between curves of group 1 and group 2. Then compute pairwise cross-distances $d(G1, G1)$ within the curves of group 1 and, separately, pairwise cross-distances $d(G2, G2)$ within the curves of group 2. Then, test for the equality of the means of the distances intra-group 1 and inter-group 1 and 2, and for the equality of the means of the distances intra-group 2 and inter-group 1 and 2, versus the alternative that intra-group distances are lower than inter-group distances:

$$\begin{cases} \textbf{Test 1} : H_0 : d(G1, G1) = d(G1, G2); H_1 : d(G1, G1) < d(G1, G2). \\ \textbf{Test 2} : H_0 : d(G2, G2) = d(G1, G2); H_1 : d(G2, G2) < d(G1, G2). \end{cases} \quad (1.3.3)$$

If the distances are distributed normally about their group-mean, are independent and if they have equal variance in each case, this can easily be done using a student t-test. As for the ANOVA-like F-test, those conditions have to be checked before the test is performed. Note that the assumptions of equality of variances can be relaxed by using the Welch adaptation of the t-test (Welch 1947). Interpretation of the results is then direct. If the intra-group mean distance within group 1 (resp. group 2) is significantly lower than the inter-group mean distance, it means either that the two groups are distinct or that group 1 (resp. group 2) is included within group 2 (resp. group 1) with a smaller variability. Combining the results of both tests allows one to deduce the relative positions of group 1 and group 2, as illustrated in Figure 1.9. Only if both t-tests significantly reject their null hypotheses are the two groups statistically different in mean. If only one of the t-test rejects its null hypothesis, it means that both groups have approximately the same mean, but a different variability. One group is then “included” within the other one.

1.4 Data Analyses Using Bagidis

Real and simulated applications are proposed in this Section. A simple simulated example of supervised classification is proposed in Subsection 1.4.1. It aims to get an insight into the behavior of the BAGIDIS semi-distance. In particular, attention is paid

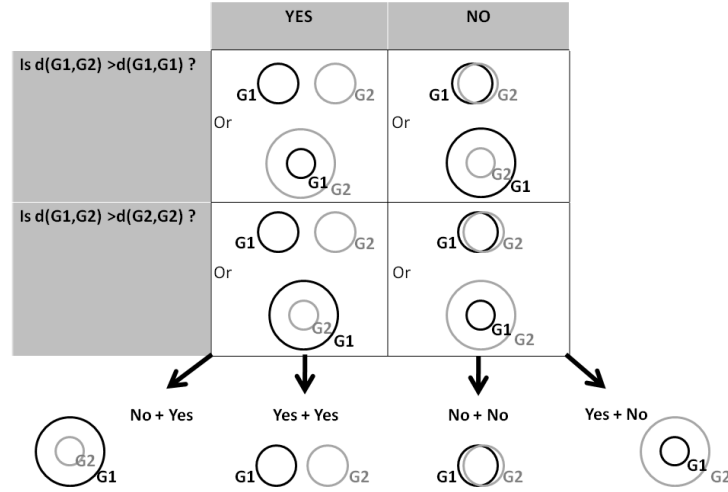


Figure 1.9: *Left: Summary table for the interpretation of the double t-test for the relative position in space of two groups of curves $G1$ and $G2$, according to a given distance or semi-distance.*

to its sensitivity with respect to the noise level. Another example involving sparse curves is then provided in Subsection 1.4.2. A simulated example of ANOVA F-test within the BAGIDIS framework is presented in Subsection 1.4.3. Two real datasets are then investigated. Both involve data with important sharp features. The first dataset, in Subsection 1.4.4, is made of spectrometric curves that have to be compared very finely, as the differences between them are very small. Our goal is to diagnose whether those small observed differences are due to noise or to systematic effects. Classical competitors show their limits in this problem. BAGIDIS reveals structure that was undetected. Afterwards, in Subsection 1.4.5, we study the similarity of the dynamics of solar irradiance time series. This dataset is highly noisy. Measurements have large time step. Physical knowledge of the process tells us that this feature does not allow for capturing significant shifts of the patterns. The goal is to illustrate how BAGIDIS behaves in a context where its specific ability to cope with horizontal variations of the patterns is not needed. This example acts as a test of robustness. BAGIDIS proves consistent with competitors.

1.4.1 A Simulated Example of Supervised Classification

The following test is proposed: 10 mother series are defined (Figure 1.10, *left*). Each has length 11. The features they are made of have height from 10 to 100. Each of those 10 mother series leads to a family of 1000 simulated noisy series as follows:

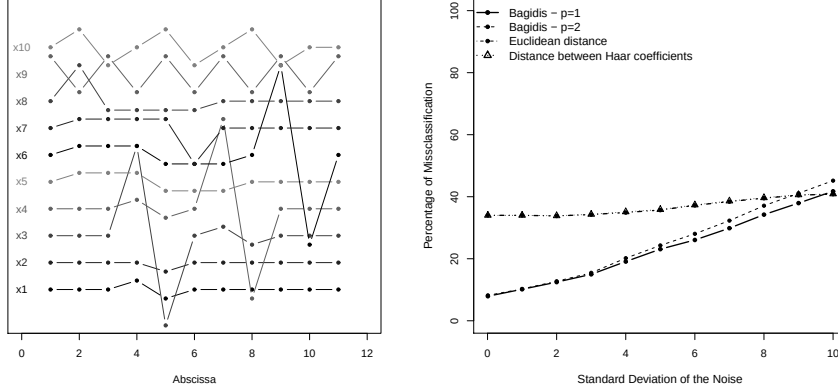


Figure 1.10: *Left*: The 10 mother series of the supervised classification example. *Right*: Percentages of misclassification as a function of the standard deviation of the Gaussian noise affecting the simulated series. Results for $d^{\text{EUC}}_{\text{L}}$ and d^{HAAR} are essentially superimposed. Results for $d^{\text{BAGIDIS}}_{1,0.5}$ are slightly better than for $d^{\text{BAGIDIS}}_{2,0.5}$, however close.

the mother series is shifted one step to the right in 20% of the simulated series; it is shifted one step to the left in another 20%, it is amplified by a factor 0.75 and 1.25 respectively for two other groups of 20% of the series; it is unchanged for the remaining 20%. Each simulated series is affected by an additive Gaussian noise with standard deviation σ . The noise level varies from $\sigma = 0.01$ to $\sigma = 10$, which is the height of the smallest peak in the noise-free data and 10% of the height of the highest one. We evaluate the dissimilarities of those 10000 series to the 10 mother series, using four different distances or semi-distances: the BAGIDIS semi-distances $d^{\text{BAGIDIS}}_{1,0.5}$ and $d^{\text{BAGIDIS}}_{2,0.5}$ used with a partial distance defined as 1-norm ($p = 1$) or 2-norm ($p = 2$) respectively and without scaling ($\lambda = 0.5$), the Euclidean distance $d^{\text{EUC}}_{\text{L}}$ (l_2 distance in the original coordinates of the series) and the l_2 distance between the expansion coefficients of the series into the usual Haar basis d^{HAAR} (series are padded with 0 up to length 2^4 in order to allow for such an expansion). The series are then classified as being part of the closest family. The percentage of misclassification is our criterion for comparing the quality of the proposed dissimilarities. Results in Figure 1.10, *right*, show that $d^{\text{BAGIDIS}}_{1,0.5}$ and $d^{\text{BAGIDIS}}_{2,0.5}$ performs very well compared to competitors, up to the maximum level of noise. Their performances are very close, although slightly better for $d^{\text{BAGIDIS}}_{1,0.5}$. Results for $d^{\text{EUC}}_{\text{L}}$ and d^{HAAR} are essentially superimposed.

Table 1.2 indicates individual misclassification rates for each family of series for three noise levels. For the sake of clarity, only rates for $d^{\text{BAGIDIS}}_{1,0.5}$ and $d^{\text{EUC}}_{\text{L}}$ are provided, those rates being essentially similar to those obtained using $d^{\text{BAGIDIS}}_{2,0.5}$ and d^{HAAR} re-

spectively. Most misclassification rates of $d_{1,0.5}^{\text{BAGIDIS}}$ start very low and increase with the noise level for each family. Results for d^{EUCL} seem mostly unaffected, having a constant high 20% or 40% rate for most series. Those specific values are related to the percentage of misaligned series (20% to the left and 20% to the right), indicating that d^{EUCL} fails at correctly dealing with misaligned features, as discussed in Section 1.2.3. $d_{1,0.5}^{\text{BAGIDIS}}$ has significantly better misclassification rates, except for families derived from x2 and x9 (in case of $\sigma = 0.01$ or $\sigma = 3$), as well as families derived from x1 and x5 (in case of $\sigma = 6$). Interpretations are as follows. Series generated from x1 and x2 are frequently mixed up when the noise level increases - this is related to the small amplitude of their significant patterns as compared with random patterns due to noise. This makes it difficult for the BUUHW algorithm to provide for a hierarchy of the features. A contrario, this fact accidentally leads the x2-family to be best identified by d^{EUCL} as this distance actually misses the significant pattern and identifies the series as being part of the “no-feature” family. Series generated from x9 are essentially oscillating and tend to be mixed up with series x10, which has also an oscillating pattern, although asymmetric. Conversely, the asymmetry of the oscillating patterns of the x10-family is correctly captured so that this family is well-identified.

Percentages of misclassification						
Family	$\sigma = 0.01$		$\sigma = 3$		$\sigma = 6$	
	$d_{1,0.5}^{\text{BAGIDIS}}$	d^{EUCL}	$d_{1,0.5}^{\text{BAGIDIS}}$	d^{EUCL}	$d_{1,0.5}^{\text{BAGIDIS}}$	d^{EUCL}
x10	0.0	40.0	7.7 (4.9 to x5)	40.0	26.4	40.0
x9	54.3 (46.2 to x10)	40.0	55.3 (42.8 to x10)	40.0	56.5 (38.3 to x10)	40.0
x8	10.5 (10.5 to x5)	40.0	10.0 (8.2 to x5)	40.0	15.0 (10.5 to x5)	40.0
x7	0.0	40.0	0.3 (0.3 to x5)	40.0	10.7	40.0
x6	3.5 (3.5 to x4)	40.0	9.9 (9.9 to x4)	40.0	10.2 (10.1 to x4)	40.0
x5	0.0	20.0	6.3 (3.6 to x1)	16.5	25.8	24.2
x4	0.0	40.0	0.7 (0.7 to x6)	40.0	8.4 (6.4 to x6)	40.0
x3	0.0	40.0	0.0	40.0	2.2 (1.5 to x4)	40.0
x2	10.5 (10.5 to x1)	0.0	38.4 (38.4 to x1)	4.3	45.9 (34.4 to x1)	18.5
x1	0.3	40.0	21.1 (18.1 to x2)	41.7	59.4 (38.5 to x2)	49.5

Table 1.2: Individual misclassification rates for the simulated example of Subsection 1.4.1. Percentages of misclassification for each family of series are presented for $d_{1,0.5}^{\text{BAGIDIS}}$ and d^{EUCL} , for three different values σ of the standard deviation of the vertical noise. When confusion occurs in the identification of the mother series, one concurrent series is mainly responsible for that. It is indicated in parentheses.

1.4.2 A Simulated Example With Sparse Data

In this Subsection, we provide for an example involving sparse curves and investigate its sensitivity to an additive noise, as compared with competitors. The dataset consists in step signals of length 31 with a moving jump between the right, positive, part at level 1 to the left, negative, part at level -1 . An additive Gaussian noise is superposed to the

signals. The signals having a sparse encoding in the BSUHWB, with only one non-zero coefficient, we make use of a weight function with $w_1 = 1$ and $w_k = 0$, $k = 2 \dots 29$. We fix $\lambda = 1$ as we know that the problems concerns the capture of an horizontal variation.

For each tested level of noise ($\sigma \in (0, 0.1, 0.25, 0.5, 0.75, 1)$), we generate a dataset of 30 curves, with one curve for each location of the jump. Figure 1.11 shows example series out of the datasets generated for each level of the noise. We compute the dissimilarity matrix for the datasets using successively the BAGIDIS semi-distance, the Euclidean distance, the Euclidean distance between the estimated derivatives of the curves and the distance between the coefficients of the Haar expansions. A representation of those dissimilarity matrices is provided in Figure 1.12. We observe that the BAGIDIS semi-distance is the only one which essentially preserve the empirically expected diagonal structure of the dissimilarity matrix up to the maximal level of noise tested.

1.4.3 A Simulated Example of ANOVA F-test Within the BAGIDIS Framework

A BAGIDIS-ANOVA F-test (equation (1.3.2)) is performed on simulated datasets with 3 groups of 50 curves, for different levels of noise. Curves are generated as follows. We have 3 mother series (Figure 1.13, *left*). Each series is repeated 50 times, and affected by an additive Gaussian noise of standard deviation σ . It is also affected by a horizontal shift, randomly chosen amongst -1, 0 and 1. The standard deviation of each mother series is 1. A BAGIDIS-ANOVA F-test is performed on those 150 curves and the p-value is noted. This test is performed for σ varying from 0.5 to 1.5, which is 50% of the height of the highest peak. The process is repeated 100 times for each value of σ . Figure 1.13, *middle left*, shows the statistical evolution of the significance ($1-p\text{-value}$) of the test as a function of the noise level. The difference between the groups of curves is detected more than 95% of the times when σ is lower than 90% of the standard deviation of the mother curves (signal), it is detected more than 75% of the times when the signal to noise ratio of the standard deviations equals 1, and about 50% of the times when the signal to noise ratio is 1.1. The proportion of detection of a difference decreases then rapidly. This is the behavior we expect for such a statistical test. Assumptions for normality and equal variances have been checked for those tests and, on average, there is no evidence for rejection (Figure 1.13, *middle right* and *right*). Finally, we note that ANOVA tests are quite robust to non-normality (Geng et al. 1982) and that some procedures have been proposed in order to relax the

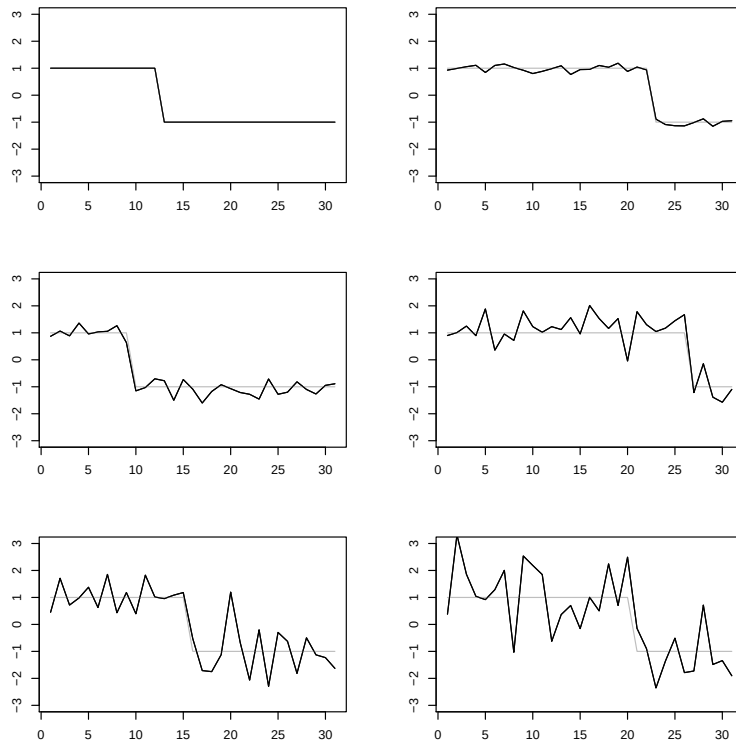


Figure 1.11: **Illustration of the simulated data used in Subsection 1.4.2.** From left to right and top to bottom: simulated series with standard deviation fixed at $\sigma = 0, 0.1, 0.25, 0.5, 0.75$ and 1 respectively, and different location of the jump. The true underlying signal is plotted in gray.

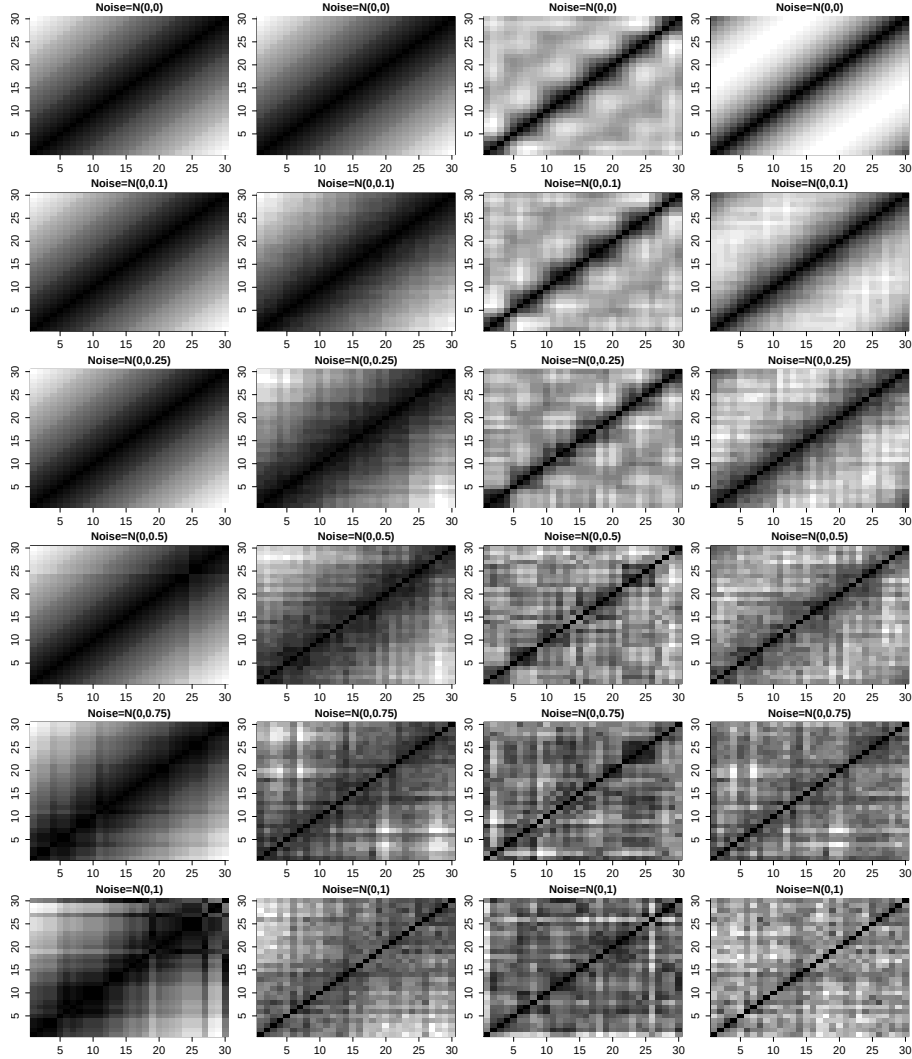


Figure 1.12: Images of the dissimilarity matrices between possibly noisy sparse curves with a moving jump from level 1 to level -1 . The indices identifying the curves in the representation of the dissimilarity matrices encode the location of the jump. Black indicates a distance 0, white indicates the largest distance. Images of the dissimilarity matrices are provided for different distances or semi-distances and different level of noise. *From left to right:* the BAGIDIS semi-distance, the Euclidean distance, the Euclidean distance between the estimated derivatives of the curves, the distance between the coefficients of the Haar expansions. *From top to bottom:* The standard deviation of the noise added to the curves is successively fixed at $\sigma = 0, 0.1, 0.25, 0.5, 0.75$ and 1.

assumption of equal variance in each group (Bishop and Dudewicz 1978; Weerahandi 1995), which could easily be applied here.

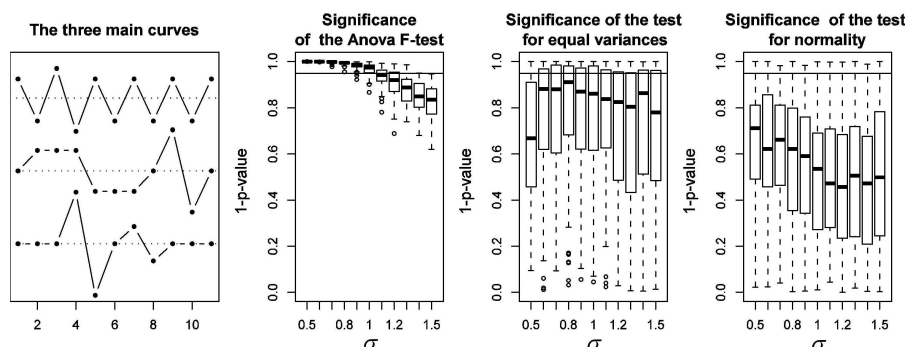


Figure 1.13: From left to right: The three mother curves for the simulation study. Boxplots of the significances ($= 1 - p\text{-values}$) of the BAGIDIS ANOVA F-test for each tested value of the noise. Boxplots of the significances of the Bartlett test for equal variances amongst the groups. Boxplots of the significances of the Jarque-Bera normality test for the residuals.

1.4.4 Searching for Factors Effects on a Set of Spectro-metric Curves

Dataset and Experimental Design. The dataset we consider is made of 24 spectra of biological serum (Figure 1.14). It has been collected (Laboratoire de Chimie Pharmaceutique, Université de Liège) in the framework of a study of the reproducibility of the acquisition and preprocessing of ^1H NMR data in the field of metabonomics (Dubois 2009; Rousseau 2011). The spectra have been acquired using ^1H NMR spectroscopy on two different samples of serum, on three different days and with two different delays after the samples have been unfrozen. All combinations of the levels of those three factors have been tested and repeated twice. The 24 resulting spectra have all been similarly preprocessed using the automated tool BUBBLE (Vanwinsberghe 2005), followed by a given sequence of actions: suppressing the peaks corresponding to lactate (1.80 to 1.98 ppm) and water (4.57 to 4.98 ppm) which are not relevant for this study of reproducibility, setting all negative values to zero and normalizing the data so that the area below the spectrum is 1. All spectra are made of 750 data points.

Goal of the Analysis. We aim at determining whether the experimental design affects the spectra or not. In case it does, we want to investigate the main effects of the

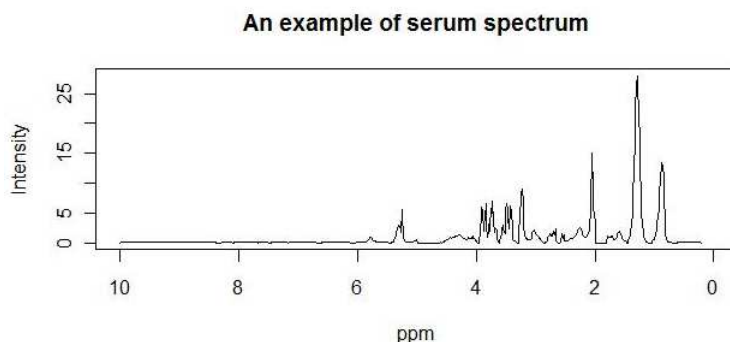


Figure 1.14: An example of preprocessed serum spectrum, obtained using ^1H NMR spectroscopy, with a pulse sequence called CPMG that leads to the suppression of the proteinic markers in the spectra - otherwise, proteins mask the signals of other molecules of interest: the metabolites (Laboratoire de Chimie Pharmaceutique, Université de Liège).

design, in view of diagnosing which parts of the spectra are concerned by the changes and how. This should help biochemists to better understand and control the sources of variations of their ^1H NMR studies of serum spectra.

Visualization of the Dataset. We build multidimensional scaling representations (Cailliez 1983; Cox and Cox 2008, for instance) of the dataset, for several kind of pairwise distances and semi-distances between the spectra. For our purpose, we will only mention that multidimensional scaling is a projection technique that aims at preserving distances between the observations in the dataset, so that the proximities of the data in the plane of projection can be interpreted - up to a certain degree - as “real” proximities of the data according to the chosen distance. Multidimensional scaling used jointly with the Euclidean distance d^{EUCL} is nothing else than the projection of the dataset on the first plane of a principal component analysis. Figure 1.15, *top left*, shows the result we obtain using the BAGIDIS semi-distance $d_{1,0.5}^{\text{BAGIDIS}}$ (i.e. no scaling along any axis) with a sliding window of length $\Delta = 25$. The group of spectra collected on *day A* is markedly separated from the other ones. Within *day A*, an effect of the sample is observed. No effect is detected for the time elapsed after the sample has been unfrozen. Data corresponding to repetitions of the same combinations of the levels of the factors are most of the times projected close to each other. An effect of our experimental design is thus clearly seen, except for the time after unfreezing. Principal component analysis (Figure 1.15, *top right*) and a functional semi-distance based either on the comparison of the first derivatives (Figure 1.15, *center left*) or on

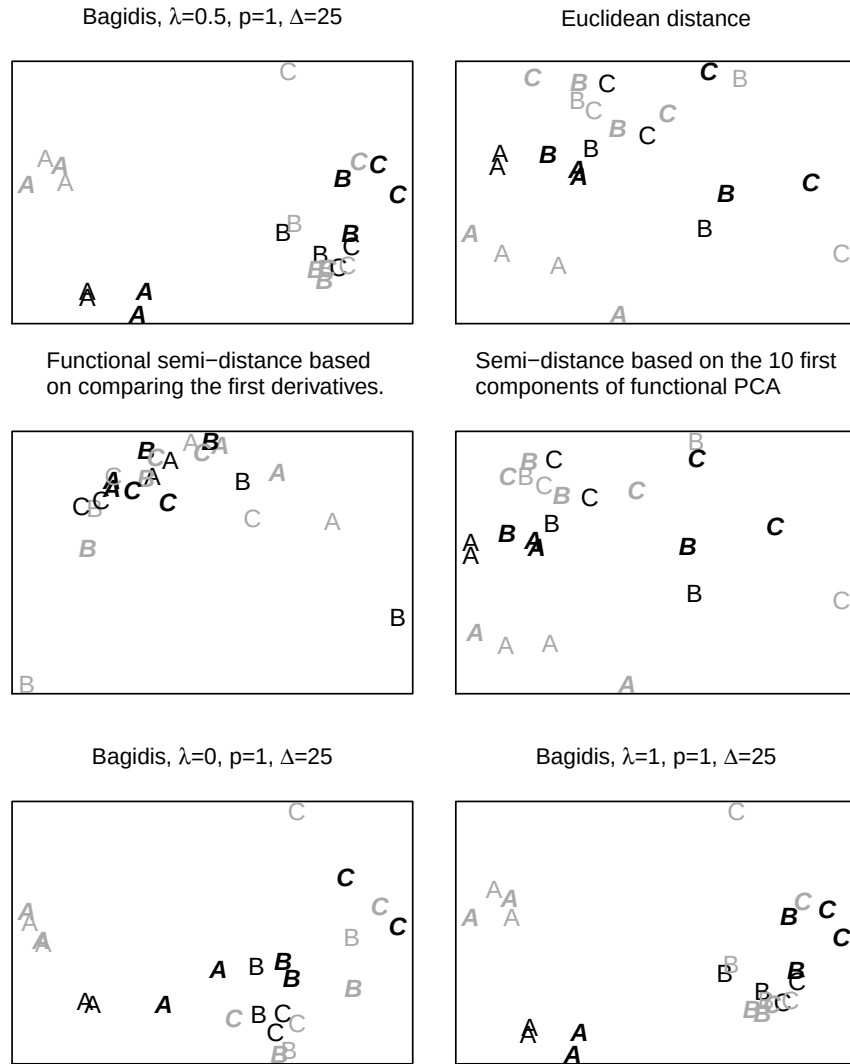


Figure 1.15: Multidimensional scaling representations of the spectrometric dataset, based on several matrices of distances or semi-distances. The projected spectra are labeled according to the levels of the three factors of the experimental design: they are denoted A, B or C depending on the day of collection of the spectrum, they are colored in black or gray depending on the serum sample, and written in normal or slanted font depending on the time waited after the sample has been unfrozen. Each label appears twice in the projections of the dataset as it corresponds to the repetitions of the acquisition of the spectra in the same experimental conditions.

the first functional principal components (Figure 1.15, *center right*) fail to detect this effect.

We also compute the pairwise semi-distances between the spectra using a sliding BAGIDIS semi-distance successively with a balance parameter $\lambda = 0$, in view of investigating the effect of differences in the details coefficients, and with $\lambda = 1$, so as to highlight the effect of the differences in the breakpoints. Multidimensional scaling representations of the dataset with those semi-distances are presented in Figure 1.15, *bottom left*, and Figure 1.15, *bottom right*, respectively. The projection we obtain with $\lambda = 1$ is clearly similar to our first result using BAGIDIS (Figure 1.15, *top left*, with $\lambda = 0.5$), while the projection we obtain with $\lambda = 0$ looks less clear. This highlights the fact that the effects of the design that were detected in the dataset are mainly related to modifications of the shapes or locations of the features (as encoded by the values of the breakpoints) rather than to modifications of the amplitudes. This is consistent with the fact that the analysis with usual methodologies, and in particular with principal component analysis, is far less clear for this dataset and is not able to detect those systematic effects.

Study of the Effect of the Day on the Spectra. As pointed out by the previous analysis, the major source of variation for our spectra is the day at which the spectra have been collected. An interesting question is then to identify which parts of the spectra are mainly affected by that change. To that purpose, we consider explicitly the vectors of sliding semi-distances, and we test for differences between them separately for each component of the vectors of distances, i.e. for each sliding segment of the spectra.

First, we perform a set of BAGIDIS ANOVA F-tests (equation (1.3.2)) for the differences between spectra as a function of their day of collection. We use a BAGIDIS sliding semi-distance with parameters $\Delta = 25$, $p = 1$, $\lambda = 1$. There is one F-test computed for each sliding segment of the spectra. Hypotheses for normality and equal variances of the residuals in each group are checked and cannot be rejected on a 5% significance level. Figure 1.16, *top*, shows the resulting vector of significances (i.e. 1- p-values of the tests). However, as this test for the locations of the differences involves multiple comparisons, a consecutive Bonferroni correction of the p-values is performed so as to control the global confidence level of the test. We see in Figure 1.16, *bottom*, that the so-corrected test lacks the power to detect differences between the spectra, except at one location, just before 2 ppm, which reflects a difference in the removal of the peak of lactate. This lack of power is not surprising as our global test is made of 726 simultaneous comparisons involving only 24 values each, which is rather small. Then, although we cannot consider simultaneously the uncorrected sig-

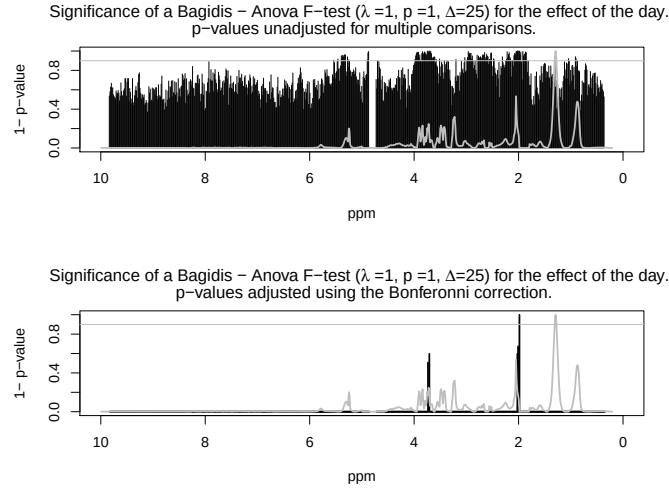


Figure 1.16: *Top*: Significances ($= 1 - p\text{-values}$) of each individual BAGIDIS ANOVA F-test for the effect of the day on the spectra. *Bottom*: Significances of those BAGIDIS ANOVA F-test corrected for 726 multiple simultaneous comparisons (Bonferroni correction). In both cases, one typical spectrum is superimposed to the graphs.

nificances of the F-tests in Figure 1.16, *top*, it is worth to have a look at it. Some parts of the spectra have clearly a higher significance for the differences between the groups. They essentially coincide with the parts detected significantly different between *day A* and *day C* by the *geometrical t-test for the relative positions of the groups* (Figure 1.17), which is discussed hereafter. It indicates that, despite their lack of power in case of multiple comparisons, the results of our F-tests are consistent.

Afterwards, we investigate the pairwise differences between the spectra collected each day. To this aim, we use sliding geometrical t-tests for the relative positions of two groups in space (expression (1.3.3)) applied with the BAGIDIS semi-distance. Technical details for those tests are presented in Appendix A. As an example, we present here our results for the comparison of the spectra collected on *day A* and those collected on *day C*, and whose differences are the major effect of the experimental design. Figure 1.17, *row 2*, shows the parts of the spectra where the shapes or locations of the peaks measured on *day A* are significantly different from the shapes or locations of those peaks measured on *day C*. Most of the differences coincide with the one detected by a classical Welch t-test for the equality of the means applied on the same sliding windows (Figure 1.17, *row 1*). There are also some differences that are not detected by the classical Welch t-test but only by the test using BAGIDIS: the rise of the peak around 5.3 ppm, or the shape of the small peak around 2.2 ppm, e.g. Some differences high-

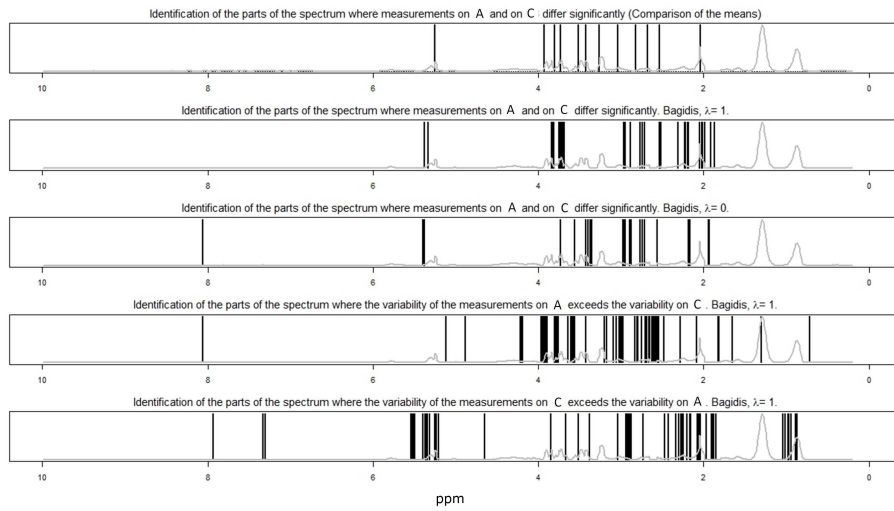


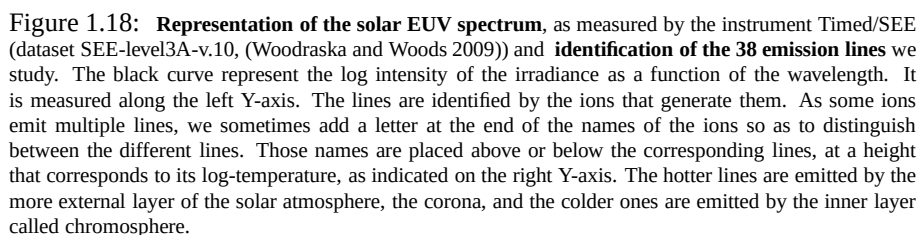
Figure 1.17: Results of multiple geometrical t-tests for the relative geometrical positions of the groups of spectra measured on *day A* and on *day C*. On *row 1*, *row 2* and *row 3*, vertical lines identify the parts of the spectrum that are detected significantly different (with a confidence level of 95%, after Bonferroni correction) between the curves of *day A* and *day C*, as computed respectively using a classical Welch t-test for the equality of the means, a test for the respective positions of the groups of spectra based on a BAGIDIS semi-distance with $\lambda = 1$ (test for the effect of the differences in the breakpoints), and the same test with $\lambda = 0$ (test for the effect of differences in the detail coefficients). On *row 4* (resp. *row 5*), vertical lines identify the parts of the spectrum where the variability of the measurements on *day A* (resp. *day C*) exceeds significantly their variability on *day C* (resp. *day A*), with a confidence level of 95%, after Bonferroni correction, based on a BAGIDIS semi-distance with $\lambda = 1$. One typical spectrum is superimposed to the graphs.

lighted by the Welch t-test are not detected anymore by our test with $\lambda = 1$ (involving the effect of differences in the breakpoints only). If they are detected by our test with $\lambda = 0$ (involving the effect of the differences in the details coefficients only), it means that the measured differences concern only the intensities of the peaks and not their shape. This is the case for the small double peak around 3.45 ppm for instance. If they are not detected neither by the test with $\lambda = 1$ nor by the one with $\lambda = 0$, it could indicate that the difference measured by the t-test are related to very small shifts of some peaks, that are not seen significant anymore when we take this possibility of shifting actually into account, as BAGIDIS does. This is in particular the case for the peak at 3.2 ppm that seems to remain unchanged in shape or intensity but was possibly only very slightly shifted, from one day to another. Finally, rows 4 and 5 of Figure 1.17, identify the parts of the spectra where the measurements are more variable on day A or on day C respectively, around approximately the same mean in both groups.

Conclusion of the Spectrometric Analysis. The BAGIDIS methodology performs well on this dataset of spectrometric curves. We highlight global effects of our experimental design that are difficult to detect using principal components analysis or usual functional methodologies. Moreover, we diagnose the parts of the spectra which are affected by the design, and we get an insight in the way they are affected (changes in amplitudes, changes in shapes, increases in the variability). Taking into account the horizontal variations of the sharp patterns in the curves reveals thus useful.

1.4.5 Clustering Highly Noisy Time Series According to their Dynamics

Context and Goal of the Analysis. Our second analysis concerns the solar atmosphere. We consider time series of solar irradiance measurements, i.e. the temporal evolution of the energy flux, at distinct extreme ultra-violet (EUV) spectral emission lines (Figure 1.19). Each emission line is emitted by a given ion and each ion only exists at a well defined temperature in the solar atmosphere. As the solar atmosphere is globally layered according to the temperature, we can deduce the altitude at which the line was emitted. Those emission lines and their related temperatures are identified in the solar EUV spectrum in Figure 1.18, each of the lines being labeled by the ion that generates it. The hotter lines are emitted by the corona (the most external layer of the solar atmosphere) and the colder ones are emitted by the chromosphere (inner layer of the solar atmosphere). Fluctuations of the emission of all the lines from a given altitude (i.e. at a given temperature) should thus be similar as they correspond to similar physical processes affecting the emission. Our goal is thus to blindly classify the time series according to their dynamics and see if the resulting classification leads to



The Dataset. The 38 time series we consider have been measured by the Solar EUV Experiment (SEE) on board the Thermosphere Ionosphere Mesosphere Energetics and Dynamics (TIMED) spacecraft (Woods et al. 2005). They consist of 344 measurements with a time interval of about 97 minutes, from October 14, 2003 to November 10, 2003 and are part of the SEE-Level3A-v.10 dataset (Woodraska and Woods 2009). Their emitting ions, wavelengths and temperatures of emission can also be found in the SEE-Level3A-v.10 dataset. All the series have been standardized so as to focus on their dynamics. The time series are affected by a lot of fluctuations, most of which are non-informative experimental noise. Similarities amongst the series have thus to be found in their local mean trends and in the sharp patterns related to sudden important increases of emission corresponding to solar eruptions (flares). As the time step of our analysis is not fine enough, we do not expect to capture any significant time shift between the flares in cold and hot lines - physically this would correspond to an energy transfer amongst the layers of the solar atmosphere. For our purpose, those considerations mean that we do not expect breakpoint differences to play a significant role in the BAGIDIS distance for discriminating the series. The point of this second real data example is to check how the BAGIDIS methodology behaves in a difficult

noisy context where considering horizontal shifts of the patterns in the series is not actually required.

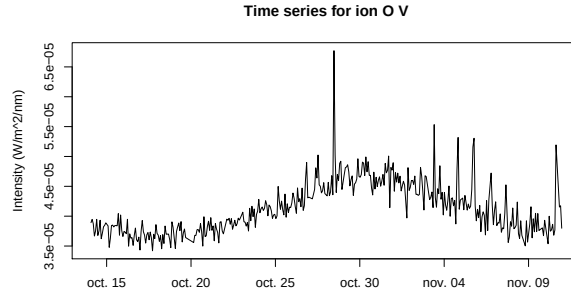


Figure 1.19: Irradiance time series for the spectral line O V at 62.97 nm from October 14 to Novembre 10, 2003, as measured by the instrument Timed/SEE (dataset SEE-level3A-v.10, (Woodraska and Woods 2009)).

Clustering of the Series. We compute the pairwise dissimilarities between the time series using several kind of distances or semi-distances. On that basis, we cluster the series using a Ward's hierarchic agglomerative algorithm (Kaufman and Rousseeuw 1990, for instance). Figure 1.20, *top right* and *top left*, shows the dendrogram that we obtain using the Euclidean distance and a functional PCA-based semi-distance respectively. If we compute such dendrograms for the BAGIDIS semi-distance, with $\Delta = 20$, $p = 1$ and a balance parameter λ decreasing from 1 to 0, we observe that the dendrograms become more and more structured as λ goes to 0, with tidier groups of series. We are thus lead to select $\lambda = 0$ - i.e. taking into account the effect of the detail coefficients only. This is consistent with the fact that we did not expect time shifts of the patterns to be significant in that dataset. The resulting dendrogram is shown in Figure 1.20, *bottom*. As expected, the three clustering methods are consistent with each other, and with a physical grouping according to the temperature (as given in Woodraska and Woods (2009)): the hot lines emitted by the corona are close to each other and are clustered within one single group in the case of BAGIDIS and the functional PCA and in two groups for the Euclidean distance. This indicates that BAGIDIS and the functional PCA are probably slightly better than the Euclidean distance. The colder but highly intense Lyman- α line (denoted H I d) is joint with the hot lines for the Euclidean distance and the BAGIDIS semi-distance. The quite tenuous line Mg X is correctly clustered with the hot lines when using functional PCA, and linked with spectrally close lines in the two other cases, possibly due to some contribution of the spectral continuum. Some groups of colder lines emitted by the chromosphere are

clearly detected in every case. Amongst those groups, the individual proximities of the spectral lines are similar whatever the dissimilarity measure we use: Si II a and b with C I a; O II, O III and O IV b; Si IV a, Si IV b and C IV, etc.

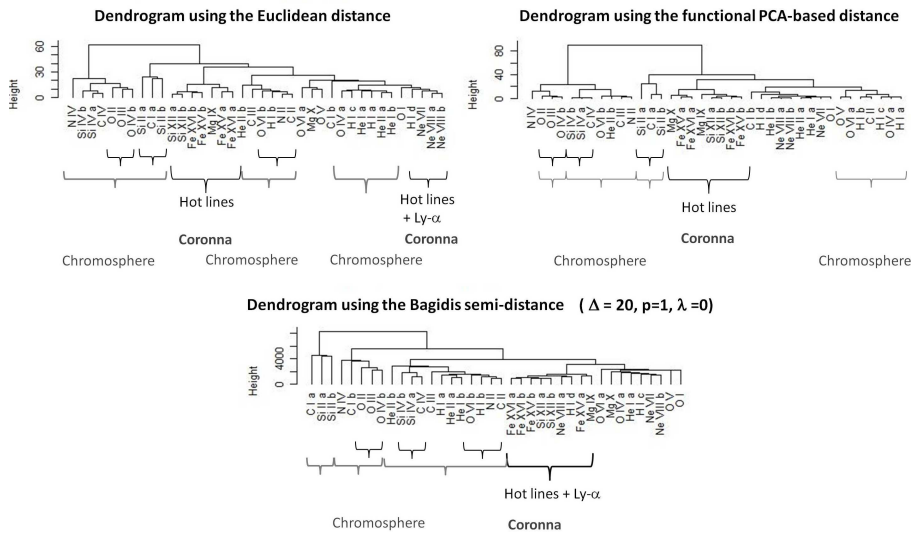


Figure 1.20: Dendrograms obtained using a Ward's agglomerative algorithm on the 38 solar irradiance time series based on various distances or semi-distances. The time series are labeled by the ions that generate them. Significant groups are underbraced and an indication of the related region of emission is provided.

Conclusion of the Time Series Analysis. In this study, the shape of the dendrograms leads to favor a balance parameter $\lambda = 0$. We get results similar to the ones obtained using the Euclidean distance or the functional PCA-based semi-distance. This example highlights the consistency of the BAGIDIS methodology as it illustrates that BAGIDIS works as well as usual competitors do even in a noisy situation where the specific properties of the BAGIDIS semi-distance are not needed.

1.5 Conclusion

In this Chapter, we introduced a new functional and wavelet-based method for comparing curves with sharp patterns. Using simulated and real data examples, we illustrated its ability for statistically detecting differences amongst curves. Its main originality is its ability to take into account both vertical and horizontal variations of sharp patterns in curves in a unified framework, which is particularly crucial when comparing

curves with sharp local features. Effects of horizontal and vertical variations can be separately diagnosed or measured simultaneously, depending on the value of a single balance parameter. An ANOVA-like F-test within the BAGIDIS context has been proposed in view of testing for differences amongst groups of curves. A geometrical test for the relative position of two groups of curves has also been developed. The use of this latter test is not restricted to the BAGIDIS framework. When used with the BAGIDIS semi-distance, the test benefits from its ability to efficiently analyze curves with sharp local patterns.

A major contribution of this work is to highlight the potential of a hierarchical description of curves in functional data analysis. Overcoming the limitation of expanding all the curves in the same basis allows for individually more efficient descriptions of the curves, those descriptions remaining comparable thanks to the hierarchy. This new paradigm opens attractive prospects for curves comparison, as illustrated in this study. Direct perspectives for future work go into two main directions: the use of BAGIDIS in Nonparametric Functional Regression (Chapter 2) and a data-driven definition of the weight function. Perspectives also include a generalization of BAGIDIS to image processing (Chapter 4).

*

Acknowledgments. *The authors thank Réjane Rousseau for the preprocessing of the spectrometric data, Matthieu Kretzschmar for interesting discussions about Timed/SEE data, and Léopold Simar for his helpful comments. Comments of an Editor, Associate Editor and two Referees of Technometrics are also acknowledged. Part of this work was completed during a stay of Catherine Timmermans at Université d'Orléans-CNRS, that was funded by the Fond National de la Recherche Scientifique (Belgium). Financial support from the IAP research network grant P 06/03 of the Belgian government (Belgian Science Policy) is gratefully acknowledged. The TIMED mission, including the SEE instrument, is sponsored by NASA's Office of Space Science.*

*

Previous versions of this work include:

Timmermans C., von Sachs, R. , 2010, *BAGIDIS, a new method for statistical analysis of differences between curves with sharp discontinuities*, ISBA Discussion Paper 2010/30.

Timmermans C., von Sachs R. and Delouille V., 2009, *Comparaison et classification de séries temporelles via leur développement en ondelettes de Haar asymétriques*, Actes des XVIe rencontres de la Société Francophone de Classification, p. 57-60.

Chapter 2

Functional Prediction

Our goal is to predict a scalar value or a group membership from the discretized observation of curves with sharp local features that might vary both vertically and horizontally. To this aim, we propose to combine the use of the non parametric functional regression estimator developed by Ferraty and Vieu (2006) with the BAGIDIS semi-distance developed in Chapter 1 in view of efficiently measuring dissimilarities between curves with sharp patterns. This association reveals powerful. Under quite general conditions, we obtain the rate of convergence of the nonparametric regression estimator in this case, as a function of the parameters of the BAGIDIS semi-distance. We propose to optimize those parameters using a cross-validation procedure, and show the optimality of the selected vector. This last result has a larger scope and concerns the optimization of any vector parameter characterizing a semi-distance used in this context. The performances of our methodology are assessed on simulated and real data examples. Results are shown superior than those obtained using competing semi-distances as soon as the variations of the significant sharp patterns in the curves have an horizontal component.

Note that, in the remaining of this Chapter, the term “semimetric” replaces the term “semi-distance”. This is for accordance with the literature in the field of nonparametric functional data analysis to which this research contributes. In the framework of this thesis, both terms are understood interchangeable.

Keywords: functional data, nonparametric regression, semimetric, wavelet, misalignment, cross-validation.

*

Using BAGIDIS in Nonparametric Functional Data Analysis: Predicting from Curves with Sharp Local Features

C. Timmermans, L. Delsol and R.von Sachs

Modern datasets often provide sets of points corresponding to discretized curves, typically time series or spectra. In this framework, the suitable information unit is the underlying curve instead of the vectorial quantity encoding the series - we refer to it as to a *functional data*. Functional statistical methods aim at taking this feature into account when extracting the information content of a dataset (see Ramsay and Silverman (2002, 2005); Bosq (2000); Ferraty and Vieu (2006); Ferraty and Romain (2010), e.g.). Amongst them, nonparametric methods often rely on the availability of a suitable metric or semimetric for measuring differences amongst the curves of the dataset.

In this framework, the BAGIDIS semimetric has been introduced in Chapter 1 as a highly adaptive wavelet-based tool for measuring dissimilarities between functional data. Its main originality is to be based upon the expansion of each series of a dataset into a *different* wavelet basis, one that is particularly suited for its hierarchical description. Measuring dissimilarities in such a way implies comparing not only the projections of the series onto the bases, as usual, but also the bases themselves. Because of this specificity, the semimetric is named BAGIDIS, which means *BAses GIVING DIS-tances*. As a consequence of this feature, the BAGIDIS semimetric has the ability to capture the variations of patterns occurring in series along both the vertical and the horizontal axis. This property makes the semimetric particularly powerful when dealing with curves that might be affected simultaneously by horizontal shifts and vertical amplifications.

Nonparametric functional data analysis techniques have been widely described in Ferraty and Vieu (2006). They include a set of prediction techniques that do not require to make an hypothesis on the form of the prediction operator - only smoothness hypotheses are made - and are able to efficiently deal with functional data provided it is used together with a semimetric able to extract the relevant features of the curves and satisfying some theoretical properties.

The purpose of the present study is to illustrate how we can advantageously make use of the BAGIDIS semimetric in the context of nonparametric functional prediction when the curves we are predicting from are characterized by some horizontally- and vertically-varying sharp local patterns. Our goal is also to show how this good behaviour is theoretically supported. Simulated examples are shown as well as a real data example, with spectroscopic ^1H NMR data.

This Chapter is organized as follows. The statistical framework of our study is first described in Section 2.1: the notion of functional models for prediction is introduced, before focusing on nonparametric functional prediction; the BAGIDIS semimetric, which is the tool we propose to use in this setting, is then presented. Then, our two main theoretical results are stated in Section 2.2. The first one gives the rate of convergence of a nonparametric functional prediction using the BAGIDIS semimetric, as a function of its parametrization. A practical cross-validation approach for optimizing the parametrization of BAGIDIS in this context is deduced therefrom and theoretically validated. This second result has a more general scope as it theoretically supports for a cross-validated optimization of the parameters of any projection-based semimetric. Finally, simulated and real data examples are investigated in Section 2.3, so as to illustrate nonparametric prediction with BAGIDIS in action.

2.1 Statistical framework

This Chapter proposes to bring together two advanced statistical tools, in view of providing a new, efficient, way to predict from curves with sharp local patterns. Those tools are, on the one hand, the nonparametric functional regression estimator provided by Ferraty and Vieu (2006) for obtaining predictions from functional data and, on the other hand, the BAGIDIS semimetric developed in Chapter 1 for comparing curves with sharp local patterns. As a support for our present work, those two tools are successively described in this Section, after a general presentation of the stakes and challenges of functional prediction.

2.1.1 Functional models for prediction

The general goal of a regression model is to link two random variables; the first one is a *response variable* Y which we are interested to explain; the second one is an *explanatory variable* X which is believed to be able to inform us about the response. Practically, this requires to estimate an unknown link operator r , based on some known pairs (X_i, Y_i) - those pairs are called the *training set* - so as to be able to predict Y for any new X using this link.

About the response variable. Regarding the response variable, two cases are most commonly encountered:

- **the response Y is a real measurement** whose we aim at predicting the conditional mean value for any given value of X . In that case, the link function r

between the two variables is an operator defined as the conditional expectation of the response variable, given the explanatory variable:

$$r(\chi) = \mathbb{E}(\mathbf{Y}|\chi),$$

which can be equivalently rewritten as

$$\mathbf{Y} = r(\chi) + \epsilon,$$

where ϵ is an error term such that $\mathbb{E}(\epsilon|\chi) = 0$. This is called a *regression model*. For any given value χ of χ , the associated scalar value Y is thus estimated as

$$\hat{Y} = \hat{r}(\chi),$$

for an estimator $\hat{r}$ of r .

- **the response $\mathbf{Y}$ is a class membership** that we have to determine for any given value of χ . This is called a *discrimination model*. The link function we are looking at in that case is the probability of being member of a given class g given a value of χ :

$$r_g(\chi) = P(\mathbf{Y} = g|\chi) = \mathbb{E}(\mathbf{1}_{[\mathbf{Y}=g]}|\chi),$$

for each class index $g \in \{g_i\}_{i=1\dots G}$, with G the number of classes and $\mathbf{1}_{[\mathbf{Y}=g]} = 1$ if $\mathbf{Y} = g$ and 0 otherwise. For any given value χ of χ , the associated class membership Y is thus estimated as

$$\hat{Y} = \arg \max_{g \in \{g_i\}_{i=1\dots G}} (\hat{r}_g(\chi)),$$

according to the Bayes rule, for an estimator $\hat{r}_g$ of r_g , $g \in \{g_i\}_{i=1\dots G}$.

As we can see, both regression and discrimination problems imply the estimation of an operator r or r_g that is defined as a conditional expectation. Consequently, the same statistical tools for its estimation may be involved in both cases.

About the explanatory variable. Classically, the explanatory variable has been a scalar or small-dimensional vector variable, and ways to estimate the operator r (or r_g) in that case have been known for a long time. New challenges and opportunities have appeared with *functional* explanatory variables taking their values in some function space $\mathcal{F}$ - i.e. explanatory variables being actually curves, typically spectra or functions of time. Of course, from a practical point of view, a curve is necessarily

observed as a discretized spectrum or a time series, that is to say a vector. Nevertheless, the classical multivariate regression framework is often not convenient anymore. There are two reasons for that:

- **The use of the complete information at hand.** The knowledge that there is a dynamic process underlying the sampled data is an information that could be exploited for an optimal estimation of the prediction model.
- **The “curse of dimensionality”.** To render the functional nature of the explanatory variable, it is often useful to make use of either a fine discretization, either a long period of data collection. As a consequence, the vectors representing the curves are rather large, leading to the need of a very large set of observations to correctly estimate the prediction model. In the nonparametric framework, this “curse of dimensionality” has been mathematically stated by Stone (1982):

Theorem. Consider a m times differentiable unknown regression function r of a $N - \text{dimensional}$ variable, and $\hat{r}$ a nonparametric estimator of r based on a training set of size n . Then the optimal rate of convergence of $\hat{r}$ to r is

$$\left(\frac{\log n}{n} \right)^{\frac{m}{2m+N}}. \quad (2.1.1)$$

To the best of our knowledge, the only way to predict from a curve nonparametrically while avoiding this curse of dimensionality consists in assuming that relevant information of reduced dimension can be extracted from the curves, the price to pay associated with this assumption being found in regularity assumptions on the regression operator. The knowledge of the functional nature of the data might help to efficiently extract this reduced dimensional information.

About the estimation of the link operator r , in case of functional data. Several ways have been recently proposed to take into account the functional nature of the explanatory variable in regression or discrimination problems. For instance, works of Ramsay and Silverman (2005), as well as Crambes et al. (2008); Ramsay and Dalzell (1991) and Cai and Hall (2006), are devoted to *parametric* functional modelling and focus on linear models. Beside, other kinds of parametric or semiparametric regression models have been considered in Sood et al. (2009); Ait-Saidi et al. (2008) and Aneiros-Pérez and Vieu (2008), for instance. On the other hand, researchers such as Ferraty and Vieu (2006) explore *nonparametric* functional prediction models - i.e. techniques allowing to dispose of the need to make an hypothesis on the form of the regression operator. Only smoothness hypotheses are required. Our work places itself in this nonparametric context.

2.1.2 Estimation in a nonparametric functional prediction model

Suppose we have got a set of curves $\{\chi_i\}_{i=1..n}$ and associated scalar values $\{Y_i\}_{i=1..n}$, and we are looking for a prediction model

$$r(\chi) = \mathbb{E}(Y|\chi),$$

for which we do not assume a particular parametric form, but only some regularity conditions. One of the basic assumptions underlying the concept of modelling is that similar values of χ correspond to similar value of Y . We need thus ways to quantify the similarity of the explanatory curves. Following Ferraty and Vieu (2006), we will say that d is a semimetric on some space $\mathcal{F}$ as soon as

- $\forall \chi \in \mathcal{F}, \quad d(\chi, \chi) = 0,$
- $\forall \chi_i, \chi_j, \chi_k \in \mathcal{F}, \quad d(\chi_i, \chi_k) \leq d(\chi_i, \chi_j) + d(\chi_j, \chi_k).$

A semimetric is thus defined in the same way as a distance, except that $d(\chi_i, \chi_j) = 0 \not\Rightarrow \chi_i = \chi_j$, and distances are particular cases of semimetrics. Semimetrics allow for measuring dissimilarities between curves through a reduced number of components. A well-suited semimetric may thus be a tool for extracting the relevant features of a set of curves.

A semimetric-based Nadaraya-Watson regression estimator. Given this definition, Ferraty and Vieu (2006) have proposed an extended Nadaraya-Watson estimator, that is able to deal with functional data provided we have a well-suited semimetric d such that some theoretical properties are satisfied. This estimator has the form:

$$\hat{r}(\chi) = \frac{\sum_{i=1}^n Y_i K\left(\frac{d(\chi, \chi_i)}{h}\right)}{\sum_{i=1}^n K\left(\frac{d(\chi, \chi_i)}{h}\right)}, \quad \text{for regression problems,} \quad (2.1.2)$$

and, as a particular case,

$$\hat{r}_g(\chi) = \frac{\sum_{i=1}^n \mathbf{1}_{[Y_i=g]} K\left(\frac{d(\chi, \chi_i)}{h}\right)}{\sum_{i=1}^n K\left(\frac{d(\chi, \chi_i)}{h}\right)}, \quad \text{for classification problems,} \quad (2.1.3)$$

where K is an asymmetric bounded kernel, n is the number of independent pairs (χ_i, Y_i) in the training set, d is a semimetric and h is the bandwidth. The well-known k -nearest neighbor (k -NN) estimator (Hastie et al. 2001) might be seen as a particular

case of the kernel estimator, with a kernel valued 1 on its support $[0; 1]$ and a suitably defined local bandwidth. It results in

$$\hat{r}(\chi) = \frac{1}{k} \sum_{i=1}^n Y_i \mathbf{1}_{[d(\chi, \chi_i) < h_{\chi, k}]} \quad \text{for regression problems,} \quad (2.1.4)$$

with $h_{\chi, k}$ being such that exactly k curves amongst the set $\{\chi_i\}_{i=1..n}$ are observed at a distance of χ smaller than $h_{\chi, k}$.

The choice of the semimetric. It is clear that the predictive qualities of the estimated regression or discrimination model depends on the features extraction capacities of the chosen semimetric d . It also relies on the regularity of the regression operator r with respect to d . Definitely, it is clear that the choice of the semimetric is crucial and must be related to the particular features on the functional dataset at hand. Commonly used families of semimetrics (Ferraty and Vieu 2006) are :

- The derivative-based family of semimetrics d_q^{deriv} based on the derivatives of order q of the curves :

$$d_q^{deriv}(\chi_i, \chi) = \sqrt{\int (\hat{\chi}_i^{(q)}(t) - \hat{\chi}^{(q)}(t))^2 dt}, \quad (2.1.5)$$

where $\hat{\chi}_i^{(q)}(t)$, $\hat{\chi}^{(q)}(t)$ are the estimations of the q^{th} derivative of χ_i and χ , respectively, at abscissa t , and where the integral has to be numerically estimated by a sum. Estimating the derivatives usually rely on a smoothing of the data. This family includes the Euclidean distance L_2 between the smoothed curves as a particular case, with $q = 0$; it will be denoted d_0^{deriv} . A contrario, the classical vectorial euclidean L_2 distance between the unsmoothed observations will be shortly referred to as d^{L_2} .

- The PCA-based family of semimetrics d_q^{PCA} based on a certain number q of principal components of the dataset:

$$d_q^{PCA}(\chi_i, \chi) = \sqrt{\sum_{k=1}^q \left(\int (\chi_i(t) - \chi(t)) \hat{v}_k(t) dt \right)^2}, \quad (2.1.6)$$

where $\hat{v}_k$, $k = 1..q$ is the k^{th} estimated eigenfunction of the principal component analysis, and where the integral has to be numerically estimated by a sum.

- The *hshift* semimetric d^{hshift} that realigns curves before computing d_0^{deriv} distances between them.

Those semimetrics have been shown useful in various problems (Ferraty and Vieu 2006; Ferraty et al. 2005; Aneiros-Pérez et al. 2004). However, they happen to fail when dealing with curves with sharp local features that might not be well aligned from one curve to another one, as discussed in Chapter 1. Besides, the computation of those semimetrics relies on a smoothing of the data, which is generally problematic for curves with abrupt patterns. This difficulty with curves with sharp patterns will be illustrated in subsequent examples of regression problems, in Section 2.3. However, such curves with sharp patterns happen to be dealt with in a large variety of scientific area (spectra, time series ...), so that it is worth thinking of the use of another, better adapted, semimetric for those data.

2.1.3 BAGIDIS, a semimetric for comparing curves having horizontally- and vertically-varying sharp local features

The BAGIDIS semimetric has been introduced in Chapter 1 so as to measure differences between regularly discretized curves that are characterized by some sharp local features. For the sake of simplicity, we denote it d^B in this Chapter, instead of $d_{p,\lambda}^{BAGIDIS}$. Readers having Chapter 1 in mind might wish to skip the remaining of this Subsection and continue to Section 2.2. Otherwise, we provide here for an overview.

The BAGIDIS semimetric is a functional data-driven and wavelet-based measure that is highly adaptive to the curves being considered. It has been proved to be a semimetric in Chapter 1. Key ideas are as follows.

Looking for a hierarchical description of the patterns of the series. We consider series observed on a regular grid $\mathbb{N}_{[1;N]}$. When we evaluate dissimilarities between series visually, we intuitively investigate first the global shapes of the series for estimating their resemblance, before refining the analysis by comparing the smaller features of the series. In other words our comparison is based upon a hierarchical comprehension of the curves. This visual approach inspired us to define our semimetric: we expand each series in a (different, series-adapted) basis that describes its features hierarchically, in the sense that the first basis vectors carry the main features of the series while subsequent basis vectors support less significant patterns; afterwards, we compare both the bases and the expansions of the series onto those bases, rank by rank, according to the hierarchy.

Expanding each series of the dataset in the *Unbalanced Haar Wavelet Basis* that is best suited for the hierarchical description of its shape. The family of *Unbalanced Haar Wavelet Bases* has been introduced by Girardi and Sweldens (1997). It consists in orthonormal bases that are made of one constant vector and a set of Haar-like (i.e. *up-and-down* shaped) orthonormal wavelets whose discontinuity point (hereafter the breakpoint) between the positive and negative parts is not necessarily located at the middle of its support. The *Bottom Up Unbalanced Haar Wavelet Transform* (BUUHW), an algorithm that was developed by Fryzlewicz (2007), allows for selecting amongst this family of bases the best basis for describing a given series hierarchically. Beside this hierarchical organization, the selected basis inherits the good capacity of Haar wavelets to efficiently capture sharp patterns.

We denote the expansion of a series χ_i in that basis as

$$\chi_i = \sum_{k=0}^{N-1} d_i^k \psi_i^k, \quad (2.1.7)$$

where the coefficients d_i^k (hereafter the *detail* coefficients) are the projections of the series χ_i on the corresponding basis vectors ψ_i^k and where the set of vectors

$$\{\psi_i^k\}_{k=0 \dots N-1}$$

is the Unbalanced Haar wavelet basis that is best suited to the series χ_i , as obtained using the BUUHW algorithm. Besides, we denote b_i^k , the breakpoint of the wavelet ψ_i^k , at every rank $k \neq 0$.

Defining a semimetric by taking advantage of the hierarchy of those expansions.

As shown by Fryzlewicz (2007), the ordered set of breakpoints $\{b_i^k\}_{k=1 \dots N-1}$ determines the basis $\{\psi_i^k\}_{k=0 \dots N-1}$ uniquely. As a consequence, the set of points

$$\{\mathbf{y}_i^k\}_{k=1 \dots N-1} = \{(b_i^k, d_i^k)\}_{k=1 \dots N-1} \quad (2.1.8)$$

determines the shape of the series χ_i uniquely - i.e. it determines the series on the grid $\mathbb{N}_{[1;N]}$, except for a change of the mean level of the series, that is encoded by the additional coefficient d_i^0 . It is the signature of the series in the *breakpoints-details* plane. Given that, and with the definition $b_i^0 = 0$ for each curve, we define the BAGIDIS semimetric as a 2-norm (weighted) distance in the *breakpoints-details* plane:

$$d^B(\chi_1, \chi_2) = \sum_{k=0}^{N-1} w_k \|\mathbf{y}_1^k - \mathbf{y}_2^k\|_2 = \sum_{k=0}^{N-1} w_k \left(|b_1^k - b_2^k|^2 + |d_1^k - d_2^k|^2 \right)^{1/2}$$

where w_k , $k = 0 \dots N - 1$, are well suited weights. As such, this semimetric takes advantage of the hierarchy of the well adapted unbalanced Haar wavelet bases: breakpoints and details of similar rank k in the hierarchical description of each series are compared to each other, and the resulting differences can be weighted according to that rank. As the breakpoints point to level changes in the series, the term $|b_1^k - b_2^k|$ can be interpreted as a measure of the difference of location of the features, along the horizontal axis. Being a difference of the projections of the series onto wavelets that encode level changes, the term $|d_1^k - d_2^k|$ can be interpreted as a measure of the differences of the amplitudes of the features, along the vertical axis. At rank $k = 0$, $|b_1^0 - b_2^0|$ vanishes and $|d_1^0 - d_2^0|$ measures the difference between the means of the curves.

Investigating the balance between breakpoints and details differences. We introduce an extension of the BAGIDIS semimetric as follows:

$$d^B(\chi_1, \chi_2) = \sum_{k=0}^{N-1} w_k \left(\lambda |b_1^k - b_2^k|^2 + (1 - \lambda) |d_1^k - d_2^k|^2 \right)^{1/2} \quad (2.1.9)$$

with $\lambda \in [0; 1]$. This parameter λ actually defines a scaling in the *breakpoints-details* plane, and hence in the original units of the problem. Setting λ at its extreme values 0 or 1 allows to investigate the contributions of the breakpoints differences and details differences separately. Moreover, the presence of this parameter allows the semimetric to be robust with respect to scaling effects: if λ is optimized according to a given criteria (such as the mean square error of a prediction model), the relative dissimilarities between the series of a dataset will remain the same, whatever the scales of measurements along the horizontal and vertical axes, so that the predictive qualities of the model will not be affected by such a change in the units of measurements. This variant (2.1.9) of the BAGIDIS semimetric is the one that we will use throughout this Chapter.

Choosing the weights. In a prediction setting, weights should ideally be positive at rank k if that rank carries information for discriminating the series, and 0 otherwise. In such a way, the weights could act as a filter that extract the part of the distances between the curves that carries relevant features. This Chapter will illustrate and validate that the weights could easily be obtained, altogether with the balance parameter and the bandwidth, using a cross-validation procedure across a set of possible values, in the framework of non parametric functional prediction.

2.2 The main results

This Section states the two main theoretical results of this Chapter, that support for the use of BAGIDIS in nonparametric functional regression. First, we obtain the rate of convergence of the nonparametric regression estimator (2.1.2) used with the BAGIDIS semimetric under suitable conditions. We see that this rate of convergence is related to the sparsity of the weight function that parametrizes the BAGIDIS semimetric in equation (2.1.9). Consequently, we propose to use a cross-validation procedure so as to optimize this weight function, as well as the balance parameter λ in equation (2.1.9) and the bandwidth h in equation (2.1.2). Our second result is the asymptotic optimality of this method.

2.2.1 Rate of convergence of estimator (2.1.2) when used together with BAGIDIS.

Our first result is the rate of pointwise convergence of the nonparametric functional regression estimator (2.1.2) (and (2.1.3) as a particular case) used together with the BAGIDIS semimetric. Under quite general conditions, we show that we can reach the rate of convergence

$$\left(\frac{\log n}{n} \right)^{\frac{\beta}{2\beta+K}},$$

with β being a Lipschitz parameter quantifying the smoothness of r , n the number of curves in the training set and K the number of non-zero weights w_k in the BAGIDIS semimetric (2.1.9). This rate of convergence is to be compared with the rate of convergence (2.1.1) for a nonparametric multivariate regression directly based on a N -dimensional variable:

$$\left(\frac{\log n}{n} \right)^{\frac{m}{2m+N}},$$

with m the order of differentiability of r and N the length of the discretized curve. This indicates that we can reach a quite good rate of convergence with $K \ll N$, provided we can restrict ourselves to a sparse enough weight function while satisfying regularity conditions on r . This may happen if the number of significant features in the curves of the dataset is not too large.

Our result relies on the following set of assumptions and definitions.

About the random curve χ . The random curve χ is observed on a regular grid $\mathbb{N}_{[1;N]}$. It is defined as a function of the breakpoints and details parameters (2.1.8) defining its BUUHWI expansion (2.1.7):

$$\chi = \chi(\mathbf{b}, \mathbf{d}), \text{ with } (\mathbf{b}, \mathbf{d}) \in \mathbb{N}_{[0;N]}^N \times \mathbb{R}^N. \quad (2.2.1)$$

Let us recall that the random signature $(\mathbf{b}, \mathbf{d})$ characterizes the random curve χ uniquely on the grid $\mathbb{N}_{[1;N-1]}$.

About the point of prediction χ . We denote by

$$\chi = \chi(b, d) \quad (2.2.2)$$

the fixed curve for which we want to obtain a prediction. This curve is uniquely related to the fixed point (b, d) in the breakpoint-detail plane.

About the response Y . We assume that Y is a scalar variable and that there exists $\sigma_m(\cdot)$ continuous at χ such that

$$\forall m \geq 2, \mathbb{E}(|Y|^m | \chi = \chi) \leq \sigma_m(\chi). \quad (2.2.3)$$

About the dataset. We assume to have n independent observations

$$(\chi_i, Y_i)_{i=1 \dots n} \quad (2.2.4)$$

of the random pair (χ, Y) .

About the BAGIDIS semimetric. We denote by d^B the BAGIDIS semimetric (2.1.9) with given balance parameter λ and weight function $w = \{w_k\}_{k=0 \dots N-1}$. We denote by $\mathcal{K}$ the set of indexes of non-zero components in w , and K the cardinality of this set. We assume that the non-zero weights are strictly positive:

$$\forall k \in \mathcal{K}, \quad w_k > 0. \quad (2.2.5)$$

About the probability distribution of χ . We denote by $f_{d|b}^{\mathcal{K}}(\mathbf{d})$ the conditional density function of $\mathbf{d}$ given $\mathbf{b} = b$, restricted on the (b^k, d^k) such that $k \in \mathcal{K}$. We assume that, at the fixed point (b, d) , $f_{d|b}^{\mathcal{K}}(\mathbf{d})$ is strictly positive and continuous with

respect to d^B : for all ϵ positive, there exists δ_ϵ positive such that

$$d^B(\chi(b, d), \chi(b, \mathbf{d})) \leq \delta_\epsilon \text{ implies } |f_{d|b}^\mathcal{K}(d) - f_{d|b}^\mathcal{K}(\mathbf{d})| \leq \epsilon. \quad (2.2.6)$$

We also assume that the curves $\chi(b, d)$ and $\chi(b, \mathbf{d})$ can have the same breakpoints for all ranks $k \in \mathcal{K}$ with a non-vanishing probability:

$$P(\forall k \in \mathcal{K}, \mathbf{b}^k = b^k) > 0. \quad (2.2.7)$$

This is possible because the breakpoints take their values on a finite grid of values. According to Ferraty and Vieu (2006), we define the small ball probability of χ around χ as

$$\phi_{d, \chi}(h) = P(\chi \in B_d(\chi, h)),$$

where $B_d(\chi, h)$ is the ball of radius h centered on χ and defined according to the semimetric d . We assume that

$$\forall \epsilon > 0, \phi_{d^B, \chi}(\epsilon) > 0. \quad (2.2.8)$$

About the regression operator. We assume that there exists β positive such that

$$r \in Lip_{\mathcal{F}, \beta} \equiv \{f : \mathcal{F} \rightarrow \mathbb{R}, \exists C \in \mathbb{R}_0^+, \forall \chi' \in \mathcal{F}, |f(\chi) - f(\chi')| < C d^B(\chi, \chi')^\beta\}. \quad (2.2.9)$$

About the kernel. We assume that K is a kernel function from $\mathbb{R}$ to $\mathbb{R}^+$ such that $\int K = 1$. We assume that there exists positive constants C and C' such that

$$C\phi_{d^B, \chi}(h) \leq \mathbb{E}\left(K\left(\frac{d^B(\chi, \chi)}{h}\right)\right) \leq C'\phi_{d^B, \chi}(h). \quad (2.2.10)$$

About the bandwidth. The bandwidth h is chosen according to a positive sequence h_n related to n in such a way that

$$\lim_{n \rightarrow \infty} h_n = 0 \text{ and } \lim_{n \rightarrow \infty} \frac{\log n}{n\phi_{d^B, \chi}(h_n)} = 0. \quad (2.2.11)$$

About the type of convergence. Following Ferraty and Vieu (2006), we consider almost complete convergence. One says that the stochastic sequence $(\mathbf{X}_n)_{n \in \mathbb{N}}$ converges almost completely to the real random variable $\mathbf{X}$ if and only if for all ϵ positive,

we have

$$\sum_{n \in \mathbb{N}} P(|\mathbf{X}_n - \mathbf{X}| > \epsilon) < \infty,$$

and we denote it by $\lim_{n \rightarrow \infty} \mathbf{X}_n = \mathbf{X}$ *a.co.* Moreover, one says that the rate of almost complete convergence of $(\mathbf{X}_n)_{n \in \mathbb{N}}$ to $\mathbf{X}$ is u_n if and only if there exists ϵ_0 positive such that

$$\sum_{n \in \mathbb{N}} P(|\mathbf{X}_n - \mathbf{X}| > \epsilon_0 u_n) < \infty,$$

and we write $\mathbf{X}_n - \mathbf{X} = \mathcal{O}_{a.co.}(u_n)$. A direct application of the Borel-Cantelli Lemma allows to prove (Ferraty and Vieu 2006) that almost complete convergence implies almost sure convergence and convergence in probability, and that

$$\mathbf{X}_n - \mathbf{X} = \mathcal{O}_{a.co.}(u_n) \text{ implies } \mathbf{X}_n - \mathbf{X} = \mathcal{O}_{a.s.}(u_n) \text{ and } \mathbf{X}_n - \mathbf{X} = \mathcal{O}_p(u_n).$$

We note that most of those conditions are very general and are not specific to the use of (2.1.2) with BAGIDIS. Only conditions (2.2.5), (2.2.6) and (2.2.7) are specific to the BAGIDIS semimetric and to the expansion of the curves in the *breakpoints-details* plane. Given those assumptions, our result is stated as follows.

Theorem 1. *Given assumptions (2.2.1) to (2.2.11), the functional kernel regression estimate (2.1.2) used together with the BAGIDIS semimetric is such that*

$$\hat{r}(\chi) - r(\chi) = \mathcal{O}(h^\beta) + \mathcal{O}_{a.co.} \left(\sqrt{\frac{\log n}{n h^K}} \right).$$

In particular, it can reach the rate of pointwise almost complete convergence

$$\hat{r}(\chi) - r(\chi) = \mathcal{O}_{a.co.} \left(\left(\frac{\log n}{n} \right)^{\frac{\beta}{2\beta+K}} \right).$$

The proof of this Theorem is given in Appendix B. An idea is as follows.

It is shown in Ferraty and Vieu (2006) that the rate of convergence of estimator (2.1.2) is linked with the semimetric through the behavior of the small ball probability $\phi_{d,\chi}(h)$ about 0: the higher $\phi_{d,\chi}(h)$, the faster the rate of convergence. This function $\phi_{d,\chi}(h)$ measures the concentration of the functional variable χ , according to the topology defined by the semimetric. Given this, our proof relies on two steps. First, we determine the behavior of the small ball probability around 0, for the BAGIDIS semimetric, as a

function of the number of non-zero weights in its definition. We see that there exists $C > 0$ such that $\phi_{d^B, \chi}(h) \sim C h^K$, when h tends to zero. This means that χ is fractal of order K around χ . Then, we make use of a result of Ferraty and Vieu (2006) that gives the rate of convergence of estimator (2.1.2) as a function of the small ball probability of fractal random variables.

The key step in this proof is thus the study of $\phi_{d^B, \chi}(h)$, the probability for a curve χ to be in a small ball of radius h around χ , according to the semimetric d^B . As soon as we consider curves χ at distances h smaller than the minimum of the weights for $k \in \mathcal{K}$ (which is strictly positive because of condition (2.2.5)), it is clear that their breakpoints $\mathbf{b}^k$ must be the same as the breakpoints b^k of the curve χ at which we want to predict, for all $k \in \mathcal{K}$. Indeed, if this were not true, there would exist at least one $k^* \in \mathcal{K}$ such that $|b^{k^*} - \mathbf{b}^{k^*}| \geq 1$, as the step of the grid is 1, so that the distance between the curves would be strictly greater than

$$w_{k^*} |b^{k^*} - \mathbf{b}^{k^*}| \geq w_{k^*} \geq \min_{k \in \mathcal{K}} w_k,$$

which leads to a contradiction for h small enough. This is the reason for condition (2.2.7). It is also the reason why our proof requires only the continuity of the conditional density function $f_{|b}$ of $(\mathbf{b}, \mathbf{d})$ given $\mathbf{b}$ at (b, d) (equation (2.2.6)). Finally, the reason for the cardinality K of $\mathcal{K}$ to appear in the behavior of the small ball probability and hence in the rate of convergence is that the volume of a ball of radius h and dimension K is proportional to h^K .

In the second step of our analysis, we refer to Theorem 6.11 in Ferraty and Vieu (2006), that states that

$$\hat{r}(\chi) - r(\chi) = \mathcal{O}(h^\beta) + \mathcal{O}_{a.co.} \left(\sqrt{\frac{\log n}{n \phi_{\chi, d}(h)}} \right). \quad (2.2.12)$$

This Theorem relies on conditions (2.2.3), (2.2.8), (2.2.9), (2.2.10) and (2.2.11). As in the classical multivariate setting, the first component on the right hand side of equation (2.2.12) is related to the bias of the estimate and depends only on the smoothness of the operator r . This component is controlled through condition (2.2.11, *left*). Similarly, the second component on the right hand side of equation (2.2.12) is related to the variance of the estimate. It is controlled through condition (2.2.11, *right*). In the case of a fractal type variable, with $\phi_{d^B, \chi}(h) \sim C h^K$ when h tends to zero, this term becomes

$$\mathcal{O}_{a.co.} \left(\sqrt{\frac{\log n}{n h^K}} \right),$$

which proves the first part of Theorem 1. The second part of the Theorem is then proved by choosing the bandwidth used in (2.2.12) as

$$h \sim C \left(\frac{\log n}{n} \right)^{\frac{\beta}{2\beta+K}}.$$

We note that condition (2.2.11, *right*) is automatically satisfied in that case.

It is worth emphasizing the trade-off relationship between β and K in the above rate of convergence. A good rate of convergence requires K small enough. On the opposite, K has to be high enough for the BAGIDIS semimetric (which depends on K) to satisfy the Lipschitz regularity condition of order β provided as assumption (2.2.9). If K is too small, the BAGIDIS semimetric would lack its discriminative power. One might then have certain χ, χ' such that we have simultaneously:

- $d^B(\chi, \chi') = 0$, because χ and χ' have the same components of signature on the K selected ranks of the BAGIDIS semimetric.
- $r(\chi) \neq r(\chi')$, because the curves actually *are* different although BAGIDIS does not capture this difference when too few components are included in the distance.

This is in contradiction with condition (2.2.9). A higher K is then needed to hopefully achieve a correct discrimination of the curves and to satisfy the Lipschitz condition.

We finally mention here two possible ways to generalize Theorem 1. First, we could combine our result about the fractality of the random variable χ with the results of Ferraty et al. (2010), so as to obtain the rate of uniform almost complete convergence of the functional kernel regression estimate (2.1.2) used with the BAGIDIS semimetric. Second, we could consider the use of the BAGIDIS semimetric in a more general non-parametric regression setting, with a functional response variable, in the framework provided by Ferraty et al. (2011).

We also indicate the existence of some structural tests for functional nonparametric regression, that might be advantageously used in the above framework. This family of tests has been developed by Delsol et al. (2011) and might be used for testing the effect on the response Y of the K first ranks of the signatures of the curves, for instance.

2.2.2 Asymptotic optimality of a cross-validated choice of the parameters of the semimetric and the bandwidth

As a second result, we propose, and theoretically support, the selection of a relevant weight function by using a leave-one-out cross-validation procedure, with a mean square error minimization criterion .

The cross-validated leave-one-out mean square error minimization criterion in nonparametric functional regression

We consider the estimator given by equation (2.1.2). The bandwidth h has to be specified in this expression. Moreover, in case d designates the BAGIDIS semimetric, it is parametrized by the balance parameter λ and the weight function w . Consequently, $\hat{r}$ relies on a vectorial parameter $H = (h, \lambda, w) \in \mathbb{R}^{N+2}$. We propose to chose this vectorial parameter H amongst a set $\mathcal{H}_n$ of possibilities by using a cross-validation procedure with a leave-one-out mean square error (MSE) criterion.

This cross-validation based approach for optimizing the estimate $\hat{r}$ generalizes the ideas of Rachdi and Vieu (2007) as well as Ait-Saidi et al. (2008). Rachdi and Vieu (2007) use a leave-one-out cross-validated MSE minimizer for choosing the bandwidth h , once a semimetric has been fully specified. They have shown the optimality of this procedure. For our purpose, not only the bandwidth but also the parameters λ and w defining the semimetric within a given family of semimetrics have to be optimized. A leave-one-out cross-validated selection of a parameter specifying a semimetric used within $\hat{r}$ has been proved asymptotically optimal in the particular case of a single functional index model (Ait-Saidi et al. 2008). In what follows, we generalize those results to families of semimetrics determined by a vectorial parameter. This more general framework includes the optimization of $\hat{r}$ used not only with the BAGIDIS semimetric, but also with any kind of projection-based semimetric for which we aim at selecting the components (or the number of first components) upon which we project the series for their comparison. Given the importance of the class of projection-based semimetric (see Ferraty and Vieu (2006, Chapter 13)), the opportunity to derive such a general result is clear.

Notations and main ideas

We denote by $\hat{r}_H(\chi)$ the regression operator estimator $\hat{r}(\chi)$ used with the fixed parameter H . For the sake of simplicity, we denote

$$\Delta_i(\chi) = K\left(\frac{d(\chi, \chi_i)}{h}\right) \quad \text{and} \quad K_H(\chi, \chi_i) = \frac{\Delta_i(\chi)}{\mathbb{E}(\Delta_i(\chi))},$$

with d , a semimetric parametrized by $H \setminus h$. Consequently, $\hat{r}_H(\chi)$ is denoted by

$$\hat{r}_H(\chi) = \frac{\frac{1}{n} \sum_{i=1}^n Y_i K_H(\chi, \chi_i)}{\frac{1}{n} \sum_{i=1}^n K_H(\chi, \chi_i)}. \quad (2.2.13)$$

Our criterion for measuring the quality of $\hat{r}_H(\chi)$ is the Mean Integrated Square Error (MISE) defined as follows:

$$MISE(H) \equiv MISE(\hat{r}_H, r) = \mathbb{E} \left(\int (\hat{r}_H(\chi) - r(\chi))^2 W(\chi) dP_\chi(\chi) \right), \quad (2.2.14)$$

where $W(\chi)$ is a non negative weight function and P_χ is the probability distribution measure of the functional variable χ . We define $\mathcal{H}_n$, a set of possible values for the parameter H , with the cardinality of $\mathcal{H}_n$ increasing with the sample size. We aim at selecting

$$H^* = \arg \min_{H \in \mathcal{H}_n} MISE(H).$$

However H^* cannot be obtained as r is unknown in expression (2.2.14). Consequently, we propose to estimate H^* by H^{CV} , defined as follows:

$$H^{CV} = \arg \min_{H \in \mathcal{H}_n} CV(H),$$

where $CV(H)$ is the leave-one-out cross-validated criterion defined by

$$CV(H) = \frac{1}{n} \sum_{j=1}^n (Y_j - \hat{r}_H^{-j}(\chi_j))^2 W(\chi_j),$$

with

$$\hat{r}_H^{-j}(\chi) = \frac{\sum_{i=1, i \neq j}^n Y_i K_H(\chi, \chi_i)}{\sum_{i=1, i \neq j}^n K_H(\chi, \chi_i)}.$$

$\hat{r}_H^{-j}(\chi_j)$ is thus the prediction associated to χ_j based upon the regression estimator (2.2.13) applied to our dataset whose the pair (χ_j, Y_j) has been excluded, with the parameter H . $CV(H)$ is thus an estimation of $MISE(H)$ calculated over the dataset.

The main result

The main result of this Section relies on the following conditions.

About the dataset. We assume to have n independent observations

$$(\chi_i, Y_i)_{i=1 \dots n} \tag{2.2.15}$$

of the random variable (χ, Y) .

About the kernel. The kernel K is bounded with compact support $[0; 1]$, Lipschitz on $\mathbb{R}^+$, and we have that for all $j = 1, 2 \dots$, there exist $C_{1j}, C_{2j} > 0$ such that for all $H \in \mathcal{H}_n$, there exists $0 < \Phi_H \leq 1$ so that

$$\forall \chi \in \mathcal{W} \quad C_{1,j} \Phi_H \leq \mathbb{E} \left(K^j \left(\frac{d(\chi, \chi)}{h} \right) \right) \leq C_{2,j} \Phi_H. \tag{2.2.16}$$

About the probability distribution of χ . We have that

$$\exists \gamma > 0, \exists C_1 > 0 \text{ such that } \sup_{H \in \mathcal{H}_n} \Phi_H \leq C_1 n^{-\gamma}, \tag{2.2.17}$$

and

$$\exists \delta > 0, \exists C_2 > 0 \text{ such that } \inf_{H \in \mathcal{H}_n} n \Phi_H \geq C_2 n^\delta. \tag{2.2.18}$$

About the weight function. The weight function $W(\cdot)$ is non negative, of compact support $\mathcal{W} \subset \mathcal{F}$, bounded by some positive constant C_W , and such that

$$0 < \int W(\chi) dP_\chi(\chi). \quad (2.2.19)$$

The interior of $\mathcal{W}$ is non-empty and we have, for all $H \in \mathcal{H}_n$,

$$\mathcal{W} \subset \bigcup_{k=1}^{d_n} B(c_k, r_n), \quad (2.2.20)$$

where $B(c_k, r_n)$ are balls of $\mathcal{F}$, of center c_k and radius $r_n = o\left(\inf_{H \in \mathcal{H}_n} h \Phi_H\right)$, with $d_n \leq Cn^\eta$, $\eta > 0$.

About the conditional distribution of the errors. The conditional mean of the errors is zero:

$$\mathbb{E}(\epsilon|\chi) = 0. \quad (2.2.21)$$

The conditional variance of the error is positive and there exists $\sigma_0 > 0$ such that

$$\mathbb{E}(\epsilon^2|\chi) \geq \sigma_0^2. \quad (2.2.22)$$

About the regression operator. The regression operator r is bounded by some positive constant C_r . We introduce the following definitions of the bias and the integrated square bias:

$$B(\chi) = \mathbb{E}((Y_i - r(\chi))K_H(\chi, \chi_i)) \quad \text{and} \quad b_H = \int B^2(\chi)W(\chi)dP_\chi(\chi).$$

There exists a positive constant C_B such that for all $H \in \mathcal{H}_n$,

$$\forall \chi, \chi' \in \mathcal{F} \text{ such that } d(\chi, \chi') \leq h, \text{ we have } |r(\chi) - r(\chi')| \leq C_B b_H^{\frac{1}{2}}. \quad (2.2.23)$$

This property is satisfied as soon as there exists constants $C, C' > 0$ such that $\forall \chi, \chi' \in \mathcal{F}$ such that $d(\chi, \chi') \leq h$, there exists $\beta_H \geq 0$ so that

$$|r(\chi) - r(\chi')| \leq C' h^{\beta_H} \quad \text{and} \quad \int B^2(\chi)W(\chi)dP_\chi(\chi) \geq C' h^{2\beta_H}.$$

About the set of parameters. The cardinality of the set of parameters $\mathcal{H}_n$ is increasing at most algebraically fast:

$$\exists \alpha > 0, \exists C > 0 \text{ such that } \#\mathcal{H}_n \leq Cn^\alpha. \quad (2.2.24)$$

About the conditional moments of $\mathbf{Y}$. We assume that Y is such that

$$\forall k = 1, 2, \dots \exists C_k > 0 \text{ such that } \mathbb{E}(|Y|^k | \chi) \leq C_k, \quad (2.2.25)$$

and

$$\forall \chi, \chi_i \in \mathcal{W}, \forall k, l = 1, 2, \dots \exists C_{kl} > 0 \text{ such that } \mathbb{E}\left(|Y|^k K^l\left(\frac{d(\chi, \chi_i)}{h}\right)\right) \leq C_{kl} \Phi_H. \quad (2.2.26)$$

This last property is valid as soon as conditions (2.2.16) and (2.2.25) are satisfied.

Theorem 2. Assuming conditions (2.2.15) to (2.2.26), we have

$$\frac{MISE(H^{CV})}{MISE(H^*)} \longrightarrow 1 \quad a.s.$$

This Theorem states the asymptotic optimality of the cross-validated choice H^{CV} amongst the set $\mathcal{H}_N$. It is proved in Appendix B. An idea of the proof is as follows.

In a view to prove Theorem 2, we need to introduce the quantity

$$MISE^*(H) = \int \mathbb{E}\left((\hat{r}_{2H}(\chi) - r(\chi)\hat{r}_{1H}(\chi))^2\right) W(\chi) dP_\chi(\chi) \quad (2.2.27)$$

$$= \int \mathbb{E}\left(\left(\frac{1}{n} \sum_{i=1}^n \delta_{i\chi}\right)^2\right) W(\chi) dP_\chi(\chi), \quad (2.2.28)$$

with

$$\hat{r}_{1H}(\chi) = \frac{1}{n} \sum_{i=1}^n K_H(\chi, \chi_i), \quad \hat{r}_{2H}(\chi) = \frac{1}{n} \sum_{i=1}^n Y_i K_H(\chi, \chi_i)$$

and

$$\text{and } \delta_{i\chi} = (Y_i - r(\chi))K_H(\chi, \chi_i).$$

Two results are then needed for this quantity $MISE^*(H)$:

$$\sup_{H \in \mathcal{H}_n} \left| \frac{MISE(H) - MISE^*(H)}{MISE^*(H)} \right| = o_{a.s.}(1) \quad (2.2.29)$$

$$\frac{MISE^*(H^{CV})}{MISE^*(H^*)} \rightarrow 1 \quad a.s. \quad (2.2.30)$$

The proof is then rather short: first, (2.2.29) is used to deal with $MISE^*(H)$ instead of $MISE(H)$, then $\left| \frac{MISE^*(H^{CV}) - MISE^*(H^*)}{MISE^*(H^*)} \right|$ is bounded above by a sequence tending to zero when n tends to infinity, because of (2.2.30). The proof of (2.2.29) is the purpose of Lemma 13. It requires that

$$\forall H \in \mathcal{H}_n, \exists C > 0, \text{ such that } \frac{C}{n\Phi_H} + \frac{n-1}{n}b_H \leq MISE^*(H).$$

This bound from below of $MISE^*(H)$ is shown valid by Lemma 5. The proof of (2.2.30) is given by Lemma 14. This Lemma relies on the following inequality:

$$\left| \frac{MISE^*(H^{CV}) - MISE^*(H^*)}{MISE^*(H^*)} \right| (1 - T_\alpha - T_\beta - T_\gamma) \leq (T_\alpha + T_\beta + T_\gamma) \frac{2}{1 - T_\alpha} + 2T_\alpha,$$

with

$$T_\alpha = \sup_{H \in \mathcal{H}_n} \left| \frac{MISE^*(H) - ASE(H)}{MISE^*(H)} \right|, \quad (2.2.31)$$

$$T_\beta = \sup_{H \in \mathcal{H}_n} \left| \frac{ASE(H) - \widetilde{ASE}(H)}{MISE^*(H)} \right| \quad (2.2.32)$$

and

$$T_\gamma = 2 \sup_{H \in \mathcal{H}_n} \left| \frac{CT(H)}{MISE^*(H)} \right|, \quad (2.2.33)$$

and the following definitions of the Average Square Error

$$ASE(H) = \frac{1}{n} \sum_{i=1}^n \left(\hat{r}_H(\chi_i) - r(\chi_i) \right)^2 W(\chi_i),$$

of the Average Square Error of the leave-one-out predictor

$$\widehat{ASE}(H) = \frac{1}{n} \sum_{j=1}^n \left(\hat{r}_H^{-j}(\chi_j) - r(\chi_j) \right)^2 W(\chi_j),$$

and with the following definition of $CT(H)$:

$$CT(H) = \frac{1}{n} \sum_{j=1}^n (Y_j - r(\chi_j)) \left(\hat{r}_H^{-j}(\chi_j) - r(\chi_j) \right) W(\chi_j) = \frac{1}{n} \sum_{j=1}^n \epsilon_j \hat{\epsilon}_j^{-j} W(\chi_j).$$

It is thus necessary to show that T_α , T_β and T_γ converge to zero almost surely when n goes to infinity. Those terms are dealt with separately through Lemmas 10, 11 and 12 respectively. Lemma 6 validates a condition that is needed for the proofs of those three Lemmas. Lemmas 8 and 9 provide with decompositions that have a role in the proof of Lemma 10. Those decompositions allow to highlight terms that have a form similar to the ones identified by Marron and Härdle (1986) for a similar problem in the multivariate setting. Those terms can be shown to converge to 0 in a way similar to the one proposed by those authors, but with conditions adapted to our functional setting. The purpose of Lemma 7 is to show that the conditions for our Theorem are enough to obtain conditions that play the same role as the ones used by Marron and Härdle (1986) in their proof.

Although our proofs follow similar ideas as the one given in Rachdi and Vieu (2007) and Ait-Saidi et al. (2008), some major differences occur. As the parameter H that we cross-validate makes changes in the semimetric, it sounds not natural to assume that the function $\hat{r}_H$ is always regular with respect to the semimetric. Opposite to Rachdi and Vieu (2007) and Ait-Saidi et al. (2008), we can thus not make use of this condition anymore. This translates into the fact that the Lipschitz parameter β_H might be equal to zero in condition (2.2.23). As a consequence, the bias of the estimate might not necessarily go to zero, for some choice of the parameters. This translates into the need to use a more precise inferior bound

$$\frac{C}{n\Phi_H} + \frac{n-1}{n}b_H$$

for the $MISE^*(H)$ (Lemma 5), instead of $\frac{C}{n\Phi_H}$ in Rachdi and Vieu (2007) and in Ait-Saidi et al. (2008). In particular, the proof of the convergence of T_γ in Lemma 12 requires a much more careful treatment. Nevertheless, the assumptions for Theorem 2 to hold are quite general and very close to the ones required by Rachdi and Vieu (2007) and Ait-Saidi et al. (2008).

2.3 Applications and discussions

In this Section, we assess the performances of BAGIDIS in nonparametric functional prediction using estimator (2.1.2), as compared with results relying upon the usual semimetrics defined in Subsection 2.1.2. We also illustrate the efficiency of the cross-validation procedure described in Subsection 2.2.2 for selecting the weights of the BAGIDIS semimetric in this context. Simulated and real data examples involving curves with sharp patterns are studied therefore.

We observe on simulated examples that BAGIDIS shows prediction performances highly superior to competitors as soon as the model involves curves whose significant variations of sharp local patterns have an horizontal component. In case the model involves a sharp pattern whose variation in amplitude is significant but which remains well-aligned across the dataset, the PCA-based semimetric is best, but performances of BAGIDIS achieve nearly the same order of magnitude when the noise on the curves is not too high. The cross-validated selection of the non-zero weights of the BAGIDIS semimetric, and hence of the significant ranks in the BUUHW expansion (2.1.7) of the curves, proves very efficient on those examples, with very few selections of un-significant ranks. This also holds when the significant sharp pattern to capture is a secondary pattern and is thus not encoded in the first ranks of the BUUHW expansion. Such a cross-validated selection allows to further improve the predicting performances of BAGIDIS as compared with a non-optimized version of BAGIDIS and with competing semimetrics.

A real data example involving ^1H NMR serum spectra is then studied. The goal is to discriminate healthy patients from patients suffering from a given illness, according to the composition of their blood serum. Again, BAGIDIS proves highly efficient for that purpose.

Those analyses are performed using the R software (R Development Core Team 2008), and using the R implementations of estimators (2.1.2) and (2.1.3), provided on the companion website of Ferraty and Vieu (2006), slightly adapted for their use together with the BAGIDIS semimetric.

2.3.1 A systematic simulated study of the prediction capacities of BAGIDIS for datasets of curves having horizontally shifted and/or vertically amplified sharp patterns

We investigate the potential of using the BAGIDIS semimetric in regression by studying simulated datasets involving curves having a single significant sharp pattern that is either horizontally shifted across the dataset, or vertically amplified, or both simultaneously. The related responses in those regression problems derive from the amplitude and/or location of that sharp pattern. Those examples allow for diagnosing the ability of BAGIDIS to deal with different kinds of differences amongst the curves of a dataset. We emphasize the fact that the method does not make use of the prior information of the nature of the variation amongst the curves. One of our goals in this work is precisely for our method to automatically adapt to this nature, through an optimal choice of the balance parameter λ , of the weights w , and of the bandwidth h .

Definition of the simulated models. The models we study are as follows. First, we investigate how BAGIDIS handles horizontal shifts and vertical amplifications of patterns separately, through the analysis of curves generated from the following models:

- **Model 1: an *up-and-down* horizontally shifted pattern is related to its delay.**
The first example involves series of length 21, being zero-valued except for the presence of an *up-and-down* pattern $(10, -10)$ that is horizontally shifted from one series to the next one. Each series is related to the delay at which the *up-and-down* pattern occurs. This is illustrated at Figure 2.1 (*top, left*).
- **Model 2: an *up-and-down* vertically amplified pattern is related to its height.**
The second example involves series of length 21 being zero-valued except for an *up-and-down* pattern located at abscissas $(10, 11)$, that is more or less amplified from one series to the next one, from amplitude 1 to amplitude 20. Each series is associated with the height of the *up-and-down* pattern. This is illustrated at Figure 2.2 (*top, left*).

We then study a model that combines horizontal shifts and vertical amplifications of sharp patterns:

- **Model 3: an amplified and shifted *up-and-down* pattern is related to a value depending on both its height and delay.** We consider series of length 21, being zero-valued except for the presence of an *up-and-down* pattern $(1, -1)$. That pattern appears after a certain delay and is affected by a certain multiplicative

σ_χ :	0.25	0.5	1	2	3
$\frac{s}{\sigma_\chi}$	4	2	1	0.5	0.3
$\frac{\bar{s}d(\chi)}{\sigma_\chi}$ for Model 1: shifted patterns	12.8	6.4	3.2	1.6	1.1
$\frac{\bar{s}d(\chi)}{\sigma_\chi}$ for Model 2: amplified patterns	13.2	66.6	3.3	1.6	1.1
$\frac{\bar{s}d(\chi)}{\sigma_\chi}$ for Model 3: randomly shifted and amplified patterns	13.2	66.6	3.3	1.6	1.1
$\frac{\bar{s}d(\chi)}{\sigma_\chi}$ for Model 4: secondary shifted patterns	28	14	7	3.5	2.3

Table 2.1: **Signal-to-noise ratio for the simulation study of Subsection 2.3.1.** $\frac{s}{\sigma_\chi}$ is the ratio of the smallest difference (vertically or horizontally) between the model curves s and the standard deviation of the noise applied to the curves σ_χ . It is the same for all models, as s is always fixed to 1. $\frac{\bar{s}d(\chi)}{\sigma_\chi}$ is the ratio of the standard deviation of the model curves, averaged on a sample set of curves, $\bar{s}d(\chi)$ and the standard deviation of the noise applied to the curves σ_χ . Although common, those values $\frac{\bar{s}d(\chi)}{\sigma_\chi}$ have to be taken with caution in our study, as $\bar{s}d(\chi)$ includes thus variations of the curves that are either significant or non-significant (Model 4) and do not take horizontal shifts into account (Models 1, 3 and 4).

amplification factor, both being randomly generated in $1 \dots 20$. Sample curves generated according to this model are illustrated in Figure 2.3 (*top, left*). The responses associated with those curves are the sum of the delay and the amplitude.

The last model we consider involves two sharp patterns, the main one being non-informative, the secondary one being the only one whose variation carries significant information:

- **Model 4: an horizontally shifted secondary up-and-down pattern is related to its delay.** We consider series of length 21, being zero-valued except for the presence of a main *up-and-down* pattern $(-20,20)$ located at abscissas $(10,11)$, as well as the presence of a secondary *up-and-down* pattern $(-10,10)$ that is horizontally shifted along the series (it is thus possibly combined to the main pattern, for certain delays). Response values are defined as the delay at which this secondary pattern occurs. This is illustrated at Figure 2.4 (*top, left*).

The simulated series we generate according to those four models are affected by a Gaussian noise with a standard deviation σ_χ taking its values in $(0.25, 0.5, 1, 2, 3)$ - depending on the simulation, and the responses are affected by a Gaussian noise with standard deviation $\sigma_Y = 1$. The related signal-to-noise ratio are provided in Table 2.1.

Description of the analysis. The following test is performed T times, for each model and each value of σ_χ . We generate m noisy pairs $(\chi_i, Y_i)_{i=1 \dots m}$ according

to the chosen model, each model value of the delay and/or height having the same probability to appear in the dataset. Then, we randomly select n pairs out of those m , and use them as a training set to calibrate the regression model. Using the model for predicting the responses associated with the $m - n$ remaining series and comparing it with their “true” simulated noisy response, we calculate the associated mean square error of prediction (MSE).

The performances obtained using the BAGIDIS semimetric with estimator (2.1.2) on those problems are compared with the one we obtain using the functional PCA-based semimetric d_q^{PCA} with various number of principal components, the derivative-based semimetric d_q^{deriv} with various order of derivation (including no derivation) and the *hshift* semimetric d^{hshift} . The use of a vectorial L_2 -distance d^{L_2} as a semimetric is also considered. Besides, a *no effect* prediction is provided - i.e. a prediction by the mean of the response values of the training set, which acts as a benchmark for the performances.

Our analysis actually proceeds into two steps.

- **Step 1: Studying the performances of BAGIDIS in nonparametric functional regression as a function of the balance parameter λ , with a prior, sub-optimal, choice of the weights and a cross-validated-choice of the bandwidth h .** In order to get a first insight into the behavior of the BAGIDIS-based regression estimator as a function of λ , only an adaptation of the bandwidth is considered, and BAGIDIS is used with a prior weight function defined as

$$w_0 = 0; w_k = \frac{\log(N + 1 - k)}{\log(N + 1)} \text{ for } k = 1 \dots N - 1, \quad (2.3.1)$$

as proposed in Chapter 1. This allows to associate a large weight to the comparison of features encoded at the first rank of the hierarchy, and a decreasing weight to the smaller features at the end of the hierarchy, which is empirically what we expect for relatively sparse noisy curves. Values from λ from 0 to 1 with a step of 0.1 are tested. The bandwidth h is optimized through a set of values defined as a sequence of 20 equispaced values from the quantile 0.05 to the quantile 0.5 of the observed distances between the curves, which is the default behavior of the R function provided by Ferraty and Vieu (2006). With this first step, we investigate thus how those “sub-optimal” versions of the BAGIDIS semimetric behave compared with “classical” semimetrics, depending on the value of λ , which gives a first idea of the potential of our method. It also allows to identify the best competitors of BAGIDIS in each setting.

This analysis is performed with $T = 100$ and $m = 60$, $n = 45$ for **Model 1**, **Model 2** and **Model 4**. Because of the more important complexity of **Model 3**, a larger training set ($m = 180$, $n = 160$) is to be used if we wish to achieve an

explained percentage of the no-effect MSE that is about 90 for $\sigma_\chi = 0.25$, as for the other models. Smaller size of the training sets (e.g. $m = 60$, $n = 45$ as for the other models) leads to the same relative performances of the semimetrics as those presented here-above, but with a systematically higher MSE.

- **Step 2: Optimizing the weights, the balance parameter and the bandwidth in nonparametric regression using a *leave-one-out* cross-validation procedure.** As theoretically supported by Theorem 1, having a sparse weight function will significantly improve the rate of convergence the estimator, and hence the performance of BAGIDIS compared to competitors. With this second step, we illustrate how we can further improve the predicting performances of BAGIDIS on the above-defined models by optimizing the weights w , the balance parameter λ as well as the smoothing parameter h . Practically, this is done using a cross-validation procedure and a *leave-one-out* mean square error criterion, as suggested by Theorem 2. The set of parameters $H = (w, \lambda, h)$ over which we optimize the *leave-one-out* MSE criterion is defined as follows: values of λ are tested from 0 to 1 with a step of 0.1; h is allowed for taking its values in a sequence of 20 equispaced values from the quantile 0.05 to the quantile 0.5 of the observed distances between the curves; values of the weights w_k are only allowed for being 1 or 0 - i.e each rank k can be *activated* or *unactivated* in the semimetric. However, not all the possible combinations of weights are actually tested, as a *forward selection approach* is favored in order to reduce the optimization time (Guyon and Elisseeff 2003). This means we first compute the *leave-one-out* MSE on the training set for any possible combination of λ and h , for each of the possible single activated weights. The rank k^* whose activation leads to the smaller *leave-one-out* MSE is selected. If the best so-obtained *leave-one-out* MSE is strictly smaller than a “no effect” *leave-one-out* MSE, the weight w_{k^*} is set to 1. We then do the same for selecting another activated rank in the weight function. This procedure is iterated while the resulting *leave-one-out* MSE decreases. The best set of parameters $H^{Opt} = (w^{k^{Opt}}, \lambda^{Opt}, h^{Opt})$ is thus selected as the minimizer of our criterion amongst the tested sets of parameters. Note here that such a *forward selection procedure* is a very common approach for selecting among a large set of parameters (see Guyon and Elisseeff (2003), for instance). Once the optimal set of parameters H^{Opt} is selected, the mean square errors of prediction is evaluated on the validation set, using the optimal so-parametrized predictor.

This analysis is performed with $T = 30$ and m and n defined as for Step 1.

Presentation and discussion of the results. Resulting distributions of the MSE obtained at **Step 1** for each of the tested semimetrics are presented in Figures 2.1 to 2.4, for each model and each value of σ_χ . When interpreting those results, we have to

keep in mind that the BAGIDIS results are sub-optimal here, as the parametrization of the semimetric is not optimized in this first analysis. Summary results extracted from those graphs about the MSE obtained using BAGIDIS and using its best competitor semimetric are shown in Table 2.3.1. Resulting distribution of the MSE obtained at **Step 2**, when the parametrization of BAGIDIS is fully optimized using a cross-validation procedure, are then summarized in Table 2.3. MSE distributions obtained with the optimized BAGIDIS semimetric for each model are confronted with the best competitor MSE distributions on the same model obtained by the analysis of **Step 1**, this best competitor distribution being either BAGIDIS with unoptimized weights and with the best expected value of λ , or another semimetric -typically the PCA.

Analysis of Model 1: Capturing the location of an horizontally shifted sharp pattern. A look at Figure 2.1 and Table 2.3.1 (row 1) tells us that, as expected, the BAGIDIS semimetric leads to excellent performances compared to all competitors for dealing with the **Model 1: shifted patterns**, as soon as $\lambda > 0$ - i.e. as soon as the differences between the breakpoints are taken into account in equation (2.1.9). We observe that the sensitivity to the choice of the parameter λ increases with σ_χ , and $\lambda = 1$ is most systematically favored in this case. This is not surprising as we know, by construction of the model, that only the breakpoints (solely captured with $\lambda = 1$) carry significant information. Only in the least noisy case $\sigma_\chi = 0.25$, d_0^{deriv} performs better than BAGIDIS. In that case, the bandwidth is actually selected so small that quasi-only similarly aligned curves define the predictor. On the contrary, BAGIDIS is able to detect the closeness of neighbor shifted curves for building the predictor. This leads to the fact that BAGIDIS-based model explains a significant part of the no-effect MSE, up to $\sigma_\chi = 3$. The additionally explained percentage of the no-effect MSE that is explained by BAGIDIS as compared with the best competitor reaches 39.75 when $\sigma_\chi = 3$. At that level of noise, the percentage of explanation offered by the sub-optimal BAGIDIS-based model is more than twice better than the one achieved by its best competitor. Further non illustrated studies show that this advantage of the BAGIDIS semimetric remains up to a noise level $\sigma_\chi = 6$, where no model is able to do significantly better than the no-effect MSE. Moreover, as could have been expected, increasing σ_Y increases the MSE whatever the semimetric, but does not affect their relative performances.

Optimizing the parameters of the BAGIDIS allows to further improve our prediction performances. As can be seen from Table 2.3 (row 1), a significant percentage ($> 80\%$) of the no-effect MSE is now explained even in the most noisy illustrated setting with $\sigma_\chi = 3$. We observe a gain of 7 to 10 % of explanation of Y , as compared with the BAGIDIS semimetric used with its prior weight function and the optimal value $\lambda = 1$, that was noticed superior to all competitors at **Step 1** for $\sigma_\chi > 0.25$. Moreover, the BAGIDIS semimetric is now competitive as compared with d_0^{deriv} even in this situation of small noise $\sigma_\chi = 0.25$. The number of selected weights remains

small in all examples. Ranks 1 and/or 2, the ones that carry significant, redundant, information about the shifted pattern, are selected most of the time. On average less than 2 non-zero weights are selected for $\sigma_\chi < 3$, as expected. In conclusion, the BAGIDIS semimetric is clearly better than competitors for capturing the shift of a sharp pattern, and its performances are further improved by optimizing its parameters.

Analysis of Model 2: Capturing the amplification of well-aligned sharp pattern.

Not surprisingly, the **Model 2: amplified patterns** is best tackled by the functional PCA-based semimetric, as can be seen from Figure 2.2 and Table 2.3.1 (row 2). Nevertheless, it is interesting to note that our sub-optimal BAGIDIS semimetric performs quite well too for $\sigma_\chi < 2$, with $\lambda = 0$ - i.e where only amplitude differences are taken into account. In those cases less than 2.18 percent of the no-effect MSE is additionally explained when using a PCA-based semimetric. The advantage of the PCA-based semimetric becomes really significant afterwards. However, it is interesting to note that BAGIDIS still significantly detects an effect of the amplification of the sharp pattern, up to a noise level $\sigma_\chi = 5$, what the derivative-based family of semimetric cannot do. Again, increasing σ_Y increases the MSE whatever the semimetric, but does not affect their relative performances.

As indicated by Table 2.3 (row 2), optimizing the parameters of the BAGIDIS does not significantly allow to further improve our prediction performances. The number of selected weights is higher than 4 most of the time, indicating a certain number of spurious rank selection. In summary, the PCA-based semimetric is best in case of well aligned sharp patterns variations, but the BAGIDIS semimetric remains competitive if the noise on the curves is not too important

Analysis of Model 3: Capturing the height and delay of a randomly amplified and shifted sharp pattern.

From Figure 2.3 and row 3 of Table 2.3.1, it is clear that even the sub-optimal form of BAGIDIS performs very well on **Model 3: randomly amplified and shifted pattern**, and significantly better than competitors in every illustrated case. For upper values of σ_χ ($\sigma_\chi > 4$), no model is able to do significantly better than the no effect MSE. As expected, an intermediate value of λ seems to be the best choice as both differences in the localizations and in the amplitudes are informative for the prediction. $\lambda = 0.2$ or $\lambda = 0.3$ seem to be favored. This can be interpreted because 0.25 is the ratio of the squared range for the breakpoints over the square range of the details (the square has to be taken because we use $p = 2$ in equation (2.1.9)). Again, further non illustrated studies show that increasing σ_Y increases the MSE whatever the semimetric, but does not affect their relative performances.

Optimizing the parameters of the BAGIDIS semimetric slightly improves our prediction performances, as compared with BAGIDIS semimetric used with its prior weight function and $\lambda = 0.25$. This smaller amelioration, as compared with the improvement

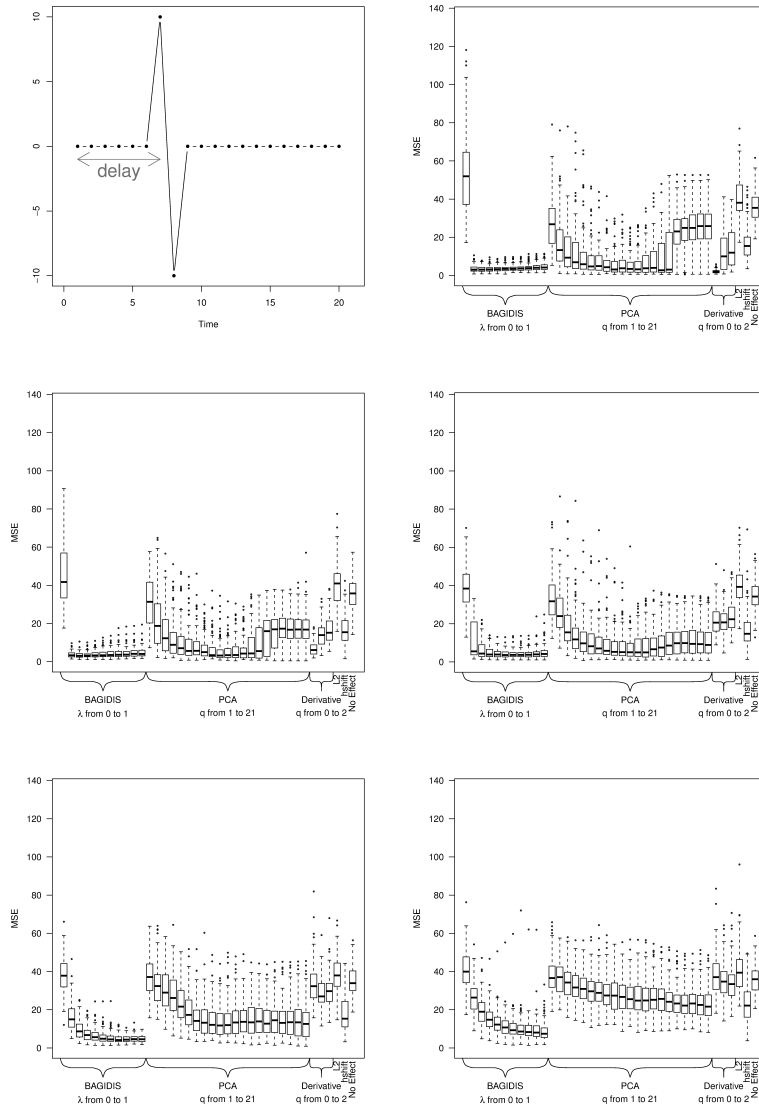


Figure 2.1: Analysis (Step 1) of the predictive performances of various semimetrics, on data generated according to Model 1: Shifted patterns. Top, left: Schematic illustration of Model 1: shifted patterns. Top, right to bottom right: Boxplot representations of the observed MSE distributions, for data generated according to Model 1, with $\sigma_Y = 1$ and $\sigma_X = 0.25, 0.5, 1, 2$, and 3 respectively (from left to right and top to bottom).

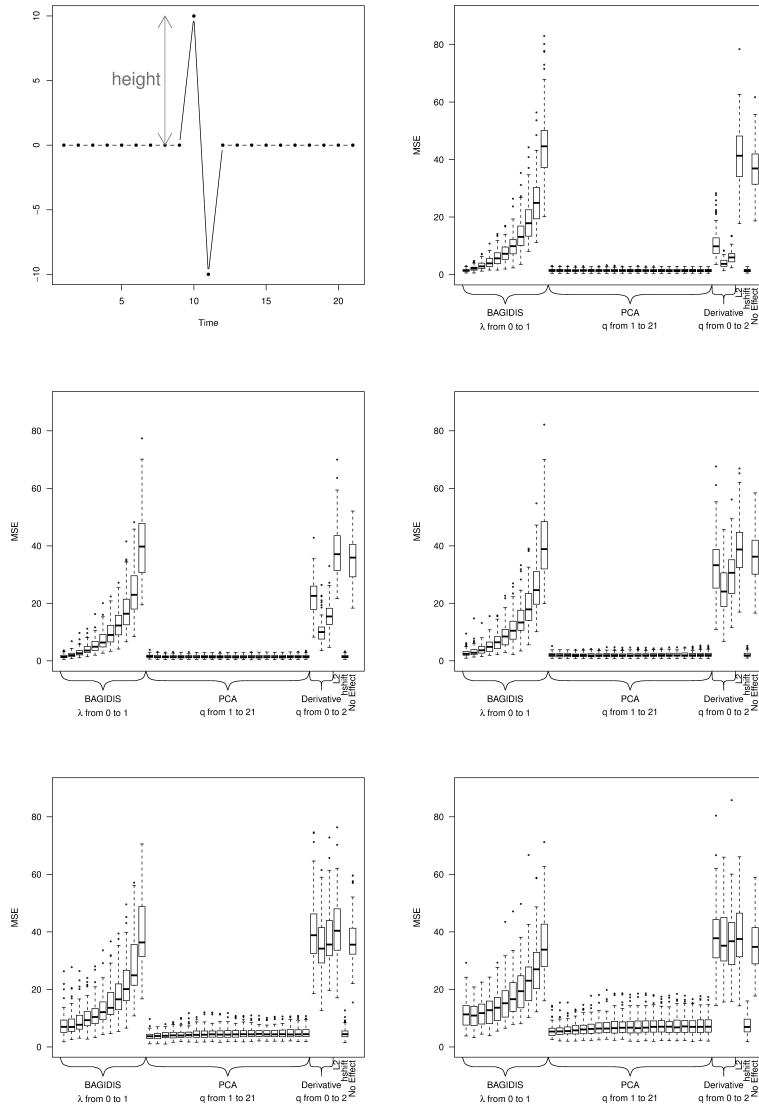


Figure 2.2: **Analysis (Step 1) of the predictive performances of various semimetrics, on data generated according to Model 2: amplified patterns.** *Top, left:* Schematic illustration of **Model 2: amplified patterns**. *Top, right to bottom right:* Boxplot representations of the observed MSE distributions, for data generated according to **Model 2**, with $\sigma_Y = 1$ and $\sigma_X = 0.25, 0.5, 1, 2$, and 3 respectively (from left to right and top to bottom).

achieved for Model 1, might be related to the fact that a larger training set is used here so that the noise that affects the prediction due to the presence of insignificant ranks in the prior weights function is better averaged to zero. The difference is more marked for a high level of noise $\sigma_\chi = 3$, as seen from Table 2.3 (row 3). The number of selected weights remains small whatever the noise level. Ranks 1 and 2, the ones that carry significant, redundant, information about the shifted and amplified pattern, are essentially selected, and few spurious selections occur. In average about 2 non-zero weights are selected, as expected. This third example shows that the BAGIDIS semimetric is clearly superior to competitors and optimizing its parametrization might help to further improve its performances.

Analysis of Model 4: Capturing a secondary shifted sharp pattern. The definition of **Model 4: secondary shifted pattern** implies that the first ranks of the BAGIDIS semimetric should encode non significant information as they essentially compare the largest uninformative main pattern, while ranks 4 and 5 should be relevant for predicting the response. This last example aims at checking that this kind of behavior is correctly handled by our optimization procedure. Results obtained at **Step 1** for this model can be found in Figure 2.4 and Table 2.3.1 (row 4). As for **Model 1: shifted pattern**, it shows that the sub-optimal non-optimized BAGIDIS semimetric has very good performances as compared with competitors, as soon as $\lambda > 0$, with the best λ significantly tending to 1 as σ_χ increases. Again, and for the same reason as discussed for **Model 1**, d_0^{deriv} is best in case of a low level of noise $\sigma_\chi = 0.25$ and has equivalent performances for $\sigma_\chi = 0.5$. For higher noise levels $\sigma_\chi > 0.5$, the BAGIDIS semimetric, even with its sub-optimal parametrization, performs clearly better than competitors. Further non illustrated studies show that this advantage of the BAGIDIS semimetric remains up to a noise level $\sigma_\chi = 6$, where no model is able to do significantly better than the no-effect MSE.

Table 2.3 (row 4) shows that the cross-validated selection of the weights leads to a significant improvement of the performances of BAGIDIS, at least for $\sigma_\chi < 3$. This is probably due to the fact that the prior weight function was not really adapted to this example, as it gave a higher weights to the first insignificant ranks. On the contrary, the optimization procedure efficiently selects ranks 4 and/or 5, the ones that carry significant information on the secondary shifted pattern, with very few spurious selection so that the number of selected weights remain small in every case. The advantage of d_0^{deriv} for small values of σ_χ disappears as soon as the selection procedure of the weights takes place, and results achieved with the optimized BAGIDIS semimetric are equivalent ($\sigma_\chi = 0.25$) or highly better ($\sigma_\chi > 0.25$) than with d_0^{deriv} .

In summary, the BAGIDIS semimetric is highly efficient for predicting from curves with a secondary shifted sharp pattern, and highly benefits from a cross-validated se-

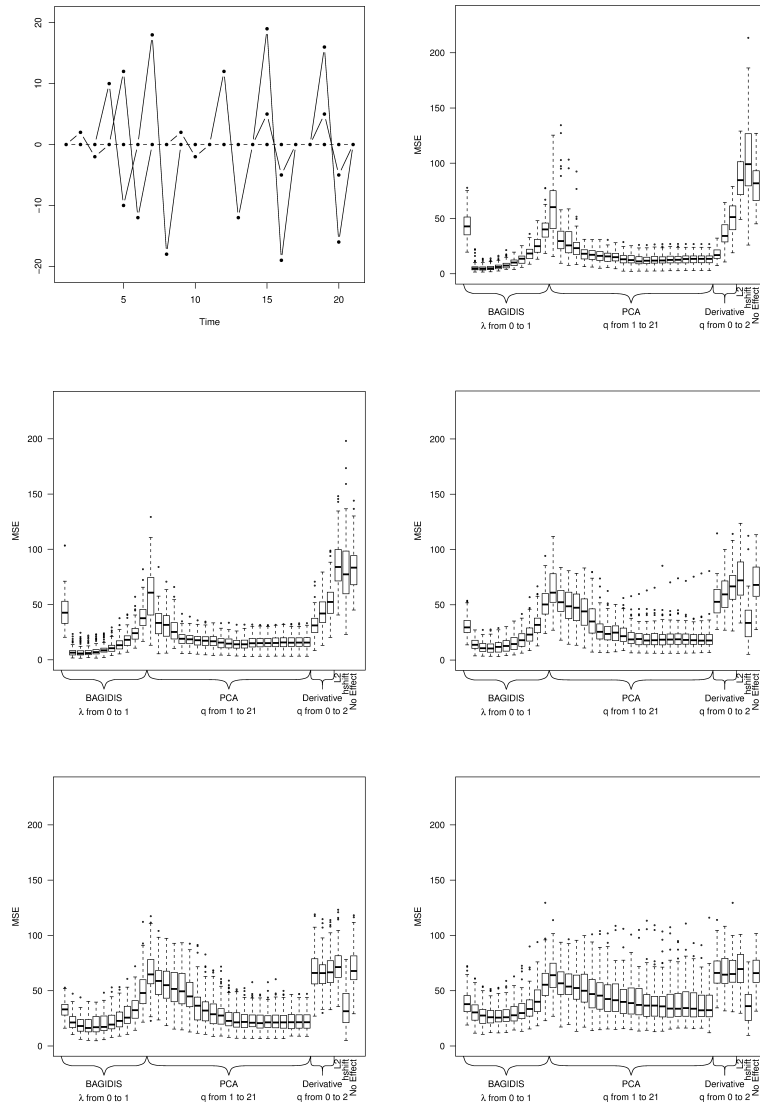


Figure 2.3: Analysis (Step 1) of the predictive performances of various semimetrics, on data generated according to Model 3: randomly shifted and amplified patterns. Top, left: Schematic illustration of sample curves generated from **Model 3: randomly shifted and amplified patterns**. **Top, right to bottom right:** Boxplot representations of the observed MSE distributions, for data generated according to **Model 3**, with $\sigma_Y = 1$ and $\sigma_X = 0.25, 0.5, 1, 2$, and 3 respectively (from left to right and top to bottom).

lection of the activated weights in this case. The significant ranks are directly selected with very few spurious selection.

Conclusion of the simulated study. Those simulated examples illustrates the potential of using BAGIDIS for nonparametric prediction of curves. Even used with its non-optimized prior weight function, performances of BAGIDIS shows superior performances compared to classical semimetric as soon as variations of sharp local patterns in curves have an horizontal component. Those performances may be further improved by a cross-validated selection of the parameters of BAGIDIS. The mean number of activated weights is then always small. This means we reach quite good rates of convergence in these examples. Moreover, the performances of BAGIDIS are equivalent to the ones of its best competitor, the PCA-based semimetric, in case no horizontal variation of the significant pattern occur, provided that the noise is not too high. It remains “acceptable” for higher noises, meaning that a prediction is still possible (which is not the case for the derivative-based semimetric for instance). This means that BAGIDIS could be used quite confidently on datasets with sharp patterns whose kind of variation might not be known in advance. A specificity of BAGIDIS is indeed that the semimetric can adapt itself to the kind of variation to detect in the dataset, through the optimization of the balance parameter λ .

2.3.2 Analysis of a real spectroscopic dataset

This last Subsection presents a real data example of prediction from spectroscopic curves (de Tullio, Frédéricich and Lambert, Université de Liège). We consider 193 ^1H NMR serum spectra of length 600, as illustrated in Figure 2.5, 94 of which corresponding to patients suffering from a specific illness, the other ones corresponding to healthy patients. More details about this dataset are provided in Chapter 3, as well as further discussion and investigation of the results.

In this Subsection, we aim at predicting from the spectrum if a patient is healthy or not. A training set of 150 spectra is randomly selected and a functional nonparametric discrimination model is adjusted, with various semimetrics. In each case, the number of misclassification observed on the remaining 43 spectra is recorded.

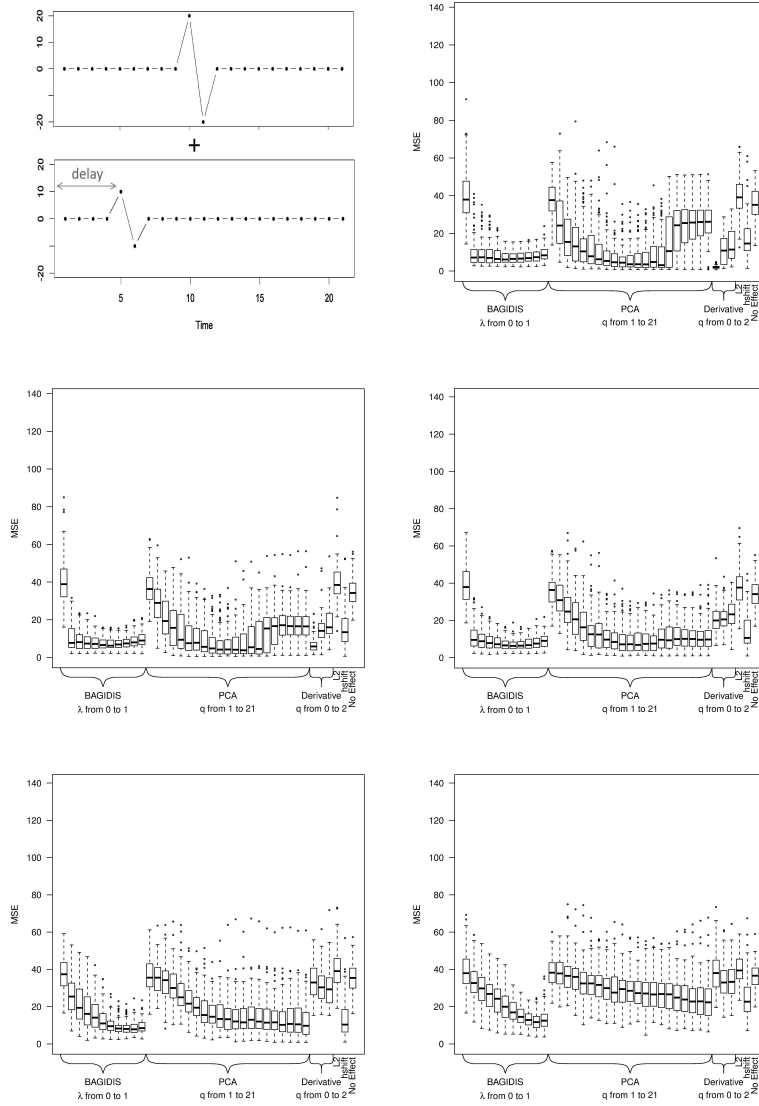


Figure 2.4: Analysis (Step 1) of the predictive performances of various semimetrics, on data generated according to **Model 4: second order shifted patterns**. *Top, left*: Schematic illustration of **Model 4: second order shifted patterns**. *Top, right to bottom right*: Boxplot representations of the observed MSE distributions, for data generated according to **Model 4**, with $\sigma_Y = 1$ and $\sigma_X = 0.25, 0.5, 1, 2$, and 3 respectively (from left to right and top to bottom).

Model 1: Shifted Patterns, $\sigma_Y = 1$

σ_X	No-effect MSE	BAGIDIS			Competitor			Comparison of performances	
		Best λ selected	Mean MSE	Percentage of no-effect MSE explained by the model	Best selected competitor	Mean MSE	Percentage of no-effect MSE explained by the model	Additionally explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	36.09	0.2	3.34	90.74	Deriv-0	2.14	94.06	-3.32	0.96
0.5	35.54	0.2	3.45	90.29	PCA-11	5.23	85.28	5.01	1.06
1	34.80	0.5	4.08	88.28	PCA-11	7.46	78.56	9.72	1.12
2	34.77	0.8	4.62	86.71	PCA-21	13.92	59.96	26.75	1.45
3	35.59	1	8.83	75.15	PCA-21	22.99	35.40	39.75	2.12

Model 2: Amplified Patterns, $\sigma_Y = 1$

σ_X	No-effect MSE	BAGIDIS			Competitor			Comparison of performances	
		Best λ selected	Mean MSE	Percentage of no-effect MSE explained by the model	Best selected competitor	Mean MSE	Percentage of no-effect MSE explained by the model	Additionally explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	36.96	0	1.46	96.04	PCA-11	1.41	96.18	-0.14	1.00
0.5	35.20	0	1.56	95.57	PCA-7	1.43	95.94	-0.37	1.00
1	36.26	0	2.77	92.36	PCA-5	1.98	94.54	-2.18	0.98
2	37.17	0	7.65	79.42	PCA-1	3.83	89.97	-10.55	0.88
3	35.86	0.1	11.18	68.82	PCA-1	5.57	84.47	-15.65	0.81

Model 3: Random Amplification and Location of the Patterns, $\sigma_Y = 1$

σ_X	No-effect MSE	BAGIDIS			Competitor			Comparison of performances	
		Best λ selected	Mean MSE	Percentage of no-effect MSE explained by the model	Best selected competitor	Mean MSE	Percentage of no-effect MSE explained by the model	Additionally explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	82.16	0.2	4.91	94.02	PCA-12	12.30	85.03	8.99	1.11
0.5	82.57	0.2	6.98	91.5	PCA-13	14.59	82.33	9.17	1.11
1	70.55	0.3	11.35	83.91	PCA-19	19.38	72.53	11.38	1.16
2	70.14	0.3	18.55	73.55	PCA-15	22.40	68.06	5.49	1.08
3	67.29	0.3	26.93	59.98	hshift	36.11	46.34	13.64	1.29

Model 4: Second Order Shifted Patterns, $\sigma_Y = 1$

σ_X	No-effect MSE	BAGIDIS			Competitor			Comparison of performances	
		Best λ selected	Mean MSE	Percentage of no-effect MSE explained by the model	Best selected competitor	Mean MSE	Percentage of no-effect MSE explained by the model	Additionally explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	35.65	0.6	7.04	80.25	Deriv-0	2.08	94.16	-13.91	0.85
0.5	35.13	0.6	7.118	79.56	Deriv-0	6.94	80.24	-0.68	0.99
1	34.29	0.6	6.91	79.85	PCA-11	8.86	74.16	5.69	1.08
2	35.54	0.9	8.55	79.94	PCA-21	12.49	64.85	11.09	1.17
3	36.04	0.9	12.36	65.70	PCA-21	23.63	34.43	31.27	1.91

Table 2.2: Summary analysis of the examples of Step 1 analysis in Subsection 2.3.1. The percentage of no-effect explained by the model is calculated as $100(1 - \frac{\text{Mean no-effect MSE}}{\text{Mean MSE}})$. The difference between this percentage for the BAGIDIS semimetric and for its best competitor is given as the *Additionally explained percentage of no-effect MSE*. This percentage is thus negative in case the competitor semimetric performs better than BAGIDIS. The last column of the table is the ratio of the percentages of no-effect MSE explained using BAGIDIS and using its best competitor. BAGIDIS is superior as soon as this ratio exceeds 1.

Model 1: Shifted Patterns, $\sigma_Y = 1$

σ_X	Optimized BAGIDIS			Percentage of no-effect MSE explained by the model	Competitor		Percentage of no-effect MSE explained by the model	Comparison of performances	
	Mean number of activated weights	Most activated weights	Mean MSE		Best selected competitor	Mean MSE		Additionally explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	1.4	1 and 2	1.48	95.11	BAGIDIS-1	4.16	85.99	9.13	1.11
0.5	1.2	1 and 2	1.62	94.92	BAGIDIS-1	4.48	85.73	9.19	1.11
1	1.5	1 and 2	1.46	95.34	BAGIDIS-1	4.01	86.37	8.97	1.10
2	1.5	1 and 2	2.48	92.76	BAGIDIS-1	4.89	85.16	7.60	1.09
3	2.3	1 and 2	5.77	80.86	BAGIDIS-1	8.16	73.75	7.11	1.10

Model 2: Amplified Patterns, $\sigma_Y = 1$

σ_X	Optimized BAGIDIS			Percentage of no-effect MSE explained by the model	Competitor		Percentage of no-effect MSE explained by the model	Comparison of performances	
	Mean number of activated weights	Most activated weights	Mean MSE		Best selected competitor	Mean MSE		Additionally explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	4.6	2 and 3	1.47	95.44	PCA-1	1.49	95.36	0.07	1.00
0.5	4.1	2 and 3	1.56	95.08	PCA-1	1.46	95.41	-0.33	1.00
1	4.1	2 and 3	2.57	91.70	PCA-1	1.92	93.65	-1.95	0.98
2	4.3	2 and 3	5.56	80.34	PCA-1	3.65	87.11	-6.77	0.92
3	3.2	2 and 3	9.93	68.40	PCA-1	5.82	81.68	-13.28	0.84

Model 3: Randomly Shifted and Amplified Patterns, $\sigma_Y = 1$

σ_X	Optimized BAGIDIS			Percentage of no-effect MSE explained by the model	Competitor		Percentage of no-effect MSE explained by the model	Comparison of performances	
	Mean number of activated weights	Most activated weights	Mean MSE		Best selected competitor	Mean MSE		Additionally explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	2.0	2 and 1	2.37	96.19	BAGIDIS-0.25	5.94	90.58	5.61	1.06
0.5	2.6	2 and 1	5.72	91.58	BAGIDIS-0.25	8.91	87.06	4.52	1.05
1	2.4	2 and 1	11.41	80.67	BAGIDIS-0.25	11.95	80.61	0.06	1.00
2	2.7	2 and 1	13.83	75.35	BAGIDIS-0.25	16.92	69.88	5.46	1.08
3	2.6	2 and 1	24.06	60.75	BAGIDIS-0.25	31.07	49.87	10.88	1.22

Model 4: Second order Shifted Patterns, $\sigma_Y = 1$

σ_X	Optimized BAGIDIS			Percentage of no-effect MSE explained by the model	Competitor		Percentage of no-effect MSE explained by the model	Comparison of performances	
	Mean number of activated weights	Most activated weights	Mean MSE		Best selected competitor	Mean MSE		Additionally explained percentage of no-effect MSE	Ratio of explained percentages of no-effect MSE
0.25	1.5	4	2.60	90.27	BAGIDIS-1	11.01	66.68	25.45	1.38
0.5	1.5	4 and 5	3.18	92.13	Deriv-0	2.21	93.31	-1.18	0.99
					BAGIDIS-1	10.51	67.85	22.42	1.33
1	1.7	4 and 5	3.75	88.16	Deriv-0	7.95	75.68	14.52	1.19
					BAGIDIS-1	9.24	70.83	17.33	1.24
2	2.7	4 and 5	6.83	78.57	Deriv-0	17.87	43.59	44.57	2.02
					BAGIDIS-1	9.48	70.25	8.32	1.12
3	3.0	4 and 5	13.34	58.30	Deriv-0	33.93	-6.46	78.57*	*
					BAGIDIS-1	13.28	58.49	-0.19	1.00
					Deriv-0	35.34	-10.47	58.30*	*

Table 2.3: Summary analysis of the examples of Step 2 analysis in Subsection 2.3.1. The percentage of no-effect explained by the model is calculated as $100(1 - \frac{\text{Mean no-effect MSE}}{\text{Mean MSE}})$. The difference between this percentage for the BAGIDIS semimetric and for its best competitor is given as the additionally explained percentage of no-effect MSE. This percentage is thus negative in case the competitor semimetric performs better than BAGIDIS. The last column of the table is the ratio of the percentages of no-effect MSE explained using BAGIDIS and using its best competitor. BAGIDIS is superior as soon as this ratio exceeds 1. Values denoted by * in the Comparison of performances for Model 4 appear when the percentage of no-effect MSE explained using the competitor semimetric is observed negative, indicating the the no-effect prediction is better - i.e. that the semimetric does not capture anything about the significant variations in the curves. In this case, the additionally explained percentage of no-effect MSE is the actual percentage of no-effect MSE explained by BAGIDIS, and the Ratio of explained percentages is not computed.

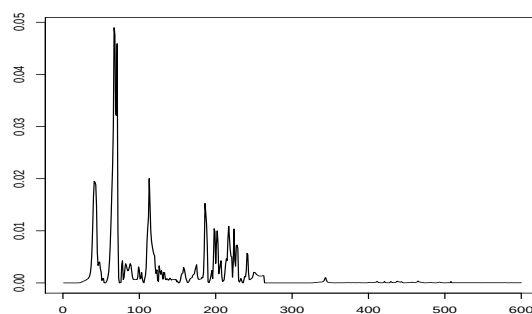


Figure 2.5: An ^1H NMR serum spectra for a ill patient.

In order to avoid a confusion of the features in such long series, we make use of the BAGIDIS semimetric together with a sliding window of length 25 (as suggested in Chapter 1). This allows for comparing the variations of one or few given peak(s) at a time. A specific R function for using estimator (2.1.3) in a discrimination setting has been provided by Ferraty and Vieu (2006). However, it makes use of a slightly different version of the nonparametric estimator (2.1.3): a local bandwidth is used, which is defined through a number of nearest neighbor that have to be included in the support of the kernel function. This is a k -NN estimator, as defined by eq. (2.1.4). Consequently, our good properties stated in Section 2.2 are not strictly proved in this case. However, we believe that the good convergence properties of the kernel-based *leave-one-out* MSE minimizer might be extended to a k -NN-based *leave-one-out* MSE minimizer, by generalizing our proofs to this case using similar arguments as those found in Burba et al. (2009).

Our test for the prediction of the health status from the spectra is repeated 80 times, with different randomly selected training sets, using the sub-optimal BAGIDIS semimetric with its prior weight function, with $\lambda = 0.5$ and with a cross-validated bandwidth h , and with the competitor semimetrics identified in Subsection 2.1.2 and a cross-validated bandwidth. Results are summarized in Table 2.3.2, for BAGIDIS and its best competitor, being a PCA-based semimetric with at least 6 components. We observe that the non-optimized BAGIDIS obtains *no error* 10% more often than the PCA-based semimetric. Afterwards, we optimize the weights and the λ parameter of the BAGIDIS semimetric using a cross-validation procedure within the training set, and the resulting model is tested on the remaining 43 series. This test is repeated 18 times on different randomly selected training sets, and no prediction error occurs. At each repetition, only 1 non-zero weight is selected. We observe no prediction error in every case, indicating a risk of misclassification that is probably smaller than 0.05.

	Occurrences of 0 error out of 43 predictions	Occurrences of 1 error out of 43 predictions
PCA-based semimetric with $q \geq 6$	40 times out of 80 50%	40 times out of 80 50%
Non-optimized BAGIDIS semimetric with prior weights and $\lambda = 0.5$	48 times out of 80 60%	32 times out of 80 40%
Optimized BAGIDIS semimetric (1 non zero weight is selected)	18 times out of 18 100%	0 times out of 18 0%

Table 2.4: **Summary results for the prediction of the health status from the spectra.** Training sets of 150 curves are randomly selected. Predictions are obtained for the 43 remaining spectra and compared with the true health status of those 43 patients. The number of prediction errors is computed. The process is repeated several times for different randomly selected training sets.

This indicates a very good capacity of discriminating the serum spectra from ill and healthy patients.

2.4 Conclusion

The key idea of this Chapter is to combine the nonparametric functional framework provided by Ferraty and Vieu (2006) with the highly adaptive BAGIDIS semimetric (Chapter 1), for predicting from curves with sharp patterns. This association proves highly pertinent. Applications on simulated data have shown in a systematic way the ability of BAGIDIS to take into account both horizontal and vertical variations of the patterns, as well as its flexibility in the use of this information. Predictions using BAGIDIS appear to be clearly better than predictions using competing semimetrics, as soon as the variations of the significant sharp patterns have an horizontal component. Those performances concerns both high and relatively low signal-to-noise ratio, which makes the method really attractive. The method also proves really powerful for prediction based on ^1H NMR spectra, issued from biomedical research.

A theoretical support for those very good observed performances has also been provided in this Chapter. It was shown that a really competitive rate of convergence of the prediction estimator can be achieved, provided that the multidimensional parametrization of the BAGIDIS semimetric is sparse enough. It is also shown that this multidimensional parametrization can be chosen using a cross-validation procedure, with a mean-square-error minimization criterion. This method is proved to be asymptotically optimal, and is shown highly efficient on the proposed data analyses examples. The related theoretical results also support for a cross-validated choice of the multidimensional parametrization of others semimetrics, which opens a large scope of

perspectives when using projection-based semimetrics in non parametric functional prediction, for instance.

Given all those elements, we think that the BAGIDIS semimetric really worth to have a place amongst the semimetrics used in nonparametric functional data analysis. Its automatic adaptivity to the nature of the variations of the patterns in the curves, its ability to deal with horizontal shifts and its capacity to detect the signal in even quite noisy data make it a competitive tool for predicting from curves with sharp patterns.

*

Acknowledgments. The ^1H NMR database used in Subsection 2.3.2 was collected and preprocessed for a study lead by P. de Tullio, M. Frédérich and V. Lambert (Université de Liège). Their agreement for us to use this database is gratefully acknowledged. The name of the concerned illness remains temporarily confidential. We had useful discussions on this project with Frédéric Ferraty and Philippe Vieu, in particular during the Second International Workshop on Functional and Operatorial Statistics. Their comments and encouragements are gratefully acknowledged. The participation of Catherine Timmermans to this workshop has been partially supported by the Fond National de la Recherche Scientifique (Belgium). Part of this work was completed during a stay of Catherine Timmermans at Université d'Orléans-CNRS, that was funded by the Fond National de la Recherche Scientifique (Belgium). Part of this work was completed during a stay of Laurent Delsol at Université catholique de Louvain, that was funded by the Institut de Statistique, Biostatistique and sciences Actuarielles (Université catholique de Louvain). Financial support from the IAP research network grant P 06/03 of the Belgian government (Belgian Science Policy) is gratefully acknowledged.

*

Previous versions of this work include:

Timmermans C. , Delsol L. and von Sachs R., 2011, *Using BAGIDIS in nonparametric functional data analysis: predicting from curves with sharp local features*. ISBA Discussion Paper 2011/20.

Timmermans C., Delsol L. and von Sachs R., 2011, *Bases Giving Distances. A New Semimetric and its Use for Nonparametric Functional Data Analysis*. in Ferraty F. (ed.), *Recent Advances in Functional Data Analysis and Related Topics*, Contributions to Statistics, Physica-Verlag, p. 307-313.

Chapter 3

Spectroscopic case study

The BAGIDIS methodology proposes a distance measure between spectra, that takes into account, in a unified framework, both horizontal shifts and amplitudes variations that might affect spectral peaks. The method relies on the expansion of the spectra in unbalanced Haar wavelet bases. Its opportunity for investigating ^1H NMR spectra in metabonomics is illustrated here in the framework of a study of an eye disease: age-related macular degeneration. Visual analysis, disease detection model and search for biomarkers are proposed here and compared with known methods.

Keywords: spectroscopy, metabolomics, misalignement, wavelet, distance, classification.

*

Advantages of the BAGIDIS methodology for metabolomics analyses.

Illustration with a spectroscopic study of Age-related Macular Degeneration

Catherine Timmermans, Pascal de Tullio, Vincent Lambert, Michel Frédérich, Réjane Rousseau and Rainer von Sachs

Metabonomics and metabolomics studies are analyses which aim at the simultaneous detection of every small weight molecule present in a biofluid, an organ or an organism (see Lindon et al. (2007), for instance, for an introduction to this field). Those “small weight molecules”, the molecular weight of which being typically less than

1500 daltons, are referred as *metabolites*. Numerous biological processes affect the concentrations of metabolites. Compounds of which the concentration is specifically modified by a given process are called *biomarkers* for that process. They can be seen as the *fingerprint* of the biochemical reactions underlying the given process. Identifying biomarkers leads to a better understanding of biological processes, and might help at designing efficient tools for its detection or prediction.

Metabolomics studies become more and more frequent in various area of scientific research (Lindon et al. (2007) gathers some example applications): detection of origin in the food industry, cultivars discrimination in agronomy, toxicological studies in environmental and pharmaceutical sciences, screening of drug candidates in pharmaceutical sciences, diagnostic tool in medicine, etc. The present work is concerned about metabolomics data investigation for biomedical purposes: we aim at discriminating blood serum samples between patients suffering from Age-related Macular Degeneration (AMD) and healthy patients. AMD is an ocular disease, that is a leading cause of vision loss in western countries amongst people aged fifty or older (see Noël et al. (2007), for instance). However, behind this specific application, the methodology we describe has a larger scope and might advantageously be applied on various metabolomics datasets.

Metabolomics datasets often consist in nuclear magnetic resonance spectra (an introduction to this technique can be found in Ross et al. (2007)). From a statistical point of view, those spectra are curves with *sharp local patterns* (“spectral peaks”). Not only their amplitudes but also their locations and shapes are affected by noise, this noise arising from the biological variability of the samples but also from unavoidable changes of the experimental conditions of spectra acquisition. However, most multivariate statistical methods rely on the good alignment of the peaks to be compared (see Chapter 1, for a discussion). Otherwise, false differences might be detected between the spectra. Realignment techniques, such as *dynamic time warping* (Giorgino 2009) have thus been developed, which can be applied as a preamble to the statistical analysis. Those realignment techniques are however imperfect.

In this context, the BAGIDIS methodology (Chapter 1) aims at explicitly and simultaneously taking into account both amplitude variations and horizontal shifts that might affect the patterns in a curve. This methodology relies on the definition of a semi-distance based upon the expansion of the curves in unbalanced Haar wavelet bases. For each curve, an unbalanced Haar wavelet basis is selected so as to hierarchically encode the patterns the curve is made of: the main patterns are supported by the first basis vectors, while subsequent basis vectors support less important ones. Every basis

vector is associated to a specific level change in the curve. Such wavelet bases are associated to each of the spectra, using a sliding window to focus on successive smaller spectral zones. The distance between two spectra is measured as a weighted sum of hierarchically computed differences in both the locations and the amplitudes of the pattern from one spectra to another. Visualization tools, classification procedure and statistical tests can be used, that take into account the BAGIDIS semi-distance.

Given this, we investigate the AMD dataset as follows: we blindly discriminate blood serum samples from healthy and diseased patients; we build a nonparametric model for predicting the AMD health status from blood serum; we select statistically discriminative spectral peaks with a aim to identify AMD biomarkers; we discuss whether statistically discriminative spectral peaks are related to systematic amplitude changes or horizontal variations, or both simultaneously. At each step of our analysis, we discuss how BAGIDIS compares to a recently published statistical analysis of the AMD dataset (Rousseau 2011) and show how this methodology can be used as a useful complement to statistical tools usually used in metabolomics (Rousseau et al. 2008).

This Chapter is organized as follows. Section 3.1 describes the AMD dataset. Section 3.2 discusses its statistical analysis. The BAGIDIS methodology itself is not recalled in this Section, and the reader is invited to refer to Chapter 1 before reading this Section. Section 3.3 concludes.

3.1 The AMD dataset

AMD is an ocular pathology, that can lead to rapid vision loss. It affects central fine vision, needed for reading, driving and face recognition, for instance. This disease exists in two distinct forms, one of which arising from an uncontrolled formation of new blood vessels (*angiogenesis*) under the *macula*, a part of the retina at the rear of the eye. The misknowledge of AMD anthology motivates the search for biomarkers in blood serum samples through a metabolomics approach (Noël et al. 2007). The AMD database was originally collected for a study lead by *de Tullio, Frédérick and Lambert* (Université de Liège). It consists in 200 blood samples, 100 of which arising from AMD patients and the other 100 arising from non-AMD patients (“control” patients), the AMD health status being diagnosed by an ophthalmologist. All AMD patients are aged over sixty and are followed by an ophthalmologist at *Centre Hospitalier Universitaire* in Liège, Belgium. Control patients are age-matched patients in the same hospital, without any sign of ocular disease and not having a known history of AMD.

The database also contains some additional general and clinical information. A complete description of this database can be found in Rousseau (2011).

A one-dimensional ^1H NMR spectrum was acquired from each blood sample, using a 500 MHz Bruker Avance spectrometer. A CPMG sequence with water pre-saturation was applied to attenuate broad signals arising from protein and water. Due to spectral acquisition problem, 6 AMD samples and 1 control sample were removed from the study. The resulting product of an ^1H NMR spectrum acquisition is a time signal called *Free Induction Decay* (FID). In order to chemically interpret the signal, each FID is converted in a spectral signal using a Fourier transform. It results in a set of spectra, which can be seen as kind of periodograms. Before and after this Fourier transform, several other pre-treatment of the signals are also needed for the data to be statistically exploitable. In this study, we used the automatic pre-treatment procedure for metabolomics data which is advised and validated by Rousseau (2011). It includes first order phase correction, suppression of the solvent, apodization by a scale function, apodization by an exponential function, Fourier transform, zero-order phase correction, setting to zero of negative values, warping, conversion in ppm scale, spectral window selection, bucketing, removal of undesired regions, spectral zone aggregation and normalization. More details on the acquisition procedure and on the pre-treatments can be found in Rousseau (2011).

As a result of this procedure, the AMD dataset contains 193 spectra, of which 94 comes from AMD patients. Each spectra is a curve of 600 consecutive intensity measures in a spectral range going from 10 to 0.2 ppm. An example is provided in Figure 3.1.

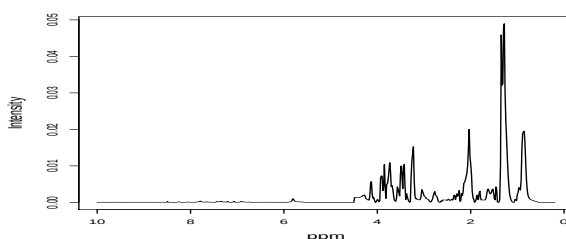


Figure 3.1: A ^1H NMR spectrum of an AMD blood serum, obtained from the FID using the pre-treatment procedure advised by Rousseau, 2011.

3.2 Statistical Analyses

Except if mentioned otherwise, we make use of the BAGIDIS semi-distance with parameters $\Delta = 25$ observational points (i.e. $\Delta = 0.4$ ppm), $p = 2$, $\lambda = 0.5$ and the default value of w_k (Chapter 1, Subsection 1.2.1). Results we obtain are compared with the recent results obtained by Rousseau (2011). We see that additional insight into the data is gained by using the BAGIDIS methodology.

3.2.1 Visual analysis

Figure 3.2, *top left*, provides with the projections of the spectra on the first plane of a principal component analysis (PCA). This representation is to be compared with a *multidimensional scaling* (MDS) representation of the dataset, based upon the BAGIDIS semi-distance (with $\lambda = 0.5$) in Figure 3.2, *top right*. Multidimensional scaling is a projection technique that aims at preserving given distances between the observations in the dataset, so that the proximities of the data in the plane of projection can be interpreted -up to a certain degree- as “real” proximities of the data according to the chosen distance. MDS used jointly with the Euclidean distance corresponds to a PCA. In both representations, points are colored in different values according to their AMD health status.

We see that using BAGIDIS allows for a nearly optimal discrimination between AMD and non-AMD serum spectra, the distinction being essentially encoded by one single axis, while PCA detects an effect of the AMD health status but does not achieve such a clear discrimination. Furthermore, four outliers were detected in the data by Rousseau (2011) using visual inspection of this PCA representation, and a PCA for group-centered data. Those spectra are marked by triangles instead of points in Figure 3.2. They were removed of the dataset in Rousseau (2011). We do not observe such an aberrant behavior for those spectra when using BAGIDIS. This might indicate that a problematic warping, resulting in misalignment, may be the cause of the aberrant behavior observed in the PCA. In this study, we discovered that it is not necessary anymore to discard those spectra from the database. This is in accordance with the fact that visual inspection shows no aberrant behavior.

Figures 3.2 *bottom left and right* are MDS representations, obtained using a balance parameter λ fixed as 0 and 1 respectively. This allows to diagnose the effect of detail differences and breakpoint differences in the distance separately. Although some

information on the AMD health status is clearly contained in the detail differences, we observe the major role of breakpoints location for discriminating the spectra. This might indicate a systematic peak appearance, shape modification of a peak, horizontal shift, or change of sign in the difference of amplitudes of neighbor peaks.

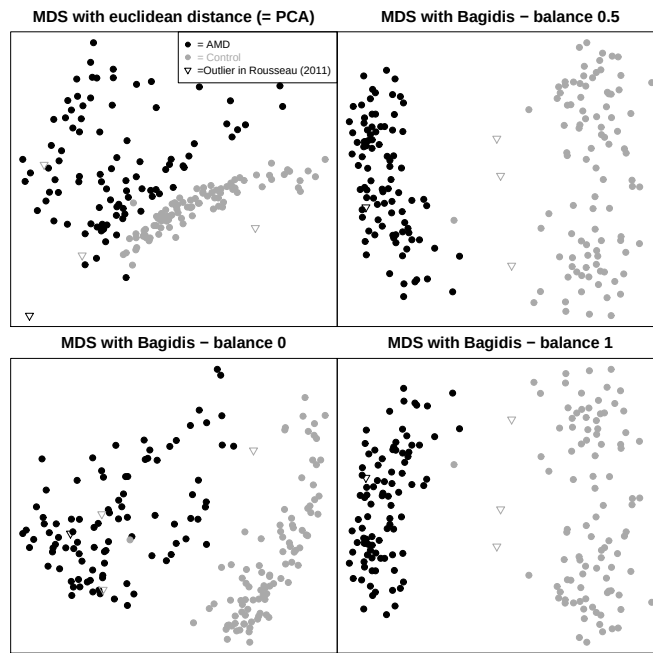


Figure 3.2: PCA representation, as compared with MDS representation using BAGIDIS with balance parameter $\lambda = 0.5, 0$ and 1 respectively. Points are colored according to the AMD health status of the corresponding projected spectrum.

3.2.2 AMD detection

We aim at predicting from the spectrum if a patient is affected by AMD or not. This can be done using the model discussed in Subsection 2.3.2. Here is a reminder of the method and of the achieved results. Readers having that Subsection in mind might prefer skip this part and go to Subsection 3.2.3.

A training set of 150 spectra is randomly selected and a functional nonparametric discrimination model is adjusted (Ferraty and Vieu 2006). This model is a k -nn predictor relying on a matrix of semi-distances between the spectra, with k being cross-

	Occurrences of 0 error out of 43 predictions	Occurrences of 1 error out of 43 predictions
PCA-based semi-distance with $q \geq 6$	40 times out of 80 50%	40 times out of 80 50%
Non-optimized BAGIDIS semi-distance with prior weights and $\lambda = 0.5$	48 times out of 80 60%	32 times out of 80 40%
Optimized BAGIDIS semi-distance (1 non zero weight is selected)	18 times out of 18 100%	0 times out of 18 0%

Table 3.1: Summary results for the prediction of the AMD health status from the spectra.

validated. We consider the adjustment of such a model using BAGIDIS and compared its performances with those obtained using the same model with several other semi-distances: the Euclidean distance, the PCA-based distance, a derivative-based semi-distance, a semi-distance that realigns before computing an Euclidean distance (see Ferraty and Vieu (2006) or Chapter 2 for definitions). In each case, the number of misclassification observed on the remaining 43 spectra is recorded. This test for the prediction of the AMD health status from the spectra is repeated 80 times, with different randomly selected training sets. Results are summarized in Table 3.1, for BAGIDIS and its best competitor, being a PCA-based semi-distance with at least 6 components. We observe that the non-optimized BAGIDIS obtains *no error* 10% more often than the PCA-based semi-distance. Furthermore, we can optimize the weights and the λ parameter of the BAGIDIS semi-distance using a cross-validation procedure within the training set, and the resulting model is tested on the remaining 43 series. This test is repeated 18 times on different randomly selected training sets, and no prediction error occurs. At each repetition, only 1 non-zero weight is selected by the cross-validation procedure. We observe no prediction error in every case, indicating a risk of misclassification that is probably smaller than 0.05. This indicates a very good capacity of discriminating the serum spectra from AMD and healthy patients.

3.2.3 Search for biomarkers

As a last step of the analysis, we aim at identifying AMD biomarkers in the spectra. Six advanced statistical methods for the discovery of metabolomics spectral biomarkers from ^1H NMR spectra have been identified in Rousseau et al. (2008): multiple hypothesis testing (MHT), supervised principal component analysis (s-PCA), supervised independent component analysis (s-ICA), discriminant partial least squares (PLS-DA), linear logistic regression (LLR) as well as classification and regression tree (CART). A description of those methods can be found in Rousseau et al. (2008),

as well as an assessment of their relative performances: recommendation is given to use s-PCA with caution due to its low general efficiency; use of CART is discouraged due to its noise sensitivity; the other four methods are diagnosed promising. All those methods have been applied to the AMD database (with outliers excluded) by Rousseau (2011). For each method, the 20 most significant biomarkers have been identified. Results are presented in Figure 3.3. Here, we compare those results to the ones we obtain using *double geometrical t-tests* (Chapter 1, Subection 1.3.3) based upon the BAGIDIS semi-distance.

The idea of *double geometrical t-tests* is as follows. We first restrict the dataset to sliding segments of the data located in a given range of ppm values, and compute the BAGIDIS distance matrix between those segments. Then, we test for the equality of the means of the distances between two AMD segments (*intra-group distances*) and the distances between one AMD and one non-AMD segment (*inter-group distances*). We also test for the equality of the means of the distances between non-AMD segments (*intra-group distances*) and between one AMD and one non-AMD segment (*inter-group distances*). In both case, the alternative is that intra-group distances are lower than inter-group distances. These tests are performed using Welch t-tests, assuming independence and normal distribution of the distances about their group mean. Combining the results of both tests allows to deduce the relative positions of AMD segments and non-AMD segments. Figure 1.9 might help for the interpretation. Only if both t-tests significantly reject their null hypotheses are the two groups statistically different in mean. In this way, we detect if the selected sliding segment is significantly discriminant with respect to the AMD health status, by requiring a significance $\alpha = 1e - 10$ for both t-tests. This test is actually performed for each sliding segment of the dataset. A Bonferroni correction is thus applied to each p-value to account for the 576 simultaneous comparisons.

Detected differences are differences in the windowed spectral segments, and not at a specific spectral location. This is a difference with competitor methods. A given spectral location contributes therefore to a number of segments equals to Δ . For each spectral location, the number of significantly discriminant segment at which it contributes is computed and reported in Figure 3.4 for different parametrization of the BAGIDIS semi-distance. This number of significances can be used to search for biomarkers. The higher the number of significances, the higher the indication that the related spectral location might be a biomarker. A number of significances equals to Δ ($\Delta = 25$ here) for a given spectral location indicates that each segment where this location contributes is significant with respect to the AMD health status. This clearly

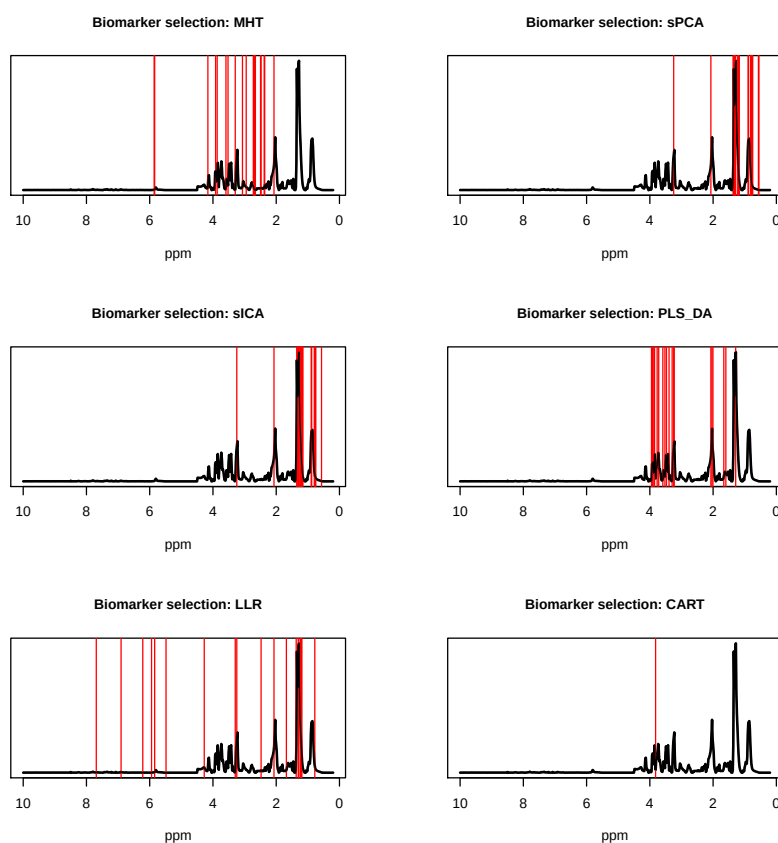


Figure 3.3: Search for AMD biomarkers using MHT, s-PCA, s-ICA, PLS-DA, LLR and classification and CART. For each method the 20 most significant biomarkers identified by Rousseau (2011) are marked by a vertical line. A typical spectrum of the AMD database is superimposed to ease the interpretation of biomarker detection.

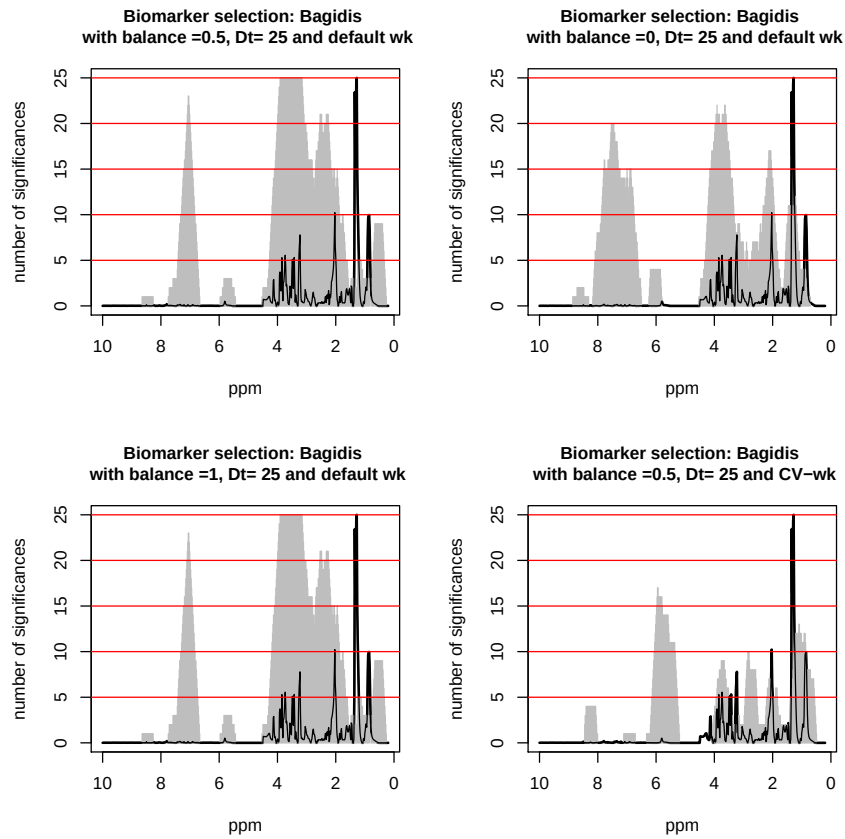


Figure 3.4: Search for AMD biomarkers using BAGIDIS with different parameterizations. For each spectral location, we compute the number of times when a spectral segment significantly discriminant for AMD covers this location. This is indicated in gray in the Figure. A typical spectrum of the AMD database is superimposed (black curve) to ease the interpretation of biomarker detection.

indicates for a biomarker.

Spectral zones from 3.99 to 3.06 ppm, as well as 2.48 and 2.27 ppm are strongly identified as biomarkers in Figure 3.4 (*top, left*). Those spectral zones are also detected by MHT and PLS-DA. A contribution in this zone, located at 3.24 is also detected by LLR, s-ICA and s-PCA, while CART has its only detection at 3.82. A highly significant detection of BAGIDIS also takes place in the spectral zone 7.24 to 6.89 ppm, which also contains a significant detection of LLR at 6.90 ppm. Some detection, although slightly less significant, is also found in the spectral zone 0.64 to 0.39, which is also detected by s-PCA, s-ICA, and, at one single location, by LLR. A detection at 2.07 also occur, which is also identified by all competitor methods except for CART. Possible, less significant, biomarkers are pointed out around 8.47 (no detection by other methods), 5.75 (also with LLR and MHT), 4.27 (also with LLR), 4.13 (also with MHT), 1.67 (also with LLR), 1.59 (also with PLS-DA), 1.36 (also with LLR and s-ICA), 1.22 (also with s-PCA, s-ICA and LLR) and 0.90 ppm (also with s-ICA and s-PCA). Setting the λ balance parameter to 1 (breakpoints only) in the BAGIDIS semi-distance does not significantly modify these results. This is in accordance with the visual analysis in Subsection 3.2.1. Setting $\lambda = 0$ (amplitudes only) suppresses the detection of the spectral zone at ppms lower than 0.64, while the detection of the spectral zone at 1.22 ppm becomes more clear. The relative contributions of the spectral peaks around 5.75 and 2.07 ppm increase. This helps gaining an insight in the way the detected spectral zones do differ in AMD and non-AMD spectra. Finally, Figure 3.4 (*bottom, right*) identify spectral zones of significant differences between AMD and non-AMD spectra when w_k is set to its cross-validated value, as obtained in Subsection 3.2.2. This allows for finding a discriminant, sufficient but not exhaustive, set of biomarkers for AMD, as those biomarkers are the only one which contributes to the AMD detection model calibrated in Subsection 3.2.2.

We summarize this analysis by observing that BAGIDIS detects in one single study nearly all the spectral zones which were detected as biomarkers by at least one of the competitor methods. This emphasizes its consistency and its large scope of detection, which might be valuable for reducing the number of different statistical tools needed in a metabolomics study. Very few detection occur that are not detected by at least one other method, which might be an indication that the method does not increase false detection rate relative to the combined use of competitor methods. From a biological point of view, this study has identified some biochemical pathways that could be implied in the anthology of AMD. Some additional biological experiments are required in order to validate these results.

3.3 Conclusion

This metabolomics study of the AMD dataset using the BAGIDIS methodology has been shown to be a useful complement to recent statistical analyses in the same field (Rousseau 2011). It provides more informative visual discrimination of AMD and non-AMD blood samples which does not highlight outliers with respect to the semi-distance used. It allows for building a detection model for the AMD health status whose performances are shown to be really good. Finally, it allows for detecting - in a single analysis - a large set of biomarkers, this detection otherwise requiring the combination of six advanced statistical methods for biomarker search.

*

Acknowledgments. *The authors gratefully thank Bernadette Govaerts for her interest and encouragements about this work. The statistical analyses are performed using the R software (R Development Core Team 2008), and using the R implementations of the regression estimators provided on the companion website of Ferraty and Vieu (2006), slightly adapted for their use together with the BAGIDIS semi-distance. Financial support from the IAP research network grant P 06/03 of the Belgian government (Belgian Science Policy) is gratefully acknowledged.*

*

Previous versions of this work include:

Timmermans C., de Tullio P., Lambert V., Frédéric M., Rousseau R., von Sachs R., 2012, *Advantages of the BAGIDIS methodology for metabonomics analyses: application to a spectroscopic study of Age-related Macular Degeneration*, Proceedings of the twelve European Symposium on Statistical Methods for the Food Industry - Agrostat 2012, 399-408.

Chapter 4

BAGIDIS and images

This Chapter is concerned with gray level images, with an aim to image comparison. The challenge for extending BAGIDIS to image data is twofold. First, we need an algorithm for automatically selecting a 2-dimensional unbalanced Haar wavelet basis that is well-suited to the image according to a principle of hierarchy. Second, we need a sparse way to encode the resulting wavelet basis, so that unique and reversible signatures of the images can be defined. In other words, the main requirement is an extension of the BUUHWI transform to image data. This extension is not trivial as the notion of neighborhood between the pixels of the image is more general than it is with curves. A graph structure accounting for the neighborhood information is therefore introduced. Moreover, the definition of the signature of an image requires a more careful treatment as the breakpoints between the positive and negative parts of each unbalanced Haar wavelet becomes a “breakline” which adapts to the shapes of the patterns in the image and is not necessarily straight.

In accordance with the above discussion, the key proposal of this Chapter is a transform which encodes an image in a unique, invertible and hopefully sparse way, as a sum of hierarchically ranked pairwise differences of intensity between zones of the image. As expected, the transform is related to the selection of a particular basis within the family of 2-dimensional unbalanced Haar wavelet bases, and this transform leads to the definition of a unique signature for the images. This is the starting point for the definition of the highly flexible image-adaptive BAGIDIS semi-distance between images, that can take into account a possible misalignment of the patterns the images are made of. More generally, this extension applies to any data with an underlying graph structure.

This transform is termed Generalized Bottom-Up Unbalanced Haar Wavelet Transform (G-BUUHWI) so as to underline its relationship with the BUUHWI presented in Chapter 1. It has been reported elsewhere under the shorter name SHape Adaptive Haar (SHAH).

Keywords: multiscale decomposition, edge detection, distance, image, graph, wavelet, regression, classification, functional data.

*

The Generalized Bottom-Up Unbalanced Haar Wavelet Transform with Application to Prediction from Image

C. Timmermans and P. Fryzlewicz

Processing images and efficiently encoding their information content has become an extensive need in our numeric world. In this Chapter, we develop a new paradigm for investigating gray level images. A numeric gray level image is made of a set of pixels with given intensities on a black-to-white scale. In other words, an image is a two-dimensional signal of intensity $z = f(x, y)$, where the pair (x, y) defines the location of a pixel. This work is rooted in the basic observation that the *set of pixel intensities* is necessary but not sufficient to fully describe the image. Indeed, the *geometric organization of the pixels* with respect to each other is a complementary key information to "see" patterns in the image. The method we present here fully takes this into account.

We start by describing an image through an object called *Intensity Network* (IN). Basically, this object is made of a *graph*, of which the nodes are the pixels and the edges define a topology accounting for neighborhood relationships between the nodes, a *codebook* which accounts for the spatial anchorage of the nodes of the graph, and a set of *intensities* associated to the nodes.

In this Chapter, we make use of this description of an image to define a transform encoding the same information in a *sparse, unique* and *reversible* way. It results in what we call the NET-ID, which stands for *Network Intensity Differences*. Because of its strong relationship with unbalanced Haar wavelet basis selection, the transform of the IN into a NET-ID is named *Generalized Bottom-Up Unbalanced Haar Wavelet Transform* (G-BUHWWT).

The NET-ID describes the image in a *hierarchical* way, i.e. the image is described as a ranked sum of level differences between pairs of zones in the image, the most informa-

tive ones being ranked first. Thereby, it provides a natural decomposition of the image as a set of features ordered according to their importance for the image description. This decomposition is as sparse as possible for a gray-level piecewise-constant image. It highlights the edges of the image. In case the image is not piecewise-constant, the NET-ID is an attempt to achieve the same property. Indeed, the G-BUHWHT is defined so as to concentrate the energy of the image as much as possible in the first ranked features of the image decomposition, so as to bring forward significant edges and prominent features of the image. This may lead to a sparse encoding of the significant sharp features of noisy piecewise-constant images or piecewise-smooth images.

Several ways to describe the NET-ID are proposed in this Chapter, each of them opening different perspectives. The first one directly responds to the Intensity Network, as it consists of a graph, a codebook, and a set of intensity differences. This is the more direct way to get the NET-ID in terms of computational load. The second expression of the NET-ID is through an image-adapted orthogonal basis and the projection of the image onto it. This is the best way to visualize the hierarchic decomposition of the image that is induced by the transform, and to realize its sparsifying potential. From this point of view, the transform may be seen as a particular unbalanced Haar basis expansion, and our work might be seen as a follow-up of Fryzlewicz (2007). Finally, the third expression of the NET-ID is the more compact and mathematically efficient as it takes the form of a matrix. This third expression of the NET-ID is the starting point for defining a highly adaptive measure of dissimilarity between images. Because of this key role, we give this expression a distinct name: we refer to it as to the *signature* of the image.

The dissimilarity measure we propose uses a weighted norm to compare the signatures of two images. Although highly flexible, it can be shown to be a semi-distance. Taking into account the wavelet basis expression of the NET-ID, it may be interpreted as follows. Each image is encoded in a different basis that is best suited to it, in the sense that it hierarchically encodes the patterns it is made of. We then compare both the bases themselves and the projection of the images onto it, the comparison being meaningful due to the hierarchy. Therefore, the semi-distance is named BAGIDIS which is an acronym for *Bases Giving Distances*. This original way to proceed for image comparison actually extends the ideas developed in Chapter 1 for curve comparison.

The BAGIDIS semi-distance between images has the following advantages. It compares both the intensities and the locations of the patterns in the image. Thereby, it may be used to capture the similarity of images that might be slightly misaligned, which

is often desirable in practical problems. In the same time, when the signal precisely lies in the location of the patterns, BAGIDIS is able to account for it in a consistent way, what usual methods relying on the preliminary registration of the images can not do. BAGIDIS is also highly flexible and adaptive to the nature of the significant signal in an image dataset: according to the problem at hand, the focus of the distance can be turned onto intensity differences, location differences along one axis or another, or any combination of those at different hierarchic levels in the description of the images.

BAGIDIS might be used to compare images for a wide range of statistical purposes (prediction, classification, dataset visualization, testing, ...) because it can be used in association with any distance-based algorithm. For instance, it can be used together with multidimensional scaling (Cox and Cox 2008), functional nonparametric regression and discrimination models (Ferraty and Vieu 2006) or k-NN algorithms (Hastie et al. 2001). Moreover it is worth emphasizing that the method we propose has a more general scope than images, although we restrict ourselves to images in the present work. Both the G-BUHWWT transform and the BAGIDIS semi-distance can actually be applied to any data that can be encoded as an Intensity Network, i.e. a graph of which the nodes are associated to a given value of intensity and are rooted in a Cartesian space. One may think about a network of communication antennas or, in higher dimension, to video data or measures of temperatures in a three-dimensional space. As far as images are concerned, it is also worth to mention that the G-BUHWWT transform, and hence the BAGIDIS semi-distance, can take into account neighborhood relationships between pixels which are more flexible than the rectangular grid reflecting the matrix-like geometry of the usual storage of a numeric image. The graph structure may reflect the geometry of a disk for instance.

The Chapter is organised as follows. Section 4.1 defines the G-BUHWWT algorithm and presents its properties. Section 4.2 introduces the BAGIDIS semi-distance for comparing images and discusses how this semi-distance might be used together with general distance-based algorithms. Defining the parametrization of the semi-distance is a key point of this Section. Section 4.3 discusses how our work is situated within the huge image processing and graph theory literature. Section 4.4 investigates some simulated and real data examples. Section 4.5 concludes.

4.1 The Transform

The Generalized Bottom-Up Unbalanced Haar Wavelet Transform (G-BUHWWT) is an algorithmic tool allowing to encode images in a sparse data-driven hierarchic way.

Subsection 4.1.1 gives the core ideas of the transform, while Subsection 4.1.2 provides its algorithmic details. Subsection 4.1.3 discusses several useful mathematical and visual properties.

4.1.1 Core ideas

The G-BUHWHT transform requires three pieces of information to describe an image: first, the intensities of the pixels of the image, second a notion of neighborhood between the pixels and third a spatial anchorage for the pixels in the two-dimensional space. It consists in a data-driven iterative smoothing procedure for dimension reduction, in which the steps of the reduction process are encoded and returned as the output of the transform. A schematic illustration of the process is found in Figure 4.1. Here are the core ideas.

An Intensity Network. Consider a gray level image, that is stored as a real matrix of dimension $N \times M$. Then, draw a *network* on this image. Each pixel is a *node* of this network, and each node is related by *edges* to its four nearest neighbor (in the left, right, top and bottom direction respectively). Border nodes have less neighbors. A square lattice is thus created. This graph structure mathematically encodes the idea of neighborhood between the pixels of the image. Associate an orientation to the edges so that each of them consists in an input node and an output node. For the sake of simplicity, give a unique label to each node. Store the mapping relating those labels to the Cartesian coordinates of the pixels in a codebook. Moreover, associate a uniform weight to all nodes of the network, as the information they store (i.e. the value of the related pixel) is *a priori* equally important in the image description. The object resulting from the concatenation of the pixel values (the $N.M$ real values stored in a $N \times M$ matrix) and the graph structure (which consists in $N.M$ nodes and $2N.M - N - M$ edges) rooted in space through the codebook is what we call an *Intensity Network* (IN). Although more details will be provided later in this Section, an example of such a IN can already be found in Figure 4.2.

The Smoothing Process. The idea is to progressively smooth the image. Therefore, replace the value of two neighbor pixels by their (weighted) average value, under the constraint to keep as much information as possible about the initial image. In practice, compute (weighted) differences between pairs of neighbor nodes along each edge. Those differences are called *details*. Identify the smallest difference (in absolute value) and replace the values of the linked pixels by their (weighted) average. Then, reduce those two pixels to a single node in the network, which is given a larger weight due the increased number of pixels it encodes. Finally, update the graph structure

of the network consequently, by suppressing the edge linking the two nodes reduced to one. This reduction process is iterated $NM - 1$ times, up to the point where the image is finally reduced to a single pixel. Figure 4.1 illustrates the first iteration of this smoothing process. Figure 4.4 shows an example of how the graph structure might evolve along the reduction process.

Encoding the Transform. At each iteration of this algorithm, store the labels of the nodes that are reduced as well as the (weighted) difference between them. Each iteration returns thus three values: the input node label, the output node label and the selected detail, the latter being the (weighted) value at the output node minus the (weighted) value at the input node of the edge. There are $NM - 1$ iterations for reducing an $N \times M$ image to a single pixel, associated to a unique real value for the reduced image. The complete reduction process can thus be stored through two vectors: one of them encodes $(NM - 1)$ edges and the other encodes $(NM - 1)$ detail coefficients, that can be interpreted as intensity differences. Both of the vectors are built row by row, from top to bottom. As a top row for them, one adds respectively a degenerated edge linking the remaining node to itself, and the associated value of intensity. Those two vectors additioned with the spatial information stored in the codebook defines the NET-ID of the image. Although more details on the NET-ID will be provided later on in this Section, an illustration can already be found in Figure 4.3.

Overview of the Key Properties. The G-BUHWIT is unique and invertible. It contains the complete information needed to reconstruct the image. It provides an efficient data-driven and highly adaptive encoding of an image, where both the pixels intensities and the image topology are accounted for. It describes the image as a sum of images supporting up-and-down patterns that are hierarchically organized according to the importance of the feature they encode for the description of the image. It is coupled with the definition of a basis of which the basis elements extract the different zones of the image and rank them according to their “importance” for the image description. If the G-BUHWIT is applied to a noiseless image with edges, then those edges and the region of constant intensity they limit are highlighted by the basis elements. This provides for a sparse way to encode the structures in the image. In case the image is noisy, the transform is an attempt to reach the same result and can be used to identify sparsifying projections of the image. The algorithm can be applied to more general geometries than a rectangular image with a quadratic grid: more complex grids might be considered, other shapes for the image might be considered (an image on a disk with a grid related to polar coordinates for instance). Not only images but any multidimensional field of data measured on a grid might be considered (a cubic

volume with a quadratic grid for instance, or on the opposite, a regularly discretized curve).

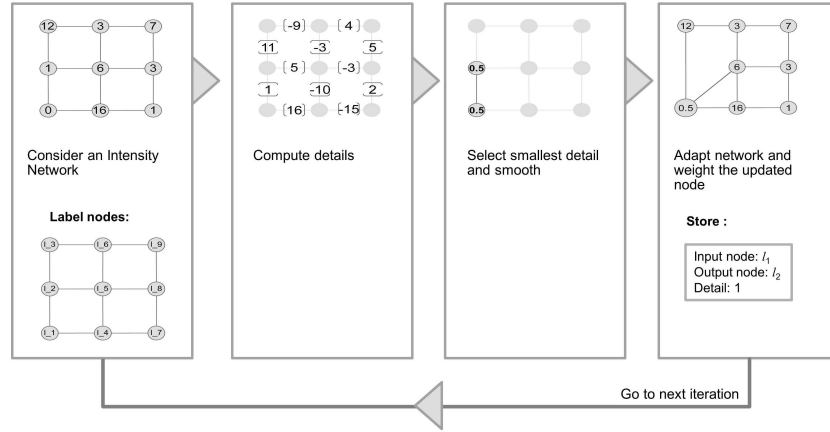


Figure 4.1: A schematic illustration of the G-BUHWHT algorithm. Illustrations refer to the first iteration of the algorithm.

4.1.2 The G-BUHWHT Algorithm

This Subsection is intended to provide algorithmic details on how the G-BUHWHT transform of an image is obtained. The input and output of the algorithm are defined in a formal general way. The iterative procedure at the core of the transform is then described step by step.

Input: an image described as an Intensity Network (IN). The IN of an image $\mathcal{I}$ is defined as a set $\{\mathcal{D}^{(p)}, \mathcal{E}^{\text{IN}}, X\}$, where

- $\mathcal{D}^{(p)}$ is a *codebook*. It encodes the coordinates of p points in the image, identified by labels $\ell_l, l = 1 \dots p$. Those points are the location of the p nodes of the network.
- $\mathcal{E}^{\text{IN}}$ is a *graph*. It is a ranked set of E oriented edges $\epsilon_l = (j, k), l = 1 \dots E$, with $j, k \in \{1, \dots, p\}, j \neq k$ identifying the linked nodes. In the case where no natural orientation exists for the edges, any choice is equally convenient but an orientation is required for the transform to be invertible.

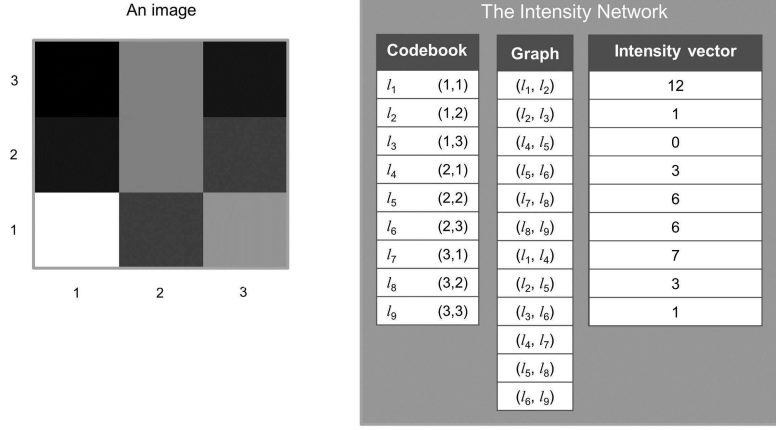


Figure 4.2: On the right: **a typical IN**. On the left: the image which it refers to. The codebook encodes the *location* of the pixels. The graph structure encodes the *neighborhood relationships* between the pixels. The intensity vector encodes the *values* of the pixels.

- $X = \mathcal{I}(\mathcal{D}^{(p)})$ is a *vector of intensities*. It is a real vector of length p encoding the intensities of the image $\mathcal{I}$ at the successive points defined in $\mathcal{D}^{(p)}$.

The IN is thus *a network with p weighted nodes and E unweighted edges*. In summary, we may say that the vector of intensities encodes the *values* of the pixels in an image, the codebook encodes the *locations* of those pixels and the graph structure encodes the *neighborhood relationships* between them. A typical example is as follows.

- The image $\mathcal{I}$ is a gray level image of $N \times M$ pixels encoded as a matrix A .
- $\mathcal{D}^{(p)} = \{(j, k)\}_{j=1 \dots N, k=1 \dots M}$ with j, k defining row and column indexes in A . The points are labeled $\ell_l, l = 1 \dots p$, with $p = N.M$, columnwise in A
- $X = \{X_l\}_{l=1 \dots p}$, where $X_l = a_{jk}$ is the gray level of the pixel with coordinates $\{(j, k)\}$ associated to the label l in A .

An example is provided in Figure 4.2. This typical form of IN will be used for all examples and applications in this Chapter. However, a more complex network might be considered, that would translate into a more complex graph structure $\mathcal{E}^{\text{IN}}$. Different geometries might also be accounted for in the definition of $\mathcal{D}^{(p)}$ and $\mathcal{E}^{\text{IN}}$ when the image has specific geometric features. We might think of an image with a geometry that is best supported by polar coordinates on a disk, for instance. Points sampled randomly in an image might also be considered, with edges defining polygons between them. The key point is actually to remember that the role of the network definition

The Network of Intensity Differences		
Codebook	Graph	Intensity differences
l_1 (1,1)	$\langle l_1, l_1 \rangle$	13
l_2 (1,2)	$\langle l_1, l_2 \rangle$	8.13
l_3 (1,3)	$\langle l_2, l_4 \rangle$	-4.69
l_4 (2,1)	$\langle l_4, l_8 \rangle$	4.04
l_5 (2,2)	$\langle l_4, l_7 \rangle$	-1.73
l_6 (2,3)	$\langle l_4, l_5 \rangle$	-2.44
l_7 (3,1)	$\langle l_8, l_9 \rangle$	1.41
l_8 (3,2)	$\langle l_2, l_3 \rangle$	0.70
l_9 (3,3)	$\langle l_5, l_6 \rangle$	0

Figure 4.3: **A typical NET-ID.** This NET-ID was obtained from the IN described in Figure 4.2.

in the processing of the image is to account for the topological neighborhood between the points of measurements. Caution is thus certainly needed for interpreting the G-BUHWHT when the IN is made of randomly located nodes linked by randomly selected edges.

A degenerated example of IN consists in a set of equispaced points of measurement along a curve, with edges linking consecutive points of measurements. Such a degenerated setting allows to study time series, for instance, or a scalar field along a curve. On the other hand, scalar field of data in a multidimensional space might also be considered. One might think of a cubic volume with a quadratic grid of points associated with a measurement of temperature for instance.

Only the 2-dimensional case (images) is explicitly discussed in this Chapter. However, the extension to the general multidimensional case is direct. As far as the 1-dimensional case (time series) is concerned, the reader is referred to Fryzlewicz (2007) and Chapter 1 where the 1-dimensional BUHWHT is defined, then discussed. Links with those previous works are discussed later on in this Chapter.

Output: A Network with Intensity Differences (NET-ID). The NET-ID of an image $\mathcal{I}$ is defined as the set $\{\mathcal{D}^{(p)}, \mathcal{E}^{\text{OUT}}, d\}$, where

- $\mathcal{D}^{(p)}$ is the same *codebook* as for the IN.
- $\mathcal{E}^{\text{OUT}}$ is a graph. It is a ranked set of p oriented edges $\epsilon_l = (j, k)$, $l = 0 \dots p-1$, with $j, k \in \{1, \dots, p\}$, $j \neq k$ identifying the linked nodes. Edge ϵ_0 is degenerated and $j = k$ for this edge.

- d is a vector of *intensity differences*. It is a real vector of length p encoding the intensity differences associated to the edges successively defined in $\mathcal{E}^{\text{OUT}}$. The components of this vector are labeled with an index $l = 0 \dots p - 1$, as for the edges in $\mathcal{E}^{\text{OUT}}$, with d_0 being an intensity instead of an intensity difference.

As such, the NET-ID of an image is *a network with p weighted oriented edges and p unweighted nodes*. A typical example is provided in Figure 4.3

The algorithm. Algorithmic details of the algorithm are provided below. An illustration can be found in Figure 4.4.

The G-BUHW algorithm

Input

$$\text{IN} = \{\mathcal{D}^{(p)}, \mathcal{E}^{\text{IN}}, X\}$$

Output:

$$\text{NET-ID} = \{\mathcal{D}^{(p)}, \mathcal{E}^{\text{OUT}}, d\}$$

Notations:

Index i designates the iteration of the algorithm;

$\mathcal{E}^{(i)}$ is the set of edges in the network at iteration i : $\mathcal{E}^{(i)} = \{\epsilon_l = (j, k)\}$

$X^{(i)}$ is the value of the nodes remaining in the network at iteration i .

$\omega^{(i)} = \{\omega_l\}_{l=1 \dots p-i}$ is a set of weights associated to the $p - i$ nodes remaining in the network at iteration i .

Initialization:

$$i := 1;$$

$$\mathcal{E}^{(1)} := \mathcal{E}^{\text{IN}}$$

$$X^{(1)} := X$$

$$\text{for } j = 1 \dots p, \omega_j := 1.$$

Iteration #i:

1. Compute details d_l along each of the edges $\epsilon_l = (j, k)$ in $\mathcal{E}^{(i)}$:

$$d_l := \frac{\omega_j}{\sqrt{\omega_j^2 + \omega_k^2}} X_k - \frac{\omega_k}{\sqrt{\omega_j^2 + \omega_k^2}} X_j.$$

2. Select an edge ϵ_{l^*} with minimum absolute value of detail:

$$l^* := \arg \min_l |d_l|.$$

In case of multiple equal minimum values $|d_l|$, select the smallest index l . Note $\epsilon_{l^*} = (j^*, k^*)$.

3. Smooth:

$$X_{j^*} := \frac{\omega_{j^*} X_{j^*} + \omega_{k^*} X_{k^*}}{\sqrt{\omega_{j^*}^2 + \omega_{k^*}^2}}.$$

$$X_{k^*} := X_{j^*}.$$

4. Encode the partial value of the NET-ID:

$$\mathcal{E}_{p-i}^{\text{OUT}} := (j^*, k^*).$$

$$d_{p-i} := d_{l^*}.$$

5. Reduce the network and prepare next iteration:

Update $\mathcal{E}^i$ by replacing all indexes k^* by j^* .

Discard duplicated edges in $\mathcal{E}^i$, keeping only the first occurrence of each edge.

This defines $\mathcal{E}^{(i+1)}$.

$$\omega^{(i+1)} := \{\omega_1, \dots, \omega_{j^*-1}, \sqrt{\omega_{j^*}^2 + \omega_{k^*}^2}, \omega_{j^*+1}, \dots, \omega_{k^*-1}, \omega_{k^*+1} \dots \omega_{p-i+1}\}.$$

$$X^{(i+1)} := \{X_1, \dots, X_{j^*-1}, X_{j^*}, X_{j^*+1}, \dots, X_{k^*-1}, X_{k^*+1} \dots X_{p-i+1}\}.$$

$$i := i + 1.$$

6. Back to Step 1, until $\text{length}(X^{(i)}) = 1$.

Final-step:

$$\mathcal{E}_0^{\text{OUT}} := (j^*, j^*).$$

$$d_0 := \frac{X^{(p)}}{\sqrt{p}}.$$

Computational complexity. Suppose, there are p pixels and $E > p$ edges. The algorithm requires $p - 1$ iterations plus one final step, and $(E - i)$ edges are examined at each iteration. The computational complexity is then $\mathcal{O}(p.E)$. If the linkage is quadratic on a matrix of dimension $N \times M$, with $p = N.M$ pixels, the number of edges is $E = 2MN - N - M$, so that the computational complexity is $\mathcal{O}(p^2)$. Ways to reduce the computational complexity might go into the following directions: examining a fixed number of edges at each iteration, using a two-stage or multi-stage algorithm or removing multiple nodes at once. Here are some comments.

- *Examination of a fixed number of edges.* One might examine only a predefined number of edges (not exceeding a constant), at each iteration i . The edges can be selected in a deterministic or random way. This would result in a computational complexity $\mathcal{O}(p)$.
- *Two- or multi-stage algorithm.* For an image of size $N \times N$, one might divide the image into $(N/a)^2$ non-overlapping sub-images, each of size $a \times a$. The G-BUHW algorithm is then applied on each sub-image separately (stage

1), and then executed on the resulting $N/a \times N/a$ matrix of coefficients d_0 from each sub-image (stage 2). The computational complexity is then $\mathcal{O}((N/a)^2 a^4 + (N/a)^4)$, which attains its minimum when $a = N^{1/3}$, resulting in the complexity of $N^{8/3} = p^{4/3}$. The algorithm can be executed similarly in more stages than one, bringing the computational complexity arbitrarily close to linear, if the number of stages is large enough.

- *Removal of multiple nodes at once.* In the G-BUHWHT algorithm, one node is removed at each iteration. Multiple nodes, corresponding to a fixed number of smallest detail values, might be removed at each single iteration. For example, removing half of the remaining nodes at each iteration would bring the complexity of the algorithm within a logarithmic factor of linear.

Storage needs. As far as the storage of the NET-ID is concerned, we may note that the codebook used in the IN and in the NET-ID is actually the same. In case several images share the same codebook, as it usually is with regular matrix-like structured images, the codebook may be stored only once, and does not need to explicitly be returned in the NET-ID, which saves the storage needs of the transform.

4.1.3 Mathematical and Visual Properties

The G-BUHWHT benefits from the following properties.

Property 1. The G-BUHWHT consists in the definition of a data-driven orthonormal matrix $e = \{e_j\}_{j=1\dots p}$ for changing from basis in $\mathbb{R}^p$, and this change of basis is applied to the vector X .

Proof. The proof of this property takes place in Appendix C. $\square$

Property 1 supports for introducing the following notations and results. We denote by $e = \{e_j\}_{j=1\dots p}$ the canonical basis of $\mathbb{R}^p$ in which X is initially encoded. The data-driven matrix of basis change is denoted by ${}_{\psi}(\mathbf{1})_e$ and defines a basis $\{\psi_k\}_{k=0\dots p-1}$.

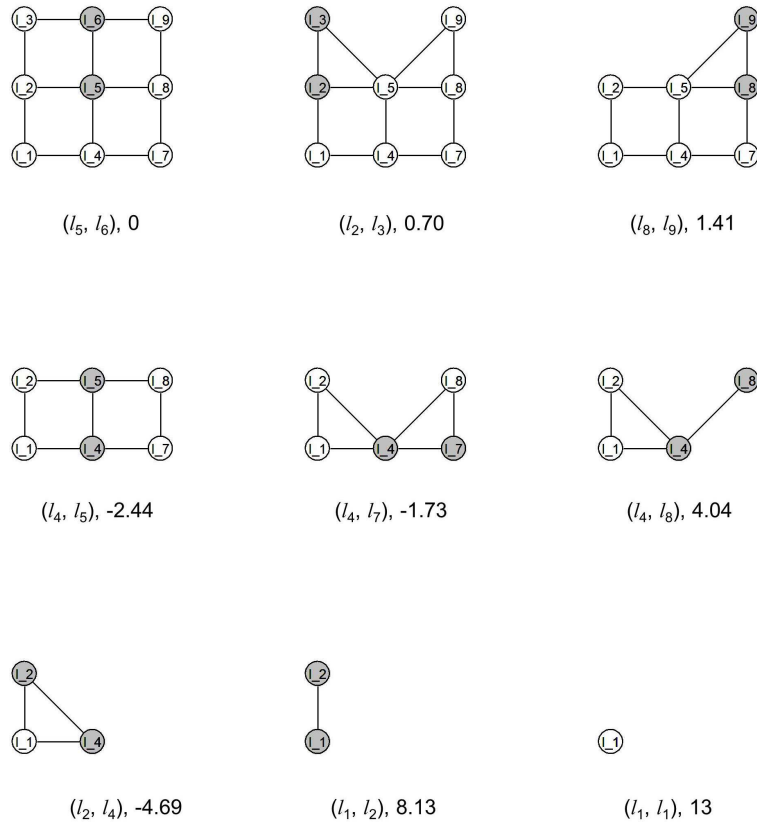


Figure 4.4: A schematic illustration of the G-BUHW algorithm. This illustrates the G-BUHW algorithm applying to the image provided in Figure 4.2. The network is iteratively reduced by one node at each iteration. The selected nodes for reduction are indicated in gray. The labels of the input and output nodes as well as the detail coefficients that are returned at iteration k are indicated below the image of the network.

The vector of intensity differences d is obtained as $d = {}_{\psi}(\mathbf{1})_e X$, and we have

$${}_{\psi}X = \sum_{k=0}^{p-1} d_k \psi_k.$$

The basis vector ψ_0 is a constant one. Each of the other basis vectors ψ_k , $k = 1 \dots p-1$, has elements having one of three distinct values: zero, a positive value α_k , and a negative value β_k . Here is an example in dimension $p = 14$:

$$\psi_k = (000\alpha_k\alpha_k0\alpha_k\beta_k\beta_k00000).$$

Values α_k and β_k , $k = 1 \dots p-1$, are fixed in such a way that all ψ_k have norm 1 and mean 0. Moreover, if $k < l$ then the support of ψ_l is entirely contained in either the support of the positive part of ψ_k , the support of its negative part, or outside both supports. The following matrix is an example.

$$\begin{array}{c} \psi_0 \\ \vdots \\ \psi_{p-4} \\ \psi_{p-3} \\ \psi_{p-2} \\ \psi_{p-1} \end{array} \begin{pmatrix} \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 \\ 0 & \dots & 0 & \alpha_{p-4} & 0 & \dots & 0 & \alpha_{p-4} & \alpha_{p-4} & \beta_{p-4} & 0 & 0 \\ 0 & \dots & 0 & \alpha_{p-3} & 0 & \dots & 0 & \alpha_{p-3} & \beta_{p-3} & \beta_{p-3} & 0 & 0 \\ \dots & 0 & \alpha_{p-2} & \beta_{p-2} & 0 & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & 0 & \alpha_{p-1} & 0 & \dots & 0 & \beta_{p-1} & 0 & \dots & 0 \end{pmatrix}$$

The resulting basis is orthonormal. Each component of the vectors ψ_k is linked to a given node of the network. By construction, the positive and negative values refer to nodes that form two contiguous zones according to the graph structure $\mathcal{E}^{\text{IN}}$.

In accordance with Property 1, the NET-ID can alternatively be expressed as

$$\text{NET-ID} = \{\mathcal{D}^{(p)}, {}_{\psi}(\mathbf{1})_e, {}_{\psi}X\}.$$

This is the second formulation we propose for the NET-ID.

Property 2. The G-BUHWHT consists in an orthonormal basis function expansion, in a basis determined by the network geometry:

$$\mathcal{I} = \sum_{k=0}^{p-1} d_k \Psi_k. \quad (4.1.1)$$

Proof. This trivially derives from Property 1 and from the fact that $\mathcal{D}^p$ defines a mapping between the nodes of the network and points in a spatial domain. The basis $\{\Psi_k\}_{k=0\dots p-1}$ is directly constructed from the basis $\{\psi_k\}_{k=0\dots p-1} = {}_\psi(\mathbf{1})_e \{e_j\}_{j=1\dots p}$ and from $\mathcal{D}^{(p)}$. $\square$

Property 2 allows to represent the transform of an image through the set of basis functions $\{\Psi_k\}_{k=0\dots p-1}$, associated with the projection of the image onto it. An example is provided in Figure 4.5. For each $k = 1 \dots p-1$, the basis matrix Ψ_k is non-zero only in a pair of adjacent zones, one of which supporting for a constant positive value α_k and the other for a constant negative value β_k . Those values α_k and β_k are exactly the same as for the related ψ_k , by construction. Each matrix Ψ_k , $k = 1 \dots p-1$ has norm 1 and mean 0. As for the vectors $\{\psi_k\}_{k=0\dots p-1}$ defined through Property 1, we have that if $k < l$ then the support of Ψ_l is entirely contained in either the support of the positive part of Ψ_k , either the support of its negative part, or outside both support.

Property 2 implies that the G-BUHWHT makes the image $\mathcal{I}$ encoded as a sum of one constant image ($d_0\Psi_0$) and a linear combination of contrast functions between contiguous zones in the image ($\sum_{k=1}^{p-1} d_k\Psi_k$). Due to the Parseval equality, the total energy of the image (i.e. the sum of the squared pixel values $\sum_{l=1}^p X_l^2$) is preserved by this decomposition and can alternatively be obtained as $\sum_{k=0}^{p-1} d_k^2$. The power of the particular contrast supported by a function Ψ_k is measured by d_k^2 , for any $k = 1 \dots p-1$.

Corollary 1. The G-BUHWHT is unique and invertible. $\square$

Starting from the second formulation of the NET-ID (as induced by Property 1), inverting the transform basically consists in inverting the orthonormal matrix of basis change ${}_\psi(\mathbf{1})_e$ and applying the inverse to ${}_\psi X$. As far as the initial form of the NET-ID is concerned, the main step of the inverse transform is to reconstruct the matrix ${}_\psi(\mathbf{1})_e$. Appendix C explains how to proceed.

Property 3. The basis $\{\Psi_k\}_{k=0\dots p-1}$ is an unbalanced Haar wavelet basis.

Proof. Unbalanced Haar wavelet bases have been introduced by Girardi and Sweldens (1997) as a family of bases that are made of one constant normalized vector and a set

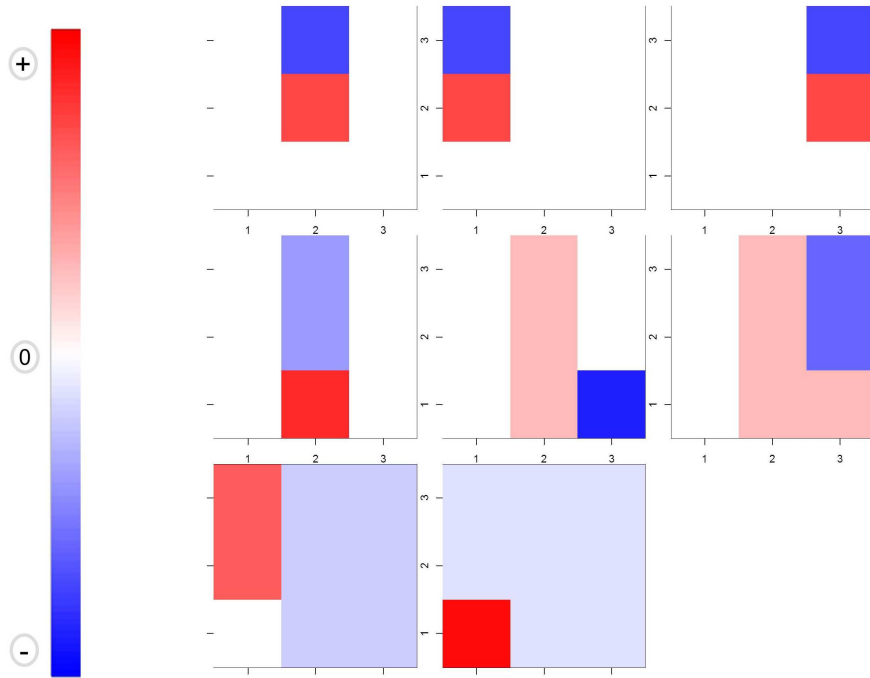


Figure 4.5: **An illustration of Property 2.** This Figure shows the basis functions $\{\Psi_k\}_{k=p-1 \dots 1}$ determined by the G-BUHWI of the image presented in Figure 4.2. The basis functions are provided in the order of their construction, from rank $k = p - 1$ (*top, left*) to rank $k = 1$ (*bottom right*). The remaining basis function Ψ_0 is constant. The basis functions are orthonormal and, except for Ψ_0 , they support level changes between contiguous zones in the image.

of Haar-like wavelets - i.e. orthonormal wavelets made of translations and dilations of an up-and-down pattern of which the level change is not necessarily located at the middle of the support. The form of the vectors of $\{\psi_k\}_{k=0\dots p-1}$ associated with the fact that the indexes of positive and negative parts in the vectors are defined from $\mathcal{E}^{\text{IN}}$, which encodes neighborhood relationships, makes it clear that the resulting basis $\{\Psi_k\}_{k=0\dots p-1}$ is made of one constant normalized function and a set of up-and-down functions encoding neighbor zones in the image. This makes $\{\Psi_k\}_{k=0\dots p-1}$ an unbalanced Haar wavelet basis. $\square$

Two-dimensional unbalanced Haar wavelets generalize well-known two-dimensional Haar wavelets and share with them the good capacity to efficiently capture sharp patterns. As for the one-dimensional case, the key difference is that two-dimensional unbalanced Haar wavelets overcome the dyadic restriction that is traditionally attached to wavelets. This implies that the supports of the basis images $\Psi_k, k = 1 \dots p-1$ may have arbitrary continuous shapes, with arbitrary “breaklines” delimiting the adjacent positive and negative zones. See Figures 4.5, 4.6 and Subsection 4.4.1 for examples. As is clear from Properties 1 and 2, the particular shape of each Ψ_k is determined adaptively from the image itself.

Property 3 opens large perspectives for the use of the G-BUHW as wavelets have been shown a very useful tool in numerous problems. In particular, unbalanced Haar wavelets, which overcome the dyadic restriction inherent to classical wavelets, have lead to several advanced methodologies (Fryzlewicz (2007), Chapter 1, Fryzlewicz (2012)). The G-BUHW opens thus the perspective to generalize those works to images or multidimensional fields. In particular, the definition of a dissimilarity measure between images of which the patterns are not necessarily aligned is investigated in this Chapter.

Corollary 2. The BUHW (Fryzlewicz 2007) is a particular case of G-BUHW.

Proof. The proof is trivial with the following choices:

- $cD^{(p)}$ associates labels $\ell_l, l = 1, 2, \dots, p$ to the successive points of measurements of a time series.
- $\mathcal{E}^{\text{IN}} = \{(1, 2), (2, 3), (3, 4), \dots, (p-1, p)\}$.
- $X = (x_1, x_2, \dots, x_p)$ is a time series. $\square$

Property 4. The G-BUHWIT induces a property of hierarchy in the resulting basis function expansion $\mathcal{I} = \sum_{k=0}^{p-1} d_k \Psi_k$. This means that the ranking of the basis vectors Ψ_k reflects the decreasing importance of the patterns they encode, for the description of the image.

Proof. The proof of this Property takes place in Appendix C. $\square$

The notion of hierarchy we are referring to is actually defined by the G-BUHWIT. As mentioned in Chapter 1, the hierarchy can be interpreted as an *ordering of the features according to the power of the supported up-and-down patterns in the image, under a functional constraint*. The G-BUHWIT may indeed be seen as an agglomerative algorithm (Härdle and Simar 2003) which groups pixels progressively by minimizing the absolute value of their difference. This implies the ordering of the basis vectors according to the power of the supported patterns. However, the agglomerative algorithm is highly constrained here by the graph structure encoding the topology of the image and only neighbor (groups of) pixels can be grouped together. This is the functional constraint on the agglomerative process. This constraint implies that the absolute values of the intensity differences are not necessarily decreasing as the index k increases along the hierarchy. It is also this constraint that provides the algorithm with the following important property.

Corollary 3. The number of non-zero detail coefficients in the G-BUHWIT transform of an image $\mathcal{I}$ is the number of zones of contiguous identical values, the notion of contiguity being defined by the linkage structure of the network. Those non zero detail coefficients are ranked before the zeros in d .

Proof. Explicitly using the mapping $\mathcal{D}^{(p)}$ in the G-BUHWIT algorithm as in the proof of Property 1 trivially shows this Corollary. $\square$

An illustration of Corollary 3 is provided in Figure 4.6. This Corollary 3 indicates that, in case of a piecewise-constant noiseless image, the edges limiting zones of constant intensities are highlighted by the transform, and the transform defines a basis where they are encoded in a sparse way. In case the image is noisy or piecewise-smooth, the algorithm is an attempt to efficiently do the same. Only the first basis elements in the hierarchy identify edges and zones that are expected to be significant in the image. The further we go in the hierarchy the more the related edges might be expected

to originate from noise. An illustration of this property will be provided in Section 4.4.

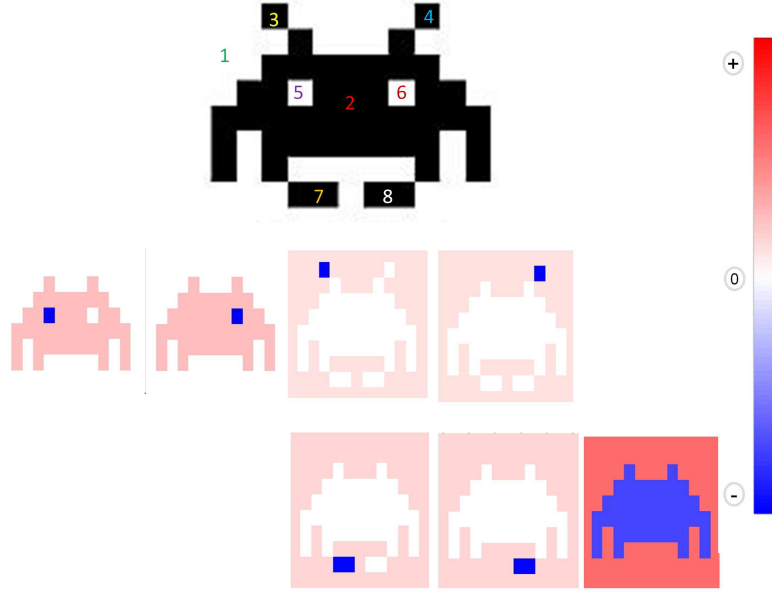


Figure 4.6: **An illustration of Corollary 3.** We consider the popular image of a “space invader” (top row). The IN for this image uses a quadratic network. From this point of view, it is a black and white image with 8 zones of contiguous identical values. Those zones are numbered in the image. The second and third rows of the Figure are the basis functions associated with non-zero detail values, from rank $k = 7$ (second row, left) to $k = 1$ (third row, right). The last basis function associated to a non-zero detail is Ψ_0 which is constant and is not represented.

Property 5. The NET-ID might be encode through a matrix in $\mathbb{R}^{p \times 5}$, that we refer to as its *signature*. This signature is uniquely determined and is an invertible representation of the image.

Proof: The signature of a NET-ID is a matrix defined by combining the information of $\mathcal{D}^{(p)}$, $\mathcal{E}^{\text{OUT}}$ and d into a single object. Each row of the matrix is related to a rank $k = 0 \dots p - 1$ in expansion (4.1.1). At each rank k , the first two columns encode the Cartesian coordinates (x_1, y_1) of the input node at rank k , as determined by $\mathcal{E}^{\text{OUT}}$ and $\mathcal{D}^{(p)}$. Columns 3 and 4 are the Cartesian coordinates (x_2, y_2) of the output node at rank k . Column 5 is the vector d . The signature is thus fully determined once the NET-ID is known. Conversely, the signature determines the NET-ID nearly uniquely - only the names used for labeling the nodes are arbitrarily chosen. The signature is

	Input (x_1, y_1)		Output (x_2, y_2)		d	Rank
Construction	l_1	(1,1)	l_1	(1,1)	13	
	l_1	(1,1)	l_2	(1,2)	8.13	
	l_2	(1,2)	l_4	(2,1)	-4.69	
	l_4	(2,1)	l_8	(3,2)	4.04	
	l_4	(2,1)	l_7	(3,1)	-1.73	
	l_4	(2,1)	l_5	(2,2)	-2.44	
	l_8	(3,2)	l_9	(3,3)	1.41	
	l_2	(1,2)	l_3	(1,3)	0.70	
	l_5	(2,2)	l_6	(2,3)	0	

Figure 4.7: A schematic illustration of Property 5. This is the signature obtained for the image presented in Figure 4.2.

thus an invertible representation of the image, as the G-BUUWT is invertible. $\square$

The signature is the third formulation we propose for the NET-ID. An example of such a signature is provided in Figure 4.7. The rows of the signature matrix are labeled by a rank index $k = 0 \dots p - 1$ referring to the hierarchy. The columns of the signature matrix are labeled (x_1, y_1, x_2, y_2, d) . It follows that the elements of the k^{th} row are denoted $(x_{1k}, y_{1k}, x_{2k}, y_{2k}, d_k)$. By construction of the signatures, the small k indexes refer to the most important features of the images. This will be used in Section 4.2 when defining the BAGIDIS semi-distance.

Property 5 requires a comment about the dimensionality of the transform. At first sight, it might seem that the encoding of an image made of p pixels through a signature of dimension $p \times 5$ is a mapping of an object in $\mathbb{R}^p$ into a higher dimensional space $\mathbb{R}^{5p}$. However, this point of view misses the fact that an image actually is a functional object, i.e. an object which is deeply anchored into a topology, inducing a notion of neighborhood between the pixels. The fact that there is a “real” image behind the matrix of pixels values is an information that is worth being taken into account for the statistical processing of the image. This is precisely what the G-BUUHWT intends to provide for a tool. The notion of neighborhood between the pixels has a central place in the algorithm and the G-BUUHWT actually reduces a double-infinite dimensional problem to a $5p$ -dimensional one without losing its functional nature.

4.2 A BAGIDIS Semi-distance Between Images

The BAGIDIS semi-distance was introduced in Chapter 1 as a data-driven and highly adaptive way for comparing curves with possibly misaligned sharp local patterns. Its performances on simulated and real datasets of curves (Chapters 1, 2 and 3) as compared with state-of-the-art competitors motivated us to explore its generalization for comparing images.

This Section is organized as follows. A definition is first given Subsection 4.2.1 followed by a discussion in Subsection 4.2.2. Subsection 4.2.3 is about methodology. It discusses how BAGIDIS can be used together with distance-based algorithms for statistically investigating datasets of images.

We mention here some properties of the BAGIDIS semi-distance for images. The semi-distance is able to take into account the variations of the patterns in an image both regarding their intensities and their locations, in a unified framework. This makes it highly efficient for comparing images of which the local features might possibly be misaligned. The semi-distance relies on a set of parameters that makes it highly adaptive. This parametrization should be used to extract relevant information of reduced dimension about the compared images. Varying this parametrization might be used as a diagnostic tool to investigate the nature of the variation of the patterns in images across a dataset. Optimizing the parametrization in the framework of a given learning problem is a data-driven way to sparsify the resulting model.

Finally, we may underline that the BAGIDIS semi-distance might be used for comparing a wide range of images as it does not require any preliminary knowledge of what the images represent. The only requirement is about their possibility to be meaningfully encoded in a hierarchic way, as will be discussed later on in this Section. This makes the method highly more general than methods based on *landmarks*, for instance, which are designed according to particular types of images and are really powerful to deal with them but cannot be extended to other datasets. The method also distinguishes from methods that define a distance between *features* computed in the images. Those methods start by reducing the dimensionality of the images and compute a distance between them afterwards. We do both simultaneously, so that the dimension reduction can be performed in an adaptive way.

4.2.1 Definition

The BAGIDIS semi-distance provides a measure of the dissimilarity of pairs of images observed on the same grid of measurements. The compared images are first described through their IN, with an identical codebook and an identical graph component. The G-BUHWHT transform is then computed for each of the images and their NET-ID are encoded in the form of signatures. The BAGIDIS semi-distance takes advantage of the property of hierarchy of the G-BUHWHT. Components $(x_{1k}, y_{1k}, x_{2k}, y_{2k}, d_k)$ of the signatures of two images are compared to each other rank by rank, and the resulting differences can be suitably weighted.

The main definition. We propose the following measure for the dissimilarity between two images $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(2)}$:

$$\begin{aligned} d^{\text{BAGIDIS}}(\mathcal{I}^{(1)}, \mathcal{I}^{(2)}) &= \sum_{k=0}^{p-1} \left(\alpha_k (x_{1k}^{(1)} - x_{1k}^{(2)})^2 + \beta_k (y_{1k}^{(1)} - y_{1k}^{(2)})^2 + \gamma_k (x_{2k}^{(1)} - x_{2k}^{(2)})^2 \right. \\ &\quad \left. + \delta_k (y_{2k}^{(1)} - y_{2k}^{(2)})^2 + \epsilon_k (d_k^{(1)} - d_k^{(2)})^2 \right)^{1/2} \\ &= \sum_{k=0}^{p-1} \left\| \mathbf{s}_k^{(1)} - \mathbf{s}_k^{(2)} \right\|_{2, \mathbf{p}_k} \end{aligned} \quad (4.2.1)$$

where

- $\mathbf{s}_k^{(i)} = (x_{1k}^{(i)}, y_{1k}^{(i)}, x_{2k}^{(i)}, y_{2k}^{(i)}, d_k^{(i)})$ is the k^{th} rank of the signature $\underline{\mathbf{s}}^{(i)}$ of the image $\mathcal{I}^{(i)}$ with $i = 1, 2$.
- $\mathbf{p}_k = (\alpha_k, \beta_k, \gamma_k, \delta_k, \epsilon_k)$ is a vector of parameters in $\mathbb{R}_+^5$.

The dissimilarity measure is thus parametrized by $\underline{\mathbf{p}} = \{\mathbf{p}_k\}_{k=0 \dots p-1}$. Notation $\left\| \mathbf{s}_k^{(1)} - \mathbf{s}_k^{(2)} \right\|_{2, \mathbf{p}_k}$ indicates that the contribution of rank k to the distance is a semi-2-norm, weighted by the parameters $\mathbf{p}_k$. The fact that it is a semi-norm instead of a norm is because, although non-negative, the parameters $\alpha_k, \beta_k, \gamma_k, \delta_k$ and ϵ_k can nevertheless be 0 so that $\left\| \mathbf{s}_k^{(1)} - \mathbf{s}_k^{(2)} \right\|_{2, \mathbf{p}_k} = 0$ does not necessarily imply $\mathbf{s}_k^{(1)} = \mathbf{s}_k^{(2)}$. Therefore, expression (4.2.1) defines a semi-distance instead of a distance. This means that it satisfies the triangular inequality and that the semi-distance between an image and itself is necessarily zero.

A more constrained form of the semi-distance. The following more constrained form of the semi-distance might be more intuitive:

$$\begin{aligned}
 d^{\text{BAGIDIS}}(\mathcal{I}^{(1)}, \mathcal{I}^{(2)}) &= \sum_{k=0}^{p-1} w_k \left(\lambda_x ((x_{1k}^{(1)} - x_{1k}^{(2)})^2 + (x_{2k}^{(1)} - x_{2k}^{(2)})^2) \right. \\
 &\quad \left. + \lambda_y ((y_{1k}^{(1)} - y_{1k}^{(2)})^2 + (y_{2k}^{(1)} - y_{2k}^{(2)})^2) \right. \\
 &\quad \left. + \lambda_D (d_k^{(1)} - d_k^{(2)})^2 \right)^{1/2} \quad (4.2.2)
 \end{aligned}$$

with w_k a non negative weight function and $\lambda_x, \lambda_y, \lambda_D$ being scaling parameters along the x, y and detail directions respectively, such that $0 \leq \lambda_x, \lambda_y, \lambda_D \leq 1$ and $\lambda_x + \lambda_y + \lambda_D = 1$. Varying the relative proportions of λ_x, λ_y and λ_D allows for a better insight into the kind of variations that are compared in images: shifts or shape modifications along the x axis, along the y axis, change in the intensities, or a combination of two or three of the latest. This second definition is a direct extension of the BAGIDIS semi-distance between curves defined in Chapter 1.

4.2.2 Discussion

We may have the following discussion.

Interpretation. The idea of the BAGIDIS semi-distance is that a signature defines the path of a point in $\mathbb{R}^5$. The distance aims thus to measure the dissimilarity between those paths. It is the hierarchy induced by the G-BUHWHT that makes the paths comparable from one image to another. Intuitively, one might expect the beginning of the path to reflect the typical features of an image as it refers to its most important patterns, according to the hierarchy. Further along the path, the behavior of the point in $\mathbb{R}^5$ is expected to become more and more related to individual features of the image as it encodes patterns of decreasing importance and ultimately noise.

Feature extraction and parametrization. This above prior intuition might be introduced in the distance through a prior choice of the coefficients $\underline{p}$. The key is that the semi-distance should extract relevant information about the discriminative patterns in the images. This relevant information is expected to have reduced dimension. Therefore, a sparse choice of the parameters $\underline{p}$ might be achieved, which is desirable. Next Subsection further discusses how the parameters can be selected in practical settings.

In a certain way parameterizing the distance implies to perform denoising and feature extraction simultaneously. The relative importance of $\alpha_k, \beta_k, \gamma_k, \delta_k$ and ϵ_k might also be chosen to tune the semi-distance for capturing specific features according to a prior idea of what makes a group of images different or similar. It can also be used as an exploratory tool for diagnosing what makes images different. Definition (4.2.2) might be more easy to deal with in this respect. These possibilities make the BAGIDIS semi-distance a very flexible tool.

Properties. The BAGIDIS semi-distance is data-driven and functional, as the G-BUHWHT transform is. It fully takes into account the notion of neighborhood between the pixels which allows to identify important zones in the image and ultimately to compare them in a simple way. A key property is that both the locations and intensities of the patterns are compared by the BAGIDIS semi-distance. Finally, we emphasize the fact that BAGIDIS does not require any preliminary smoothing or registration of the images.

Extensions and limits: the notion of hierarchy. We might imagine other kinds of functions of the signatures. The key is to rely on hierarchy. It is the hierarchy that allows for the major originality of BAGIDIS which is the comparison of images through their projections onto different image-adapted bases. Therefore, a requirement for datasets of images to be meaningfully compared using BAGIDIS is that the structures in images satisfy a certain notion of hierarchy regarding their significant patterns. This means that no confusion should occur in the hierarchic ranking of the zones of the image. This requirement ensures that, in case a similar pattern appears twice at different locations in both compared images, the more closely located patterns are compared to each other. Chapter 1 discusses this in the case of curve comparison and this discussion can directly be extended for images. Although we can not provide at the moment for a mathematical characterization of the class of functions that satisfy this property of hierarchy, we can consider the following empirical description. The image should not have two different patterns of similar non-noisy intensity and similar neighborhood, and of which the ranking of the mean intensities would randomly change because of the noise. Most of the time, cases where this difficulty occur can easily be solved by looking at a localized version of the distance. The images are splitted in smaller images, which possibly overlap and do not have problems of hierarchy. A BAGIDIS semi-distance is then computed for the related pairs of small images and the overall distance between the two original images is defined as their average. Using such a localized version of the BAGIDIS semi-distance may also be used to diagnose which parts of the images are significantly different from one image to another as one may map the so-computed local relative measures of dissimilarity. Finally one may note that it is argued in Chapter 1 that, even in case some confusion remain in the

hierarchic description, it does not make the semi-distance unable to efficiently discriminate two groups of images, for instance. It will lead to an artifact in the dataset representation. The reader is referred to Chapter 1 for a detailed discussion.

4.2.3 Methodology

The BAGIDIS semi-distance has been designed for statistically investigating datasets of images. It can fit within any distance-based method for dataset visualization, unsupervised discrimination or for prediction in a supervised setting: multidimensional scaling (Cox and Cox 2008), Ward's hierarchic algorithm (Lebart et al. 2004) and nonparametric functional regression (Ferraty and Vieu 2006) are examples of algorithms that can be advantageously used with BAGIDIS in those different areas. Furthermore, some descriptive statistics and tests have been specifically designed for being used with BAGIDIS for curves in Chapter 1, which have a direct extension for images. Nevertheless, a key point in using BAGIDIS in any of those frameworks is the selection of the parametrization to be used in either Equation (4.2.1) or Equation (4.2.2). It can be chosen as a prior for having a first insight into a problem, as discussed above, or as a diagnostic tool for the nature of the difference affecting the images. An optimization strategy has nevertheless to be defined for achieving better performances. This question is the purpose of this Subsection. One can basically distinguish two different cases : supervised learning and unsupervised learning. This distinction reflects the question whether a cross-validation based learning scheme can be implemented for selecting the parameters. Therefore, the case of a supervised learning is the easiest one.

Supervised learning. We consider here problems where we want to obtain a model for predicting a scalar value or a class membership (the response) associated to an image. The model is calibrated using a training set for which both the images and their associated responses are known. It aims to be used for predicting unknown responses associated to new images. A criterion can be defined to measure the relative quality of such models. The mean square error (MSE) of prediction or the number of misclassifications estimated on an independent set of images are such criterions. A leave-one-out cross-validation scheme can be designed to select the parametrization of the BAGIDIS semi-distance amongst a set of possible values. It consists in removing one pair image/response from the training set and predicting the associated value with different parametrization. One then compares the related value with the true known response. The process is repeated for all the images of the training set and the parametrization that performs best on average according to the criterion is selected. This procedure is quite common in supervised learning and usually reaches very good performances. In

case of the BAGIDIS semi-distance for curves, for instance, it has been successfully used in Chapter 2, and a theoretical support for its use has been provided. However, in practice, implementing such a cross-validation scheme requires a long computational time, especially in the case where the more flexible definition (4.2.1) is considered. Therefore, we tested the following simplified procedure in the examples of Section 4.4. Parameters are restricted to be either 0 or 1, meaning that a term is set *active* or *inactive* in the semi-distance without considering its fine tuning. The decision for the activity status of a term is taken by testing for its correlation with the response (in case of a regression problem) or by testing for a difference in its mean value (in case of a discrimination problem).

Unsupervised learning. We term unsupervised learning the problems where no known response is associated with some of the images, so that this information can not be used to select the parametrization. We successfully use the following automatic procedure in the third example of Section 4.4. We consider the elements $x_{1k}, y_{1k}, x_{2k}, y_{2k}, d_k$ with $k = 0 \dots p - 1$ that make the signature of the images. If the dataset consists in n images, there are n observations for each of those elements. We have thus $5 \times p$ vectors with n observations which define variables that may or may not be included in the computation of the distance, depending on the parametrization. We use vector quantization (the K-MEANS algorithm (Hartigan and Wong 1979) for instance) to define groups of variables having similar behavior. The suitable number of groups is chosen in an heuristic way using an elbow criterion. In each group one variable is randomly chosen as a representative and is set *active* (i.e. it is associated with a parameter 1), while the other ones are set *non-active* (i.e. it is associated with a parameter 0) in the semi-distance. The hope is that variables with a significant discriminative role will have a sufficiently different behavior to be a group by themselves, and will be grouped together if they are redundant. In the same way, it is hoped that the variables which do not support discriminative information will be sufficiently grouped so that they do not have too much weight in the semi-distance. This procedure has the advantage to drastically reduce the dimensionality of the semi-distance. It might be tested several times with different random choices of the representatives in each group, so as to estimate the stability of the result. Moreover, if some prior information can be defined on which terms are expected to be significant, it can be used as a probability function that tunes the random choice of the representative in a group according to this prior.

4.3 Review of the Literature

Image and network processing are huge fields of research, and the related literature is diverse and abundant. This Section aims at situating our work amongst the variety of available methods.

Multiscale representations of images. In the field of image processing, techniques involving multiscale representations appear powerful. Wavelets are a natural tool for achieving such representations, and a survey of their use in image processing can be found in Cohen (1997) or Mallat (1999) for instance. The usual discrete wavelet transform of an image consists basically in applying a one-dimensional wavelet transform separately along the vertical and horizontal directions of an image, which lacks the adaptivity to capture shapes transverse to those “preferred directions”. Some multiscale techniques have been designed to better capture those structures. For instance, *curvelets* (Candès and Donoho 2001) are localized oscillatory needle-shaped elements with many directions and positions. In this case, the “building blocks” of the transform do not adapt to the geometry of the images but they better capture the edges that structure the image.

Methods that combine multiscale ideas and adaptivity for image decomposition. Methods that automatically adapt to the geometry of the image often prove more efficient at providing sparse representations of geometrical structures. A recent review of adaptive image representations can be found in Peyré (2011). Amongst many more, it includes a survey of methods that share with ours the idea to combine the multiscale character of wavelet representations and the need for image adaptivity. The key idea behind those techniques is to exploit the geometric regularity along edges in an image, which is an anisotropic regularity that traditional wavelets can not deal with, by modifying the shape of the “building blocks” of the transform. Ideas for doing this have been diverse and have lead to a large panel of “-lets” transforms. One may mention the following.

- *Bandlets* (Le Pennec and Mallat 2005) are orthogonal elements that are elongated in the direction where the image has a maximum of regularity. Thereby, they are good at capturing “flows” in images. They have been used to decompose images, but they have also been applied to wavelet coefficients. A review of related techniques is found in Mallat and Peyré (2008).

- *Wedgelets* have been introduced by Donoho (1999) and then embedded in a multiresolution transform by Claypoole and Baraniuk (2000). A wedgelet is a small square image which is *up-and-down* valued on both sides of a linear edge of which the orientation has to be selected among a discrete set of possibilities. An image is dyadically decomposed and a wedgelet is adjusted to each subimage, with the edge oriented so as to best fit the subimage.
- *Tetrolets* have been defined by Krommweh (2010) as Haar-like wavelets whose supports are groups of four equal-sized neighbor squares in the image. An image is dyadically decomposed and a local tetrolet basis is selected in each dyadic square.
- *The Easy Path Wavelet Transform* (Plonka 2009) looks for paths of pixels in the image of which neighbor pixels are highly correlated, and applies a one dimensional discrete wavelet transform along each path.

Those transform rely on various ideas mostly different from the G-BUHWWT.

Another transform which expands an image in an adaptive orthonormal basis and encodes it through the combination of intensity differences and some geometric information about pixels neighborhood. Amongst the above-mentioned "-lets" techniques, *orthogonal grouplet bases* (Mallat 2009) seem the most similar in spirit to our G-BUHWWT transform. Those bases are obtained by applying the dyadic Haar transform together with an association field that links zones that are not necessarily "spatial" neighbors in the image. In some sense, new binary neighborhood relationships are defined between subsets of pixels in the image, at each dyadic scale, and a weighted Haar transform is then performed. In the image geometry, grouplets define thus a family of Haar-like bases of which the basis elements have a support that consists of possibly spatially discontinuous sets of pixels and which might become unbalanced in case several binary relationships involve a given set of pixels. On the opposite, unbalanced Haar bases keep the requirement for the support of the basis elements to be formed by contiguous pixels. In practice, this is however often the case for grouplets too, as the association field seems generally designed to reflect the geometry of a "flow" present in the image. In any cases, the number of possible grouplet bases associated to an image is the number of possible different association fields between sets of pixels. It is suggested in Mallat (2009) that one may chose the best grouplet basis to decompose a particular image simultaneously with the grouplet decomposition, thereby using matching pursuit ideas. A parallel can be drawn with the G-BUHWWT transform, which attempts to select the unbalanced Haar basis that is best suited to a particular image. As a consequence of these adaptive building processes, the best

grouplet basis transform and the G-BUHWHT transform lead to representations of an image which include both detail coefficients encoding differences of intensities and some geometric information about neighborhood relationships in the image. However, the notion of hierarchy that is embedded within the G-BUHWHT representation of the image is not at all present in the grouplet transform and the parallelism between the ideas of both transforms ends here.

Methods for estimating scalar responses associated to each of the pixels of an image. The various above-mentioned image-adaptive "-lets" transforms have been widely used in order to perform image denoising, inpainting, compression and texture synthesis (Peyré 2011; Mallat and Peyré 2008). Other approaches to such problems include, amongst numerous others, adaptive weight smoothing (Polzehl and Spokoiny 2000), pointwise adaptive estimation (Polzehl and Spokoiny 2003) and penalized regression on a graph (Kovac and Smith 2011). A recent review on image denoising techniques can be found in Milanfar (2011).

Methods for predicting a scalar value or a group membership associated to an image. Regarding the prediction of an external response from an image, as we are interested in, one may roughly highlight three families of methods, mostly originating from image classification and pattern recognition literature. First, there are methods which compare images through *landmarks*. Landmarks are specific features of the images that are identified on each image across a dataset and of which the locations are recorded and compared. Those methods are highly purpose-designed according to the particular problem at hand. For instance, face recognition problems involve facial features (eyes, nose, mouth) as landmarks (Beumer et al. 2005). We do not compare with this class of methods. Second, there are methods which compare images through suitably chosen vectors of *features*. Various sets of features have been proposed, including color histograms, measures of image coarseness, contrast, etc (Hu et al. 2008). A more elaborated approach consists in subsetting the images by using a set of "patches", and to compute vectors of features for each of the patches (Frome et al. 2006). A distance measure is then chosen for comparing these vectors of features and the resulting matrix of distances is finally processed. A key point regarding this class of method is that a dimensionality reduction is performed *a priori* on the image and the computation of a suitably chosen distance only follows. This is a main difference with our approach that relates to the third class of methods. Methods of this third class compute distances between the images directly, without relying on a preliminary dimension reduction. One may think about the Euclidean distance (L2 distance), the Manhattan distance (L1 distance), the Hausdorff distance (Huttenlocher et al. 1993), the IMED

distance (Wang et al. 2005), the Euclidean distance between a given set of principal components (Ferraty and Vieu 2006), etc. The key element is then that the distance itself should extract relevant information of reduced dimension about the images. In certain cases, a dimension reduction embedded in the distance might be tuned to adapt to the problem, which may appear more efficient. This is the case with the BAGIDIS semi-distance, but also with the Euclidean distance between a number of principal components of the image (Ferraty and Vieu 2006), for instance, the number of component having to be selected according to the particular problem at hand.

Distances that are robust to small spatial perturbations of the images. As discussed in Chapter 1 and Chapter 2 in the framework of curve comparison, computing a distance between principal components is a powerful technique as long as the significant patterns to be compared across the dataset are perfectly aligned. This is because principal components are defined using the usual Euclidean distance, which compares pixels at fixed given locations. This limitation was the initial motivation for defining the BAGIDIS methodology (Chapter 1). Another interesting proposal for dealing with this problem of robustness with respect to small spatial perturbation of the image is the IMED distance defined by Wang et al. (2005). The basic idea of this method is to weight the Euclidean distance with a matrix of which the entries are a continuous monotonically-decreasing metric function of the distance between the compared pairs of pixels - usually a Gaussian function is proposed. Although the IMED distance is adaptive as a suitable metric function is required, the IMED distance cannot really be tuned to extract specific patterns in the image. This is a difference with BAGIDIS. Another distance between images which is quite robust to small spatial variations is the well-known Hausdorff distance (Huttenlocher et al. 1993). As far as we know, the problem of misalignment of patterns in images is often tackled by registering the images (Zitová and Flusser 2003) before computing a distance between them. However, this implies that the signal can not be found in the spatial modification of the pattern.

Methods that study images with graph structures. In order to define a distance between images, this work starts with the idea to use a graph structure for encoding the neighborhood relationships between the pixels. This idea is not new. Hammond et al. (2009) and Kovac and Smith (2011) are examples.

Methods linking graph structures and wavelets. Our proposal is to study graphs using wavelets. This association can be found in several other works, of which a rapid

discussion is provided below. As far as we know, those techniques have not lead to the definition of distance measures. Most of them approach the problem very differently from us.

- *Wavelets on graphs via spectral graph theory.* Hammond et al. (2009) on the one hand and Antoine et al. (2010) on the other hand define wavelets on graphs by researching eigenvalues of the *graph Laplacian*. The latter takes the form of a matrix encoding the connectivity of each node and edge.
- *Wavelets on graphs via a diffusion operator.* Coifman and Maggioni (2006) use the powers of a diffusion operator as the scaling tool leading to multiscale analysis. Several variant of their ideas (Szlam et al. 2005; Maggioni et al. 2005) lead to different wavelet constructions.
- *Wavelets on graphs using the n -hop distance.* Crovella and Kolaczyk (2003) define for each node a series of wavelets at different scales which are radially symmetric *up-and-down* functions centered on the node, their radius being defined by the n -hop distance, i.e. the minimal number of edges one has to travel to go from the central node to another. Those wavelets encode differences between the average value of a function in a “disk” of nodes closer to the center than n -hops and its average value in a “ring” of nodes at a distance $(n + 1)$ -hop around the same node.
- *Wavelets on graphs using the lifting scheme.* Jansen et al. (2009) uses the lifting algorithm of Sweldens (1996) to build dyadic wavelets on graphs using a bottom-up approach.
- *Wavelets on dendrograms.* Some authors also have defined wavelet transforms specifically dedicated to a particular kind of graph: the dendrogram. Murtagh (2007) uses Haar bases, while Gavish et al. (2010) generalize to unbalanced Haar.
- *Adaptive discontinuous unbalanced Haar wavelet bases.* Singh et al. (2010) consider datasets with several measures associated to each node of a graph. They compute the correlations between all of the nodes and use an algorithm very similar to ours to iteratively reduce the graph by replacing two (groups of) pixels by a single one. The way they perform this reduction consists in regrouping the (groups of) pixels that are most correlated, whatever their location in the graph. This is a major difference with the G-BUHW as the graph structure is not used in the reduction process. On the contrary, the algorithm allows to highlight disjoint groups of nodes that have similar behavior across the several observations. This approach meets different goals from ours. It is closely related to the ideas behind *treelets* (Lee et al. 2008), which concern unordered

datasets.

One may summarize this Section by saying that our work combines three very powerful tools of statistical analysis of images: the idea of adaptivity to the images being investigated, the multiscale power of a wavelet approach and its efficiency to achieve sparse representations, and the capacity of graphs to account for topological relationships. It defines a new approach for wavelet transforms on graphs and uses it to define a distance measure. This combination is quite original, as far as we know. The distance is designed to be robust to spatial perturbations of the image, which is a common need that is often solved by a preliminary registration of the images. Few methods allow to compare spatial and intensity differences of images in a single-step unified framework. Ours is one of them.

4.4 Applications

Real and simulated examples are proposed in this Section. The segmentation of a noisy image is investigated in Subsection 4.4.1. A simple simulated example in the context of a functional nonparametric regression model is then proposed in Subsection 4.4.2. It aims to get an insight into the behavior of the BAGIDIS semi-distance. In particular attention is paid to its sensitivity with respect to noise and to its comparison with competing (semi-) distances. The third example, in Subsection 4.4.3 involves a dataset of handwritten zeros and ones that are to be automatically discriminated.

4.4.1 The Multizone Noisy Image

This first Subsection illustrates the ability of the G-BUHWWT to segment a noisy multizone image. We consider the gray level image of dimension 15×15 in Figure 4.8. This image counts 6 different zones of which the intensities vary between 0 and 50. We add to this image a Gaussian noise of which the standard deviation is successively fixed at $\sigma = 1$, $\sigma = 5$ and $\sigma = 8$. For each of those noisy images, a G-BUHWWT transform is performed. Because the underlying signal counts 6 zones, we know that - if there were no noise - only the 6 first basis matrices would encode the signal. Figures 4.9, 4.10 and 4.11 present the three noisy images, with successive levels of noise, and the associated five basis matrices $\Psi_1, \Psi_2 \dots \Psi_5$ returned by the G-BUHWWT (the basis Ψ_0 is always constant, by construction). We want to see if the significant patterns are extracted correctly, as a function of the noise level. A perfect extraction is achieved for $\sigma = 1$ and $\sigma = 5$. We note that the higher valued pattern (the “triangle”

with an hole at the bottom right) has intensity 50, so that $\sigma = 5$ is 10% of it, while the smaller pattern (on the bottom left) has intensity 10, so that $\sigma = 5$ is 50% of that one. When $\sigma = 8$, the algorithm still detects the significant zones, although the shapes of the two last ones (the cross in the middle and the bottom left zones) are not exactly reproduced. The rough identification of these zones is however already quite satisfying as even their visual identification is difficult in the noisy image.

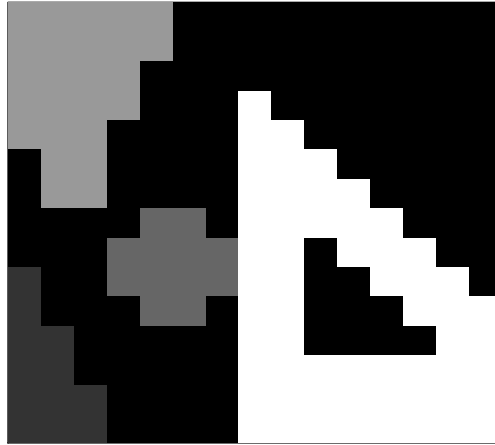


Figure 4.8: **Multizone image without noise.** The intensities of the pixels are recorded on a black-to-white scale from 0 to 50.

4.4.2 The Moving Pixel

This Subsection is intended to illustrate the ability of BAGIDIS to capture the relevant variations of a signal regarding the location or the intensity of a simple pattern. This ability is studied as a function of the noise affecting the images. The performances of BAGIDIS are assessed against those of several competing (semi-) distances: the Euclidean L_2 distance, the PCA-based semi-distance with various number of principal components (Ferraty and Vieu 2006), the IMED distance (Wang et al. 2005) and the Hausdorff distance (Huttenlocher et al. 1993).

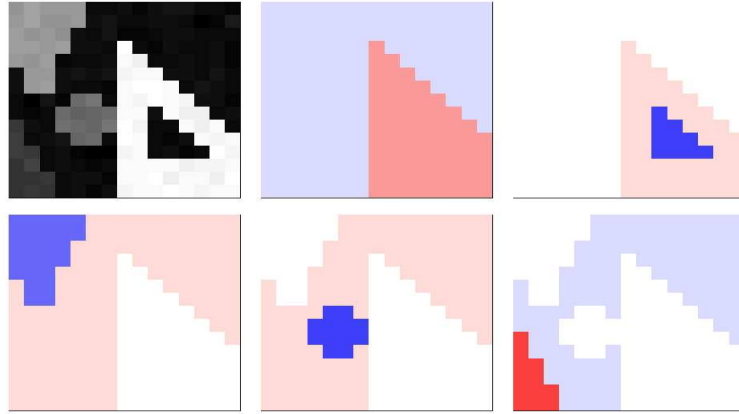


Figure 4.9: From top left to bottom right, by row: **Multizone image with noise $N(0, 1)$ and its basis matrices Ψ_1 to Ψ_5 extracting its most important features.** The intensities of the pixels of the noisy image are recorded on a black-to-white scale. The intensities of the pixels of the basis matrices are recorded on a blue to red scale with white representing the value 0.

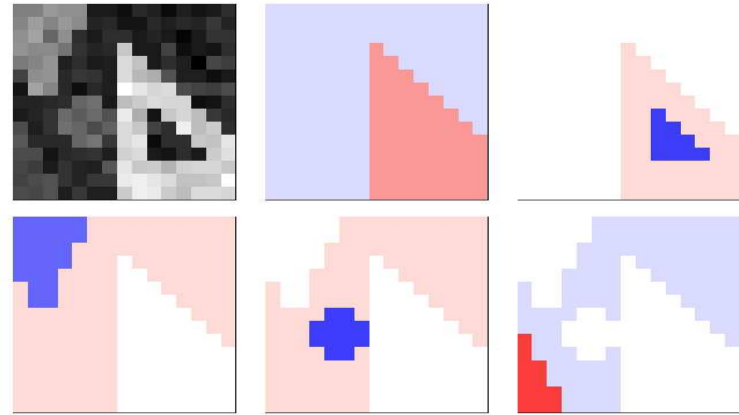


Figure 4.10: From top left to bottom right, by row: **Multizone image with noise $N(0, 5)$ and its basis matrices Ψ_1 to Ψ_5 extracting its most important features.** The intensities of the pixels of the noisy image are recorded on a black-to-white scale. The intensities of the pixels of the basis matrices are recorded on a blue to red scale with white representing the value 0.

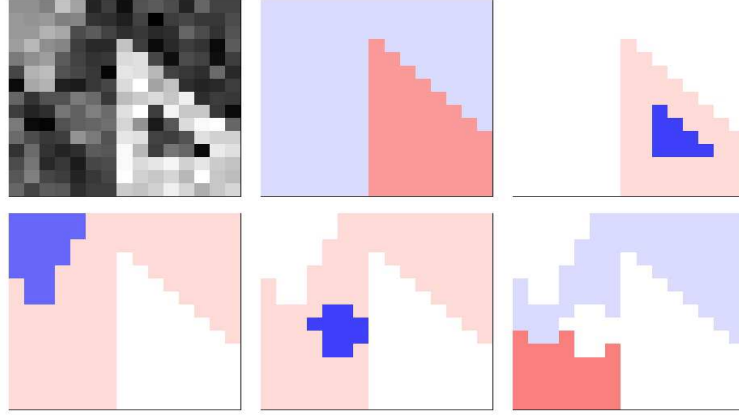


Figure 4.11: From top left to bottom right, by row: **Multizone image with noise $N(0, 8)$ and its basis matrices Ψ_1 to Ψ_5 extracting its most important features.** The intensities of the pixels of the noisy image are recorded on a black-to-white scale. The intensities of the pixels of the basis matrices are recorded on a blue to red scale with white representing the value 0.

The test scenarios. Three test scenarios are proposed. They involve 5×5 images of which all pixels are valued zero except at one randomly selected location (the “moving pixel”). The intensity of the moving pixel is defined by the test scenarios. In each case, a dataset of 90 images is generated and a Gaussian noise with standard deviation σ is added to all of them. Each image is related to a scalar value Y , determined according to the scenario. The goal is to obtain a model for predicting Y given an image. The scenarios are as follows.

- **Test 1: Capturing the location of a moving pixel.** In this scenario, the intensity of the moving pixel is fixed at 10. The predicted value Y is the row index of the moving pixel.
- **Test 2: Capturing the location of a moving pixel with varying intensity.** In this scenario, the intensity of the moving pixel is randomly chosen in the set $(5, 10, 15, 20, 25)$, with a uniform probability. The predicted value Y is the row index of the moving pixel.
- **Test 3: Capturing both the location and the intensity of a moving pixel.** In this scenario, the intensity of the moving pixel is randomly chosen in the set $(5, 10, 15, 20, 25)$, with a uniform probability. The predicted value Y is the sum of the intensity and the row index of the moving pixel.

The goal is to predict Y from the image.

The test procedure. We define a test procedure as follows. Databases of 90 images are generated according to both scenarios, for different levels of the noise. In the first scenario, the maximum tested value of the noise is 40% of the intensity of the moving pixel in the signal. In the second and third scenarios, the maximum tested value of the noise is 40% of the maximum possible intensity of the moving pixel in the signal and 200% of its minimum possible intensity. For each database, a nonparametric functional regression model (Ferraty and Vieu 2006) is trained on 70 of the images, randomly chosen in the dataset. Such a model is estimated for the BAGIDIS semi-distance on the one hand, and for the Euclidean distance on the other hand. In each case, the model is then applied to predict the Y value associated to the 20 remaining images. Those predictions are compared with the “true” known Y values, and a mean square prediction error is computed. The whole test procedure is repeated 10 times with different simulated database, for each test scenario and each value of σ .

The parametrization of the model and the (semi-)distances. The parametrization of the BAGIDIS semi-distance is selected by looking at the correlations ρ between each components of the signature (including the absolute value of the details) at each ranks and the response Y . A test of significance is applied on the correlations, with $H_0 : \rho = 0$ and $H_1 : \rho \neq 0$ with $\alpha = 0.1$ (applying a more usual threshold $\alpha = 0.05$ was too drastic in this example, as it may happen that no correlation at all is detected significant depending on the noise level). A PCA-based semi-distance, with $k = 1, 3$ or 6 components, is obtained as follows:

$$d_k^{PCA}(\mathcal{I}^{(1)}, \mathcal{I}^{(2)}) = \sqrt{\sum_{q=1}^k \left(\sum_{x=1}^N \sum_{y=1}^M (\mathcal{I}^{(1)}(x, y) - \mathcal{I}^{(2)}(x, y)) \hat{v}_q(x, y) \right)^2},$$

where $\hat{v}_q$, $q = 1..k$ is the q^{th} estimated eigenfunction of the principal component analysis, the pair (x, y) , $x = 1 \dots N$, $y = 1 \dots M$, identifies the location of a pixel, with images having dimension $N \times M$. The IMED distance is used in its nonparametrized form (eq. (6) in Wang et al. (2005)):

$$d^{IMED^2}(\mathcal{I}^{(1)}, \mathcal{I}^{(2)}) = \frac{1}{2\pi} \sum_{x, x'=1}^N \sum_{y, y'=1}^M \exp\left(-\frac{\sqrt{((x-x')^2 + (y-y')^2)}}{2}\right) \cdot (\mathcal{I}^{(1)}(x, y) - \mathcal{I}^{(2)}(x, y))(\mathcal{I}^{(1)}(x', y') - \mathcal{I}^{(2)}(x', y')),$$

where we use the same notations as above. Obviously, no parametrization is needed for the Euclidean distance and for the Hausdorff distance, the latter being computed

as

$$d^{\text{HAUSDORFF}}(\mathcal{I}^{(1)}, \mathcal{I}^{(2)}) = \max\left(\max_{p \in \mathcal{I}^{(1)}} \min_{q \in \mathcal{I}^{(2)}} |p - q|, \max_{q \in \mathcal{I}^{(2)}} \min_{p \in \mathcal{I}^{(1)}} |p - q|\right),$$

where p, q are the intensities of the pixels in the images. The bandwidth of the regression estimator is chosen by cross-validation in each case.

The results. Results are provided in Figure 4.12 as a function of the noise standard deviation σ , for each test scenario. We see that using BAGIDIS improves the prediction capacity as compared with most competitors, in each scenario. Here is a discussion.

- The Euclidean distance and the PCA-based semi-distance show bad performances. This is not surprising as they totally miss the geometry of the pixel network, which gives crucial information here. Therefore, the only information they can rely on for prediction is the one extracted from images that are strictly aligned with the one for which a prediction is desired. This makes the prediction highly difficult, a fortiori in the case where the intensity of the moving pixel also changes. Better performances would require a large and highly redundant dataset. Any semi-distance relying on pixel-to-pixel comparison would produce such bad results, because of its construction.
- The IMED distance is quite competitive for **Test 1**, at small noise levels. This is because this distance accounts for similarities of pixels that are misaligned, as BAGIDIS does. However, the IMED distance fails as soon as the intensity of the moving pixel varies (in **Test 2** and **Test 3**). The reason is that the information about intensity and location can not be decoupled by IMED, what BAGIDIS conversely does.
- The Hausdorff distance is the most successful in **Test 3**, while BAGIDIS has equivalent or slightly worse performances, depending on the noise level. However, it is very bad in **Test 1** and **Test 2**. Indeed, the Hausdorff distance is not able to capture the part of the response which concerns the location, because two images with a moving pixel of the same intensity but at different locations have a dissimilarity zero according to the Hausdorff distance. Its good performances in **Test 3** are related to the fact that the part of the response that concerns the location (valued in $\{1, 2, 3, 4, 5\}$) is mostly dominated by the part of the response related to intensity (valued > 5).

Those results illustrate both the good performances of BAGIDIS and its flexibility to capture the relevant information in the image. The first test scenario (**Test 1**) shows that BAGIDIS very efficiently captures a signal contained in the location of a pattern

up to a noise level as high as 20 to 30% of the signal intensity. The second scenario (**Test 2**) shows that BAGIDIS can achieve about the same good performance even if the intensity of the moving pixel varies, which indicate that the distance efficiently adapts to extract the relevant information, what none of the competing methods can do. Moreover, the third scenario (**Test 3**) indicates that BAGIDIS can as well capture the intensity, in addition to the location when required for the given problem. Its performance in the latter case, is comparable to the one of its best competitor at small noise levels, and slightly inferior at medium noise level. It remains nevertheless clearly better than other competitors in this case.

4.4.3 Handwritten Digits

This third Subsection concerns images of handwritten zeros and ones. Two different kinds of studies are performed. First, we visually analyze the dataset in a blind unsupervised way, using multidimensional scaling as a diagnostic tool. Second, we adjust a nonparametric functional discriminant model (Ferraty and Vieu 2006) for automatically recognizing the handwritten digits.

The database. The data consists in images of handwritten digits from the publicly available MNIST database (<http://yann.lecun.com/exdb/mnist/>). This database contains 70000 black-and-white images of digits, with the images centered in a 28x28 image by computing the center of mass of the pixels, and translated so as to position this point at the center of the 28x28 field. The database is loaded into R using a purpose-designed function by B. o'Connor provided on the following website: <https://gist.github.com/39760>. We restrict ourselves to a small subset of images of handwritten zeros and ones within this database.

Unsupervised learning. The first test we perform with this handwritten digits database is as follows. We consider 50 images of zeros and 50 images of ones. We compute the set of pairwise BAGIDIS distances between all images, with a parametrization chosen according to a K-MEANS algorithm as discussed in Subsection 4.2.3. In this example, $K = 10$ groups of variables are identified by the elbow criterion associated to the K-MEANS procedure. In these groups, one representative is then randomly selected in each group, and the associated parameters of the BAGIDIS semi-distance are set to 1, the others being set to 0. This procedure drastically reduces the dimensionality of the distance. Given the so-obtained set of pairwise BAGIDIS semi-distances, a multidimensional scaling representation is provided, in which the projected points

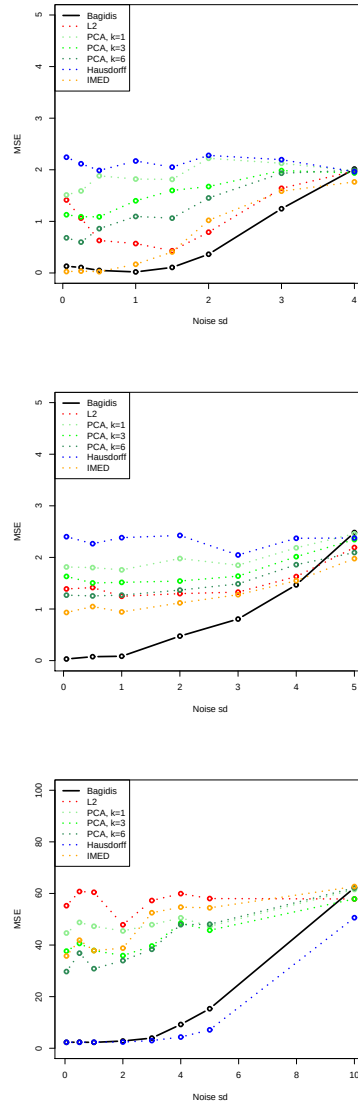


Figure 4.12: **The moving pixel.** Those figures present the average of the mean square errors (MSE) over 10 repetitions of **Test 1**, **Test 2** and **Test 3**, as a function of the noise level, for the BAGIDIS semi-distance and competing (semi-) distances. Results for the three test scenarios **Test 1**, **Test 2** and **Test 3** are presented successively, from top to bottom.

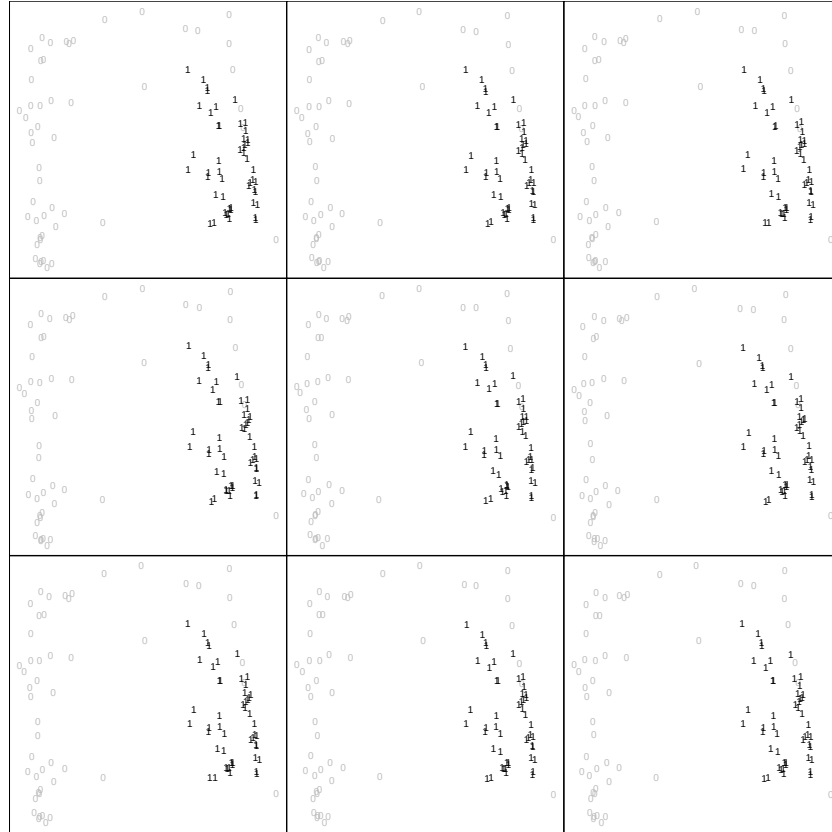


Figure 4.13: **Unsupervised Learning for the handwritten digits recognition problems.** This figure presents 9 multidimensional scaling representations of a dataset of 100 images, with different random choices of the activated coefficients in the groups determined by the K-MEANS algorithm.

are labeled 0 or 1 according to the truly represented digit. This helps visually assessing the performances of the blindly adjusted distance. The whole test procedure was applied to two different datasets and repeated 9 times for each of them with different random choices of the parametrization. Figure 4.13 provides the results obtained for the first dataset. Results for the other dataset looks pretty similar and are therefore not shown here. Although the achieved discrimination is not perfect, the results look stable whatever the selected variables in the K-MEANS groups. The similarity of all images of ones is clearly captured. Most images of zeros also form a well-identified group. Some images of zeros appear misclassified however. This might partially be an effect of the projection in the MDS representation. Those results seem therefore

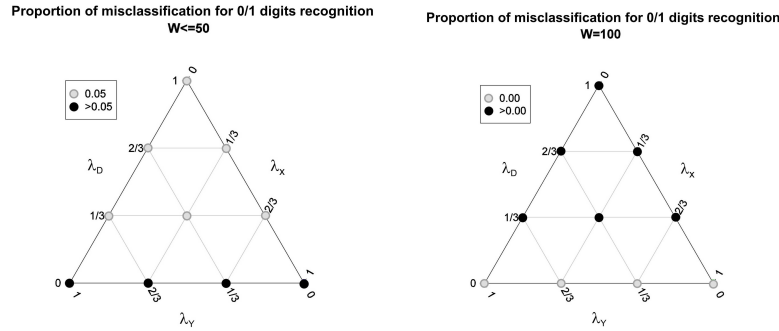


Figure 4.14: **Handwritten digits recognition: Investigation of the effects of the parameters and restricted cross-validation procedure.** The ternary plot in the left panel presents the results we obtain as a function of the parameters $(\lambda_x, \lambda_y, \lambda_D)$ when W is taken in the set $\{2; 5; 10; 20; 50\}$. The one in the right panel indicates the results obtained with $W = 100$.

quite encouraging for an unsupervised analysis with an unsupervised choice of the parametrization.

Supervised learning. The second analysis with the handwritten digits database consists in the calibration of a model that discriminates images of zeros and ones. A nonparametric functional discrimination model (Ferraty and Vieu 2006), with a k-NN algorithm, is trained on a dataset of 70 images of one and 70 images of zero. It is then tested on an independent set of 15 ones and 15 zeros. The number of neighbor k is selected using a local cross-validation (function *funopadi.knn.lcv* from the companion website of Ferraty and Vieu (2006), slightly adapted for being used with BAGIDIS.) Several parametrization are tested. The restricted form (4.2.2) is first considered for a preliminary, limited, investigation of the parametrization by cross-validation. The cross-validation is performed amongst values $(\lambda_x, \lambda_y, \lambda_D)$ taken in the set $\{0; \frac{1}{3}; \frac{2}{3}; 1\}$ under the constraint that they sum to 1. All of the possible combinations of these values are tested with several weight functions w_k , that are step functions with values 1 up to a rank $k = W$ and 0 afterwards. Tests are performed with $W = 2, 5, 10, 20, 50$ and 100. This complete experimental design is repeated twice with different training sets and test sets. The proportion of misclassification on the test set is recorded for each configuration. Results are presented in Figure 4.14. Whatever $W \leq 50$, good performances (misclassification rate: 0.05) are achieved as soon as the coefficient λ_D associated to the details is non zero, while the relative proportions of λ_x and λ_y do not matter. On the opposite, when $W = 100$, the best results occur when $\lambda_D = 0$, the relative proportions of λ_x and λ_y remaining non-influent. In this case, a perfect discrimination is achieved. The observations above mean that the first ranks of the signature

clearly carry some discriminative information in the detail component. It is however required to look further in the hierarchy, between ranks 50 and 100, to find the information needed to achieve a perfect discrimination. This information lies in the components (x_1, y_1) and (x_2, y_2) of the signature, indicating that some information about the geometry of the image is needed. Given this, a possible cross-validated choice for the parametrization might be $(W = 100, \lambda_D = 0, \lambda_x = 0.5, \lambda_y = 0.5)$. Such a cross-validation procedure, although incomplete, is rather time-requiring. We test then the following “rapid” strategy. We consider the formulation (4.2.1) of the semi-distance. Each component of the signatures of the images in the training set is studied separately, rank by rank. A t-test is performed for the difference of the mean of the values taken by this component in each group (zero or one) of images. The p-values are recorded. Only components having a p-values smaller than $1E - 10$ are given a coefficient 1 in the parametrization, the other ones being set to 0. The so-trained model is use for discriminating the images in the test set and the number of misclassification is recorded. The procedure is repeated 8 times with different training sets and test sets. We successively obtain misclassification rates 0, 0, 0, 0.03, 0.03, 0, 0.1 and 0. On average, a misclassification rate 0.02 is thus achieved.

Discussion. This study of handwritten digits recognition shows really encouraging performances of BAGIDIS on real data. We have certainly to mention that numerous classification methods have been tested on the NMIST database, with very good results. A point that has nevertheless to be highlighted is that the results we present here have been obtained with a model trained on a relatively small subset of images, with a semi-distance that has a very general range of application, which might be desirable in practical problems. No prior information about the images or about the shapes of the discriminant patterns is used in the definition of the model.

4.5 Conclusion

Our original contribution is twofold in this Chapter.

First, we describe the G-BUHW transform of an image. This algorithm combines the functional requirement to account for topological relationships between the pixels of an image, the multiscale properties of a wavelet approach and the useful flexibility of a data-driven method. It results in a sparse description of an image, through its expansion in a well-suited basis. This description is hierarchic in the sense that it provides a ranking of the features the image is made of. Therefore the G-BUHW may be used to segment the image into different pairs of zones, ranked according to

the importance of their difference.

Second, we define the BAGIDIS semi-distance between images. BAGIDIS makes use of the NET-ID of the images to compare them through both the amplitudes and locations of the patterns they encode, in a hierarchic way. This makes the distance highly powerful in problems where the patterns in the images might be misaligned. This is illustrated with real and simulated problems. Performances of the method in regression, classification and clustering problems are illustrated. Results are shown really encouraging. It is worth to emphasize that the training sets used in this experiment were relatively small which indicates a certain power of the method. Another advantage of BAGIDIS is its flexibility and its possibility to be used as a diagnostic tool for the significant differences in the dataset. The semi-distance can be parametrized so as to extract from the images some information of reduced dimension that is relevant for the particular problem investigated.

A clear asset of the method is its very general scope. It is not designed for particular kinds of images and might be used as is for images with different underlying geometries, or more generally for any problem that can be described as a network with a spatial anchorage. The only requirement for the image is to have a relevant underlying structure that can be described in a hierarchic way, at least locally.

*

Acknowledgments. *The authors thank Rainer von Sachs for his encouragements and comments about this project. Part of this work was completed during a stay of Catherine Timmermans at London School of Economics that was funded by the Institute of Statistics, Biostatistics and Actuarial sciences of Université catholique de Louvain, and was made possible thanks to the hospitality of Piotr Fryzlewicz. Part of this work was completed during a stay of Piotr Fryzlewicz at Institute of Statistics, Biostatistics and Actuarial sciences of Université catholique de Louvain. Statistical analyses are performed using the R software (R Development Core Team 2008), and using the R implementations of the functional regression estimators provided on the companion website of Ferraty and Vieu (2006), slightly adapted for their use together with BAGIDIS. Financial support from the IAP research network grant P 06/03 of the Belgian government (Belgian Science Policy) is gratefully acknowledged.*

*

Previous versions of this work include:

Timmermans C. and Fryzlewicz P., 2012, *SHAH: SHape-Adaptive Haar wavelet transform for images with application to classification*. ISBA Discussion Paper 2012/15.

Conclusion

This conclusion is organized as follows. I first give a brief summary of the thesis. I then discuss its original methodological contributions, and highlight the perspectives they open for future research. I finally get back to the original motivation of spectroscopic analyses.

*

The thesis proposes a methodology for investigating functional data with sharp local patterns. The key element of the method is a new way to measure a semi-distance between curves (Chapter 1) or images (Chapter 4). This semi-distance takes into account the functional nature of the data, in an adaptive and highly flexible way.

The thesis starts within the framework of curve comparison. The core idea (Chapter 1) consists in the expansion of each curve or image of a dataset in a *different*, curve-adapted, wavelet basis. This basis is automatically selected within the family of Unbalanced Haar Wavelet Bases (Girardi and Sweldens 1997; Fryzlewicz 2007) so as to best encode the sharp features of the data in a hierarchic way. This is a major originality. The BAGIDIS semi-distance is defined therefrom by comparing not only the projections of the curves onto their associated bases, but also the bases themselves. This specificity underlies the name of the method: BAGIDIS stands for *BAses GIving DIStances*.

The use of the BAGIDIS semi-distance in the framework of nonparametric functional data analysis is shown really promising (Chapter 2). Under some general conditions, functional variables are shown fractal with respect to the BAGIDIS semi-distance, which implies that we know the rate of pointwise convergence of the functional nonparametric regression estimator associated to the BAGIDIS semi-distance. Competitive

rates are achieved provided that the parametrization of the BAGIDIS semi-distance is sparse enough.

A spectroscopic case study assesses the advantages of the BAGIDIS methodology for applied research purposes, in the field of a metabolomics ^1H NMR spectra analysis about Age-related Macular Degeneration (Chapter 3). Visual data analysis, disease detection modeling and search for disease biomarkers are proposed. The results are compared with a recent investigation of the same dataset (Rousseau 2011) with state-of-the-art methodologies. BAGIDIS is shown a useful complement to the latter due to its important visual discrimination capacities, its very good performances at detecting the disease and its large scope of biomarker identification.

Finally, the interest of the thesis enlarges towards images (Chapter 4). Defining an extension of BAGIDIS in this case requires defining a suitable hierarchic basis function expansion. The G-BUHWHT transform (Chapter 4) provides with this. Its main idea is to see an image as an *Intensity Network* - an object combining the set of the intensities of the pixels, a graph structure encoding the neighborhood relationships between the pixels, and an explicit anchorage of the pixels in a Cartesian space. The basis function expansion is then obtained through an iterative data-driven reduction of the graph structure. The extension of BAGIDIS to images directly follows.

Thanks to its unusual building process, the BAGIDIS semi-distance takes into account the variations of the sharp local patterns in data regarding both their intensities and their locations, in a unified framework. This property is a key asset of the methodology, as classical methods usually fail to compare curves (or images) as soon as the location of a sharp local pattern slightly differs from one curve (or image) to another (Chapter 1, Chapter 2), whatever the noise or signal origin of the variation. Its flexible parametrization allows BAGIDIS to adapt itself to the kind of variations affecting the significant features of the data (variations in location, intensity, or both simultaneously). Moreover, it may act as a diagnostic tool for the nature of this variation. Because of these qualities, BAGIDIS reveals powerful in applied problems of spectroscopic data analysis (Chapter 1, Chapter 3). Promising results on image datasets also give hope for its use in image analysis research problems (Chapter 4).

An efficient measure of dissimilarity is a fundamental component for numerous statistical analyses, aimed at data mining and visualization as well as discrimination, classification or prediction modeling. BAGIDIS has been shown useful in association with multidimensional scaling representation (Chapter 1, Chapter 3), hierarchic clustering (Chapter 1), k -nearest neighbor algorithms (Chapter 1, Chapter 4) and nonparametric

functional regression models (Chapter 2, Chapter 3 and Chapter 4). Regarding the latter model, a rate of convergence for the regression estimator coupled with BAGIDIS can be provided under some general conditions (Chapter 2). A really competitive rate of convergence can be achieved provided that the parametrization of BAGIDIS is sparse enough. A related point, whatever the jointly used algorithm, is the optimal choice of this parametrization, in each given problem. In case of a supervised problem, a cross-validated choice is a classical and generally pertinent solution. An important side-result of the thesis is to prove the asymptotic optimality of the cross-validated choice of a multidimensional parameter defining a semi-distance used within non-parametric functional regression models (Chapter 2). Another side-result of the thesis is found in the tests and tools provided for statistically analyzing datasets of curves with BAGIDIS or with another distance measure (Chapter 1).

*

The starting point of the thesis is rooted in an applied question for analyzing spectroscopic datasets: comparing curves of which the sharp local patterns might possibly be misaligned. The first contribution of the thesis is thus to provide for an original tool that answers this issue. From the point of view of statistical research, I feel that the contribution of the thesis is however more diverse. Therefore, I list below the original proposals and results that come along with this thesis. I think those ideas might themselves open doors for new developments. I discuss some perspectives in relation with the several mentioned contributions. Moreover, although shown really efficient at analyzing several simulated and real datasets, the BAGIDIS methodology is not believed to be a panacea. Limitations exist and some aspects of the thesis involve heuristic arguments. These are a call for further developments, too. I indicate below tentative approaches to overcome them. Actually, I think one the main achievement of the thesis precisely lays in these openings for future work.

Here is the discussion.

The expansion of each curve in a basis that is best suited to it as a preliminary step for curve comparison. This is certainly the most unusual proposal of the thesis. The idea of individual "data-adaptivity" of the basis selection for curve comparison distinguishes from the well-known powerful idea of "dataset-adaptivity" of which principal components analysis is a leading tool (Jolliffe 2002). BAGIDIS shares with principal component analysis a projection-based character with an aim to hierarchize the differences between the curves according to their importance. The thesis shows the potential of comparing curves

through their expansion in bases that are different from one curve to another but individually more efficient.

The individually "best suited bases" used here are restricted to the family of unbalanced Haar wavelet bases. Perspectives include the generalization to other kinds of bases families. A necessary condition is the possibility to find another family of bases that benefits from a best suited basis selection algorithm which ranks the basis elements according to a criterion keeping them comparable from one curve to another. Another condition is the possibility of a sparse encoding of the selected basis, so that relatively low-dimensional signatures of the curves might be defined. Although this discussion mentions curves, it also concerns images, for which numerous bases dictionary do exist and might be investigated (see the review of some "-lets" transforms in Chapter 4).

The explicit identification and exploitation of the hierarchy in the BUUHW and G-BUUHW transforms. The key to success in the comparison of curves expanded in individually selected bases is the hierarchy of the induced expansions. The thesis highlights this notion of hierarchy and its distinction from a multi-scale character, both notions bringing complementary points of view.

The idea of hierarchy was in some sense already present in Fryzlewicz (2007), although not explicitly exhibited and used. I believe that the coupling of this notion with the recently developed "Time-Threshold Map" of Fryzlewicz (2012) counts amongst the possible promising uses of hierarchy for data investigation. The Time-Threshold Map plots curves expanded in a wavelet basis as a function of time with multiple thresholdings across a range of possible threshold values. Combining the information from such a Time-Threshold Map with the one of a similarly defined Time-Rank Map might certainly give further insight to the links between the sparsity that can be achieved by a truncation of the hierarchical description and the sparsity arising from usual wavelet thresholding for denoising purposes.

Hierarchy is currently understood as "an ordering of the basis elements according to the power of the supported *up-and-down* patterns in the signal, under a functional constraint". As a follow-up of the present work, it would be desirable to mathematically deepen this definition. This should lead to the definition of a class of functions that could be said hierarchical up to a given level of noise, in the sense that the ranking of their important features would be unambiguous

with a known high probability, up to that level of noise.

A related possible development is to couple the BUUHW algorithm with a testing-thresholding scheme that would provide some initially non-hierarchical functions with an unambiguous hierarchic description, with a given probability related to the noise level. The principle would be to test for the equality of candidate-details in the BUUHW algorithm and to favor the one related to the smallest horizontal index in case of the non-rejection of the equality. Thereby, the hierarchic description would become non-ambiguous, in the same way as it is non-ambiguous for non-noisy curves, with a certain probability associated to the Type I and Type II errors of the tests at each iteration. The same discussion might be extended towards images.

The development of a semi-distance that takes into account the variations of the patterns in a curve or an image both in location and in intensity, in a unified framework. The BAGIDIS semi-distance provides for a way to efficiently quantify the closeness of curves or images with significant sharp local features that might be modified in intensity as well as in alignment. Thereby, it provides for a solution to a problem that current methods do not handle in a satisfactory way. The simultaneous investigation of the intensity and location variations of the patterns is achieved through the definition of the signatures of the curves and images. Signatures define "paths" in $\mathbb{R}^3$ (curves) or $\mathbb{R}^5$ (images), the initial part of which being expected to reflect the main features of the given curve or image while the subsequent part is expected to reflect patterns of decreasing importance and ultimately noise.

The way BAGIDIS compares the signatures as a weighted sum of scaled rank-to-rank distances is actually quite simple, although already efficient. Further studies might benefit from improvements in the way to compare signatures. Projections on principal components or independent components of the signatures obtained for a given dataset might be analyzed. On top of that, the comparison might usefully be made functional.

A highly flexible method of which the parametrization within a distance-based learning scheme might act as a data-driven dimension reduction directed towards a given learning problem. Defining the parametrization of BAGIDIS is a continuously encountered question in the thesis. Parameters are first defined as a prior (Chapter 1). They are then adapted to act as a diagnostic tool (Chapter 1). A theoretical support for their optimization using a cross-validated

procedure is provided in the framework of nonparametric functional regression (Chapter 2). Some guidance for a pragmatic less computationally intensive selection is also given for supervised as well as unsupervised learning (Chapter 4). The performance of those approaches is shown sufficient on various simulated and real datasets.

The question remains nevertheless open on how to further improve this parameter selection. Two preeminent aspects are the computational efficiency of the optimization procedure and the development of a better theoretically supported scheme adapted to non-supervised learning. Those aspects would certainly benefit from links with approaches developed in the field of "metric learning" research (see Frome et al. (2006); François et al. (2011) for instance). Another idea is to define the parametrization from the quantity of "structure" in the $b - d$ plane (or in its extension for images) at each rank, as compared with the distribution of the components of a white noise at the same rank in the $b - d$ plane.

Descriptive statistics that embed the good properties of BAGIDIS for dealing with possibly misaligned data. A definition of the mean and variance of a dataset is provided in the framework of BAGIDIS for curves. Its extension for images is direct. Beside providing for tools for describing distributions of curves, those definitions have lead to the proposal of an ANOVA-like F-test in the BAGIDIS framework.

A distinct way to take advantage of this definition might be as follows. An assessment of the robustness of statistical methods with respect to noise is often performed as a function of a signal-to-noise ratio. Commonly used definitions of the signal-to-noise ratio take into account patterns and noise measured along the vertical (intensity) direction. As far as I know, there is no equivalent notion that takes into account signal and noise variations along the horizontal (location) direction. The definition of the moments of a distribution in the BAGIDIS framework is a step forward towards the definition of such an extended signal-to-noise ratio that would be useful in case of misaligned data.

Some tests for differences between groups that are suitable for possibly misaligned data. Two different tests are proposed (Chapter 1) for investigating the differences between two or more groups of curves within the BAGIDIS framework. One of them has a more general scope than BAGIDIS as it may be used with other distance measures.

Both tests might be extended to the comparison of images. In the same way as the usual ANOVA and t-test they derive from, those tests rely on quite restrictive assumptions of normality and homoskedasticity. It is believed that nonparametric extensions of the ANOVA and t-test might be transposed in the BAGIDIS framework so as to overcome those limitations.

The rate of convergence of the nonparametric regression estimator used with BAGIDIS. The rate of pointwise convergence of the nonparametric functional regression estimator is obtained under quite general conditions, as a function of the parametrization of BAGIDIS. This result concerns BAGIDIS for curves in the thesis, but its extension for images is straightforward.

A generalization of those results might go towards two different directions. First, one may investigate the rate of uniform almost complete convergence of the functional kernel regression estimator with BAGIDIS. This would require the combination of our work with the ideas of Ferraty et al. (2010). Second, it might be of interest to consider the extension of our results towards a functional regression with a functional response, thereby referring to the recent developments of Ferraty et al. (2011).

A proof of the asymptotic optimality of a cross-validated choice of a vectorial parameter specifying a semi-distance in functional regression. This result supports for the leave-one-out cross-validated choice of the parametrization of BAGIDIS in nonparametric functional regression. Moreover, the scope of this result is not restricted to the BAGIDIS framework. It also concerns the very general family of projection-based semi-distances for which one has to select a set of components upon which to project. The importance of this family of semi-distances has been stated in (Ferraty and Vieu 2006) so that the interest of our result is quite large.

Future developments of this result might follow the directions of Burba et al. (2009) and lead to its generalization for a nonparametric k -NN functional regression estimator.

The Generalized-Bottom Up Unbalanced Haar Transform for images. The definition of BAGIDIS for images has required the development of the G-BUHWHT algorithm. This result has also an interest by itself, although not exploited in the thesis.

A first direct perspective for future work would be the use of the G-BUHW in association with a wavelet thresholding, thereby generalizing the work of Fryzlewicz (2007), with a goal of piecewise-constant images denoising. Another possible direction is the investigation of the G-BUHW for simultaneous denoising and edge detection, thereby taking into account the fact that noise can affect both the intensity of the pixels but also the shapes of the patterns.

The following idea sounds to me really promising. An unbalanced wavelet principal components selection can be achieved by coupling the ideas of Singh et al. (2010) with the G-BUHW algorithm. It applies to datasets of images of which the pixels are observed at the same locations, and embedded within the same graph structure. The idea of Singh et al. (2010) is to compute "details" in the transform through correlations between pixels at the same location across the image dataset. Thereby a best basis selection is achieved that hierarchically describes the dataset of images instead of the individual image. However Singh et al. (2010) do not take into account the spatial neighborhood between the pixels. The proposal is to include this notion through a graph structure accounting for neighborhood relationships between the pixels, as is done for the G-BUHW. This functional constraint should lead to the segmentation of the images of the dataset into zones of which the edges are estimated so as to best fit the entire dataset. A direct application of this technique might be the segmentation of HDR images, which consists in an array of several photographs of the same scene with different speed-aperture couplings for the acquisition process.

Further developments of the G-BUHW should also include a localized version that would be the counterpart of the sliding window used in association to the BUHW in the case of long curves with numerous patterns. Another possibility would be a multi-stage version of the G-BUHW over a certain number of dyadic partitions of an image.

Another possible field of investigation with the G-BUHW might be related to texture synthesis and image generation. Indeed, it might be of interest to play with the G-BUHW by modifying the codebook of the NET-ID so as to make it slightly different from the codebook in the IN. Inverting the G-BUHW transform would then lead to a new reconstructed image. One might possibly find some interest in investigating such a transform.

Last, it is worth observing that the topological part of the G-BUHW transform consists in the definition of the evolution of graph. Other ways to compare images and networks may take this explicitly into account and benefit from in-

teractions with research about "evolving graphs".

*

To conclude this thesis, I would like to say a word on achievements and perspectives for spectroscopists.

Practitioners are provided with a tool for investigating differences between groups of spectra in a dataset, determining a statistical model for the prediction of a measured quantity or a group membership associated to a spectrum, and identifying the zones of the spectra that carry significant information for the discrimination. This tool can be advantageously used in association with or in addition to state-of-the-art methods for spectroscopic analysis (Rousseau 2011), as illustrated in the framework of a metabolomics study in Chapter 3. It proposes a solution to the common problem of misalignment of the spectra, which was not satisfactorily dealt with by existing methods. I think this contributes to tackle the continuous challenging need for new computer-aided research and diagnosis methods.

The enlargement of the interest of the thesis towards images meets the current trend in spectroscopic data acquisition to provide for bi-dimensional spectra. It is hoped that the methodology will prove efficient for analyzing such data, and may thereby continue to contribute to some biomedical and pharmaceutical research.

Summary of main notations

About datasets

$\mathbf{x}, \chi$: a discretized curve (a series) in $\mathbb{R}^N$. Notation χ is preferred in Chapter 2 for accordance with the literature in the field. In case a particular series has to be identified, an upper index ($\mathbf{x}^{(i)}$) is used, except in Chapter 2 where a lower index (χ_i) is preferred for accordance with the literature.

$\mathcal{I}$: an image.

$X = \{X_l\}_{l=1\dots p}$: a vector of intensity, related to an image $\mathcal{I}$.

N : the length of a series, the row dimension of a matrix (encoding an image).

M : the column dimension of a matrix (encoding an image).

p : the number of pixels in an image, the number of nodes in $\mathcal{E}^{IN}$ and $\mathcal{E}^{OUT}$.

n : the number of curves/ images in a dataset.

$\mathcal{D}^{(p)}$: a codebook.

$\mathcal{E}^{IN}, \mathcal{E}^{OUT}$: a graph.

E : the number of edges in $\mathcal{E}^{IN}$.

$\epsilon_l = (j, k)$: an edge of a graph, directed from node j to node k .

IN: an Intensity network.

NET-ID: a NETwork of Intensity Differences.

About basis expansions

$\{\mathbf{e}_j\}_{j=1\dots N}, \{\mathbf{e}_j\}_{j=1\dots p}$: the canonical basis of $\mathbb{R}^N$, or $\mathbb{R}^p$.

$\{\psi_k\}_{k=0\dots N-1}, \{\psi_k\}_{k=0\dots p-1}$: the unbalanced Haar wavelet basis that is best suited to a series in $\mathbb{R}^N$ (the BSUHWB), or to a vector of intensity in

$\mathbb{R}^p$ representing an image. In the latter case, the vectors ψ_k are preliminary constructions leading to the definition of the actual unbalanced Haar wavelet basis elements $\{\Psi_k\}_{k=0\dots p-1}$ which reflect the geometry of the intensity network (IN). A lower index is used throughout the thesis for identifying the vectors (ψ_k) except in Chapter 2 where an upper index (ψ^k) is preferred for concordance with the literature in the field.

$\{\Psi_k\}_{k=0\dots p-1}$: the unbalanced Haar wavelet basis that is best suited to an image of p pixels.

$d = \{d_k\}_{k=0\dots N-1}, d = \{d_k\}_{k=0\dots p-1}$: the detail coefficients, i.e. the coefficients of projection of a series of length N onto a basis $\{\psi_k\}_{k=0\dots N-1}$, or the coefficients of projections of an image of p pixels (or an intensity network with p nodes) onto the basis $\{\Psi_k\}_{k=0\dots p-1}$.

${}_e X$: a vector of intensity expressed in the canonical basis of $\mathbb{R}^p$.

${}_\psi X$: a vector of intensity expressed in the basis $\{\psi_k\}_{k=0\dots p-1}$.

${}_{psi}(\mathbf{1})_e$: the matrix of basis change from $\{e_j\}_{j=1\dots p}$ to $\{\psi_k\}_{k=0\dots p-1}$.

BSUHWB: the Best Suited Unbalanced Haar Wavelet Basis, obtain using the BUUHWB algorithm.

BUUHWB: the Bottom-Up Unbalanced Haar Wavelet Transform.

G-BUUHWB: the Generalized Bottom-Up Unbalanced Haar Wavelet Transform.

About BAGIDIS

BAGIDIS: BAses GIving DIStances.

$b = \{b_k\}_{k=1\dots N-1}$: the set of breakpoints associated to the basis vectors $\{\psi_k\}_{k=1\dots N-1}$.

$\{y_k\}_{k=1\dots N-1} = \{(b_k, d_k)\}_{k=1\dots N-1}$: the signature of a series of length N .

$s = \{s_k\}_{k=1\dots N-1} = \{(x_{1k}, y_{1k}, x_{2k}, y_{2k}, d_k)\}_{k=1\dots p-1}$: the signature of an image (or an intensity network) of p pixels (or p nodes).

$\lambda, (\lambda_x, \lambda_y, \lambda_D)$: the balance parameter (for curves), the vector of balance parameters in the horizontal, vertical and detail directions (for images).

$\{w_k\}_{k=1\dots N-1}, \{w_k\}_{k=1\dots p-1}$: the weight function for curves of length N , the weight function for images of p pixels (or intensity networks with p nodes).

$p = \{p_k\}_{k=0\dots p-1} = \{\alpha_k, \beta_k, \gamma_k, \delta_k, \epsilon_k\}_{k=0\dots p-1}$: a vector of parameter defining the BAGIDIS semi-distance.

- $\mathcal{K}$: a set of ranks related to non-zero weights w_k , with K the cardinality of $\mathcal{K}$.
- Δ : the length of the sliding window.
- p : the kind of norm used for computing the partial distances in the BAGIDIS semi-distance.

About the BUHWT algorithm

- $\mathbf{x}^{[N-i]}$: the approximation of x in dimension $N - i$ at iteration i .
- $\{\mathbf{f}_j\}_{j=1\dots N-i}$: a complete set of generating vectors for the vectorial subspace $V^{[N-i]}$ in which $\mathbf{x}^{[N-i]}$ is defined.
- $V^{[N-i]}$: a vectorial subspace of $\mathbb{R}^N$, of dimension $N - i$.
- $\{\omega_j\}_{j=1\dots N-i}$: a set of weights associated to the $N - i$ components of $\mathbf{x}^{[N-i]}$.
- $\Psi_{j,j+1}^p$: is the candidate projection-axis in the plane generated by $\{\mathbf{f}_j, \mathbf{f}_{j+1}\}$.
- $\Psi_{j,j+1}^d$: the associated detail-axis,;
- $L_{j,j+1}$: the information loss associated to a projection of the plane $\{\mathbf{f}_j, \mathbf{f}_{j+1}\}$ onto the axis $\Psi_{j,j+1}^p$.

About the G-BUHWT algorithm

- $\mathcal{E}^{(i)}$: the set of edges in the network at iteration i .
- $\mathcal{E}^{(i)} = \{\epsilon_l = (j, k)\}$: a network at iteration i with edges ϵ_l directed from a node j to a node k .
- $X^{(i)}$: the value of the nodes remaining in the network at iteration i .
- $\omega^{(i)} = \{\omega_l\}_{l=1\dots p-i}$: a set of weights associated to the $p - i$ nodes remaining in the network at iteration i .

About nonparametric functional regression

- Y : a scalar response.
- $(\mathbf{X}, \mathbf{Y})$: a random pair of explanatory and response variable.
- $\hat{Y}$: the estimated response.
- r : the true regression operator.
- $\hat{r}$: the estimated regression operator.
- $K(\cdot)$: an asymmetric kernel.

h : the bandwidth. A lower index n might be joined to emphasize a dependency on the number of curves in the dataset.

$d(\cdot)$: a semimetric.

β : the Lipschitz parameter.

$\phi_{d,\chi}(\cdot)$: the small ball probability to be in a ball $B_d(\chi, \cdot)$ centered on χ , defined according to the semimetric $d(\cdot)$, as a function of the radius of the ball.

$B_d(\chi, h)$: a ball of radius h , centered on χ , defined according to the semimetric $d(\cdot)$

H : a vector parameter defining a semimetric.

$\mathcal{H}$: a set of vector parameters. A lower index n might be joined to emphasize a dependency on the number of curves in the dataset.

MISE: the Mean Integrated Square Error.

CV: the leave-one-out Cross-Validated criterion.

MSE: the empirical Mean Square Error.

About distances

$d^{\text{BAGIDIS}}(\cdot, \cdot), d^B(\cdot, \cdot)$: the BAGIDIS semi-distance between two series or images. Lower indexes λ (balance parameter) or p (kind of norm) are sometimes written to account for the parameterization of the distance.

$d^{\text{EUC}}(\cdot, \cdot), d^{L_2}(\cdot, \cdot)$: the Euclidean (L_2) distance.

$d^{\text{HAAR}}(\cdot, \cdot)$: the Haar-based distance, i.e. the Euclidean distance between the expansion coefficients of two series in the Haar basis.

$d_q^{\text{deriv}}(\cdot, \cdot)$: the derivative-based semi-distance, i.e. the Euclidean distance between the derivatives of order q of two series.

$d_q^{\text{PCA}}(\cdot, \cdot)$: the PCA-based semidistance, i.e. the Euclidean distance between the q first principal components of two series.

$d^{\text{hshift}}(\cdot, \cdot)$: the *hshift* semi-distance between two series.

$d^{\text{IMED}}(\cdot, \cdot)$: the IMED distance between two images.

$d^{\text{HAUSDORFF}}(\cdot, \cdot)$: the HAUSDORFF distance between two images.

Bibliography

- Ait-Saidi, A., Ferraty, F., Kassa, R., and Vieu, P. (2008), “Cross-validated estimations in the single-functional index model,” *Statistics*, 42, 475–494.
- Aneiros-Pérez, G., Cardot, H., Estévez-Pérez, G., and Vieu, P. (2004), “Maximum ozone concentration forecasting by functional non-parametric approaches,” *Environmetrics*, 15, 675–685.
- Aneiros-Pérez, G. and Vieu, P. (2008), “Nonparametric time series prediction: A semi-functional partial linear modeling,” *Journal of Multivariate Analysis*, 99, 834–857.
- Antoine, J.-P., Rosca, D., and Vandergheynst, P. (2010), “Wavelet transform on manifolds: Old and new approaches,” *Applied and Computational Harmonic Analysis*, 28, 189 – 202.
- Antoniadis, A., Bigot, J., and von Sachs, R. (2009), “A Multiscale Approach for Statistical Characterization of Functional Images,” *Journal of Computational and Graphical Statistics*, 216–237.
- Berkner, K. and Wells, R. (2002), “Smoothness estimates for soft threshold denoising via translation invariant wavelet transforms,” *Applied Computational Harmonic Analysis*, 12, 1–24.
- Beumer, G., Tao, Q., Bazen, A., and Veldhuis, R. (2005), “Comparing landmarking methods for face recognition,” in *Proceedings of ProRISC 2005, 16th Annual Workshop on Circuits, Systems and Signal Processing*, pp. 594–597.
- Bishop, T. A. and Dudewicz, E. J. (1978), “Exact Analysis of Variance with Unequal Variances: Test Procedures and Tables,” *Technometrics*, 20, 419–430.
- Bosq, D. (2000), *Linear Processes in Function Spaces*, Berlin: Springer.
- Burba, F., Ferraty, F., and Vieu, P. (2009), “k-Nearest Neighbour method in functional nonparametric regression,” *Journal of Nonparametric Statistics*, 21, 453–469.

- Cai, T. T. and Hall, P. (2006), "Prediction in functional linear regression," *Annals of Statistics*, 34, 2159.
- Cailliez, F. (1983), "The Analytical Solution of the Additive Constant Problem," *Psychometrika*, 48, 343–349.
- Candès, E. J. and Donoho, D. L. (2001), "Curvelets and Curvilinear Integrals," *Journal of Approximation Theory*, 113, 59–90.
- Claypoole, R. L. and Baraniuk, R. G. (2000), "A Multiresolution Wedgelet Transform for Image Processing," in *SPIE Technical Conference on Wavelet Applications in Signal Processing VIII*, San Diego.
- Cohen, A. (1997), "Wavelets and Multiscale Methods in Image Processing," .
- Coifman, R. and Donoho, D. (1995), "Translation Invariant Denoising," in *Wavelets in Statistics*, eds. Antoniadis, A. and Oppenheim, G., Springer-Verlag, Lecture Notes in Statistics, pp. 125–150.
- Coifman, R. R. and Maggioni, M. (2006), "Diffusion Wavelets," *Applied and Computational Harmonic Analysis*, 21, 53–94.
- Cox, M. A. and Cox, T. F. (2008), "Multidimensional Scaling," in *Handbook of Data Visualization*, eds. Chen, C.-H., Hardle, W. K., and Unwin, A., Springer Berlin Heidelberg, Springer Handbooks of Computational Statistics, chap. 3, pp. 315–347.
- Crambes, C., Kneip, A., and Sarda, P. (2008), "Smoothing splines estimators for functional linear regression," *Annals of Statistics*.
- Crovella, M. and Kolaczyk, E. (2003), "Graph Wavelets for Spatial Traffic Analysis," in *Proceedings of IEEE Infocom*.
- Delouille, V. (2002), "Nonparametric stochastic regression using design-adapted wavelets," Ph.D. thesis, Université catholique de Louvain, Louvain-la-Neuve, Belgium.
- Delouille, V., Franke, J., and von Sachs, R. (2001), "Nonparametric stochastic regression with design-adapted wavelets," *Shankhya*, 328–366.
- Delsol, L., Ferraty, F., and Vieu, P. (2011), "Structural test in regression on functional variables," *Journal of Multivariate Analysis*, 102, 422–447.
- Donoho, D. L. (1999), "Wedgelets: nearly-minimax estimation of edges," *Annals of Statistics*, 859–897.
- Donoho, D. L. and Johnstone, I. M. (1994), "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, 81, 425–455.

- Dubois, N. (2009), “Utilisation de la H-NMR en métabonomique; pré-traitement des spectres et analyse de fidélité.” Master’s thesis, Université catholique de Louvain, Louvain-la-Neuve, Belgium.
- Ferraty, F. (ed.) (2011), *Recent Advances in Functional Data Analysis and Related Topics*, Contributions to Statistics, Physica-Verlag - Springer.
- Ferraty, F., Laksaci, A., Tadj, A., and Vieu, P. (2010), “Rate of uniform consistency for nonparametric estimates with functional variables,” *Journal of Statistical Planning and Inference*, 140, 335–352.
- (2011), “Kernel regression with functional response,” *Electronic Journal of Statistics*, 5, 159–171.
- Ferraty, F., Rabhi, A., and Vieu, P. (2005), “Conditional Quantiles for Dependent Functional Data with Application to the Climatic El Nino Phenomenon,” *Sankhya : The Indian Journal of Statistics*, 67, 378–398, special Issue on Quantile Regression and Related Methods.
- Ferraty, F. and Romain, Y. (2010), *The Oxford Handbook of Functional Data Analysis*, City: Oxford University Press, USA.
- Ferraty, F. and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, Springer Series in Statistics, Secaucus, NJ, USA: Springer-Verlag New York, Inc.
- François, D., Wertz, V., and Verleysen, M. (2011), “Choosing the metric: a simple model approach,” in *Meta-Learning in Computational Intelligence*, ed. Jankowski, N., SCI, Springer-Verlag, no. 358, pp. 97–115.
- Frome, A., Singer, Y., and Malik, J. (2006), “Image Retrieval and Classification Using Local Distance Functions,” in *NIPS*, eds. Schölkopf, B., Platt, J. C., and Hoffman, T., MIT Press, pp. 417–424.
- Fryzlewicz, P. (2007), “Unbalanced Haar Technique for Non Parametric Function Estimation,” *Journal of the American Statistical Association*, 102, 1318–1327.
- (2012), “Time-Threshold Maps: using information from wavelet reconstruction with all thresholds simultaneously,” *Journal of the Korean Statistical Society*, 41, 145–159.
- Gavish, M., Nadler, B., and Coifman, R. R. (2010), “Multiscale Wavelets on Trees, Graphs and High Dimensional Data: Theory and Applications to Semi Supervised Learning,” in *ICML*, pp. 367–374.

- Geng, S., Schneeman, P., and Wang, W.-J. (1982), "An Empirical Study of the Robustness of Analysis of Variance Procedures in the Presence of Commonly Encountered Data Problems," *American Journal of Enology and Viticulture*, 33, 131–134.
- Giorgino, T. (2009), "Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package," *Journal of Statistical Software*, 7, 1–24.
- Girardi, M. and Sweldens, W. (1997), "A new Class of Unbalanced Haar Wavelets that form an Unconditional Basis for L_p on General Measure Spaces," *Journal of Fourier Analysis and Applications*, 3, 457–474.
- Guyon, I. and Elisseeff, A. (2003), "An introduction to variable and feature selection," *Journal of Machine Learning Research*, 3, 1157–1182.
- Hammond, D. K., Vandergheynst, P., and Gribonval, R. (2009), "Wavelets on Graphs via Spectral Graph Theory," *Applied and Computational Harmonic Analysis*, 30, 129–150.
- Hartigan, J. A. and Wong, M. A. (1979), "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28, 100–108.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, Springer Series in Statistics, New York, NY, USA: Springer New York Inc.
- Härdle, W. and Simar, L. (2003), *Applied multivariate statistical analysis*, Berlin [u.a.]: Springer.
- Hu, R., Rüger, S. M., Song, D., Liu, H., and Huang, Z. (2008), "Dissimilarity measures for content-based image retrieval," in *ICME*, IEEE, pp. 1365–1368.
- Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. (1993), "Comparing Images Using the Hausdorff Distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, 850–863.
- Jansen, M., Nason, G., and Silverman, B. (2009), "Multiscale methods for data on graphs and irregular multidimensional situations," *Journal of the Royal Statistical Society B, Statistical Methodology*, 71, 97–125.
- Jolliffe, I. (2002), *Principal Component Analysis (Second Edition)*, Springer Series in Statistics, Secaucus, NJ, USA: Springer-Verlag New York, Inc.
- Kaufman, L. and Rousseeuw, P. (1990), *Finding Groups in Data An Introduction to Cluster Analysis*, New York: Wiley Interscience.
- Kovac, A. and Smith, A. (2011), "Nonparametric regression on a graph," *Journal of Computational and Graphical Statistics*, 20, 432–447.

- Krommweh, J. (2010), "Tetrolet Transform: A New Adaptive Haar Wavelet Algorithm for Sparse Image Representation," *Journal of Visual Communication and Image Representation*, 21, 364–374.
- Kruskal, W. and Wallis, W. (1952), "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, 47, 583–621.
- Le Pennec, E. and Mallat, S. (2005), "Sparse Geometrical Image Representation with Bandelets," *IEEE transaction on Image Processing*, 14, 423–438.
- Lebart, L., Morineau, A., and Piron, M. (2004), *Statistique exploratoire multidimensionnelle*, Dunod, 3rd ed.
- Lee, A. B., Nadler, B., and Wasserman, L. (2008), "Treelets - An adaptive multi-scale basis for sparse unordered data," *Annals of Applied Statistics*, 2, 435–471.
- Lindon, J. C., Nicholson, J. K., and Holmes, E. (2007), *The Handbook of Metabolomics and Metabolomics*, Elsevier Science, 1st ed.
- Maggioni, M., Bremer, J. C. J., Coifman, R. R., and Szlam, A. D. (2005), "Biorthogonal diffusion wavelets for multiscale representations on manifolds and graphs," SPIE, vol. 5914.
- Mallat, S. (1999), *A Wavelet Tour of Signal Processing*, Academic Press, 2nd ed.
- (2009), "Geometrical grouplets," *Applied and Computational Harmonic Analysis*, 26, 161–180.
- Mallat, S. and Peyré, G. (2008), "Orthogonal Bandlet Bases for Geometric Images Approximation," *Communications on Pure and Applied Mathematics*, 61, 1173–1212.
- Marron, J. and Härdle, W. (1986), "Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation," *Journal of Multivariate Analysis*, 20, 91–113.
- Milanfar, P. (2011), "A Tour of Modern Image Filtering," *To appear as a feature article in IEEE Signal Processing Magazine*.
- Morris, J. S. and Carroll, R. J. (2006), "Wavelet-based Functional Mixed Models," *Journal Of The Royal Statistical Society Series B*, 68, 179–199.
- Murtagh, F. (2007), "The Haar wavelet transform of a dendrogram," *Journal of Classification*, 24, 2006.

- Noël, A., Jost, M., Lambert, V., Lecomte, J., and Rakic, J.-M. (2007), "Anti-angiogenesis therapy of exudative age-related macular degeneration: current progress and emerging concepts," *Trends in Molecular Medicine*, 345–352.
- Peyré, G. (2011), "A Review of Adaptive Image Representations," *IEEE Journal of Selected Topics in Signal Processing*, 5, 896–911.
- Plonka, G. (2009), "The easy path wavelet transform: A new adaptive wavelet transform for sparse representation of two-dimensional data," *Multiscale Modeling and Simulation*, 7, 1474–1496.
- Polzehl, J. and Spokoiny, V. (2000), "Adaptive weights smoothing with applications to image restoration," *Journal of the Royal Statistical Society, Series B*, 62, 335–354.
- (2003), "Image denoising: pointwise adaptive approach," *Annals of Statistics*, 31, 30–57.
- R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
- Rachdi, M. and Vieu, P. (2007), "Nonparametric regression for functional data: Automatic smoothing parameter selection," *Journal of Statistical Planning and Inference*, 137, 2784–2801.
- Ramsay, J. and Dalzell, C. (1991), "Some tools for functional data analysis (with discussion)," *Journal of the Royal Statistical Society, Series B*, 53, 539–572.
- Ramsay, J. and Silverman, B. (2002), *Applied functional data analysis : methods and case studies*, Springer Series in Statistics, New York: Springer-Verlag.
- (2005), *Functional Data Analysis*, Springer Series in Statistics, Springer, 2nd ed.
- Ross, A., Schlotterbeck, G., Dieterle, F., and Senn, H. (2007), "NMR Spectroscopy Techniques for Application to Metabonomics," in *The Handbook of Metabonomics and Metabolomics*, eds. Lindon, J. C., Nicholson, J. K., and Holmes, E., Elsevier Science.
- Rousseau, R. (2011), "Statistical contribution to the analysis of metabonomics data in ^1H NMR spectroscopy." Ph.D. thesis, Université catholique de Louvain, Louvain-la-Neuve, Belgium.
- Rousseau, R., Govaerts, B., Verleysen, M., and Boulanger, B. (2008), "Comparison of some chemometric tools for metabonomics biomarkers identification," *Chemometric and intelligent laboratory systems*, 54–66.

- Singh, A., Nowak, R. D., and Calderbank, A. R. (2010), "Detecting Weak but Hierarchically-Structured Patterns in Networks," *Journal of Machine Learning Research - Proceedings Track*, 9, 749–756.
- Sood, A., James, G. M., and Tellis, G. J. (2009), "Functional Regression: A New Model for Predicting Market Penetration of New Products," *Marketing Science*, 28, 36–51.
- Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053.
- Sweldens, W. (1996), "Wavelets and the lifting scheme: A 5 minute tour," *Zeitschrift für Angewandte Mathematik und Mechanik*, 76 (Suppl. 2), 41–44.
- Szlam, A. D., Maggioni, M., Coifman, R. R., and Bremer, J. C. J. (2005), "Diffusion-driven multiscale analysis on manifolds and graphs: top-down and bottom-up constructions," *SPIE*, vol. 5914.
- Van der Vaart, A. and Wellner, J. (1996), *Weak convergence and empirical processes*, Springer Series in Statistics, Secaucus, NJ, USA: Springer-Verlag New York, Inc.
- Vanwinsberghe, J. (2005), "Bubble: developpement of a Matlab tool for automated H-NMR data processing in metabonomics," Master's thesis, Université de Strasbourg, Strasbourg, France, <http://www.clinbay.com/Download,007.html>.
- Wang, L., Zhang, Y., and Feng, J. (2005), "On the Euclidean Distance of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 1334–1339.
- Weerahandi, S. (1995), "ANOVA under Unequal Error Variances," *Biometrics*, 51, 589–599.
- Welch, B. (1947), "The Generalization of 'Student's' Problem when Several Different Population Variances are Involved," *Biometrika*, 34, 28–35.
- Woodraska, D. and Woods, T. (2009), *TIMED SEE Version 10 Data Product Release Notes*, main URL for Timed SEE data: <http://lasp.colorado.edu/see>. Data have been downloaded on March 12, 2010.
- Woods, T. N., Eparvier, F. G., Bailey, S. M., Chamberlin, P. C., Lean, J., Rottman, G. J., Solomon, S. C., Tobiska, W. K., and Woodraska, D. L. (2005), "Solar EUV Experiment (SEE): Mission overview and first results," *Journal of Geophysical Research (Space Physics)*, 110, 1312–1336.
- Zitová, B. and Flusser, J. (2003), "Image registration methods: a survey," *Image and Vision Computing*, 21, 977–1000.

Appendix A

Technical details for Chapter 1

This appendix details the way we obtained the results shown in Figure 1.17. We computed sliding BAGIDIS semi-distances with $\Delta = 25$, $p = 1$, $\lambda = 1$, between the spectra of *day A* ($d(A, A)$), of *day C* ($d(C, C)$) and of *day A* and *day C* ($d(A, C)$). Then, for each component of the sliding vectors of semi-distances, we used a Welch t-test for the equality of the semi-distances $d(A, C) = d(A, A)$ and $d(A, C) = d(C, C)$, versus the alternatives that $d(A, C)$ is larger than $d(A, A)$ or $d(C, C)$. Normality of the semi-distances in each group was checked and there was no reason to reject this hypothesis. The resulting p-values were then adjusted for multiple comparisons using a Bonferroni correction. Then, for both tests, we selected regions of the spectra where the p-values of one or both of the tests were smaller than 0.05. Combining those results according to the interpretation table in Figure 1.9 we obtained the results shown on rows 2, 4 and 5 in Figure 1.17. The same testing procedure was then applied using a BAGIDIS sliding distance with parameters $\Delta = 25$, $p = 1$ and $\lambda = 0$. This time, the assumption of normality was not always satisfied, so that we compared the logarithms of the distances when necessary, possibly adding a constant value to the distances in both groups so as to avoid to take a logarithm of zero. After this transformation, the normality assumption could not be rejected anymore. Results shown on row 3 in Figure 1.17, identify the parts of the spectra where spectra of *day A* and *day C* are significantly different to each other, according to the distance computed with $\lambda = 0$. All those results can be compared to the parts of the spectra that were detected significantly different from one day to the other using classical Welch t-tests on the equality of the measurements, at each ppm, with p-values corrected for multiple comparisons (Bonferroni correction) and data taken in logarithm when it was necessary so as to satisfy the assumption of normality of the residuals of the ANOVA. The corresponding significantly different parts of the spectra are identified on row 1, in Figure 1.17.

Appendix B

Proofs of the main results from Chapter 2

In what follows, we denote by C, C' positive constants, whose value might change from one line to another.

B.1 Proof of Theorem 1

Proof of Theorem 1 is a direct consequence of Lemma 3 and Lemma 4 above.

Lemma 3. *Assume conditions (2.2.1), (2.2.5), (2.2.6) and (2.2.7) to be satisfied. Then, the random variable χ defined by $(\mathbf{b}, \mathbf{d}) \in \mathbb{N}_{[0;N-1]} \times \mathbb{R}^N$ is fractal of order K with respect to the BAGIDIS semimetric at point $\chi \equiv (b, d)$. This means that the small ball probability function $\phi_{d^B, \chi}(\cdot) = P(\chi \in B_{d^B}(\chi, \cdot))$ of χ about χ is such that there exists a positive constant C such that*

$$\phi_{d^B, \chi}(h) \sim C h^K, \quad \text{when } h \text{ tends to } 0.$$

Proof. We have

$$\phi_{d^B, \chi}(h) = P(d^B(\chi, \chi) \leq h) \tag{B.1.1}$$

$$\begin{aligned} &= P(d^B(\chi, \chi) \leq h \cap \forall k \in \mathcal{K}, \mathbf{b}^k = b^k) \\ &\quad + P(d^B(\chi, \chi) \leq h \cap \exists k \in \mathcal{K}, \mathbf{b}^k \neq b^k) \end{aligned} \tag{B.1.2}$$

If $\exists k \in \mathcal{K}$, $\mathbf{b}^k \neq b^k$, it implies that $\exists k \in \mathcal{K}$ such that $|\mathbf{b}^k - b^k| \geq 1$, where 1 is the step of the grid $\mathbb{N}_{[0;N-1]}$ on which the curve is observed. In such a case, we have

$$\begin{aligned} d^B(\chi, \chi) &= \sum_{k \in \mathcal{K}} w_k \left(\lambda |b^k - \mathbf{b}^k|^p + \underbrace{(1 - \lambda) |d^k - \mathbf{d}^k|^p}_{\geq 0} \right)^{1/p} \\ &\geq \sum_{k \in \mathcal{K}} w_k \lambda^{1/p} |b^k - \mathbf{b}^k| \\ &\geq \lambda^{1/p} \min_{k \in \mathcal{K}}(w_k). \end{aligned}$$

Consequently, the right hand side term of equation (B.1.1) has probability zero for $h < \lambda^{1/p} \min_{k \in \mathcal{K}}(w_k)$. Thus, when h tends to zero, we have

$$\begin{aligned} \phi_{d, \chi}(h) &= P\left(\underbrace{\sum_{\forall k \in \mathcal{K}} w_k (1 - \lambda)^{1/p} |\mathbf{d}^k - d^k|}_{\equiv \|\mathbf{d} - d\|_{\lambda, w}} \leq h \cap \forall k \in \mathcal{K}, \mathbf{b}^k = b^k \right) \quad (\text{B.1.3}) \\ &= \underbrace{P(\forall k \in \mathcal{K}, \mathbf{b}_k = b_k)}_{>0} \cdot P(\|\mathbf{d} - d\|_{\lambda, w} \leq h \mid \forall k \in \mathcal{K}, \mathbf{b}_k = b_k) \end{aligned}$$

By definition, the last term of this expression is

$$P(\|\mathbf{d} - d\|_{\lambda, w} \leq h \mid \forall k \in \mathcal{K}, \mathbf{b}^k = b^k) = \int_{B(d, h)_{\lambda, w}} f_{|b}(b, s) ds.$$

Then, by the continuity condition of the conditional density, we have $\forall \epsilon > 0, \forall h < \min(\delta_\epsilon, \min(w_k))$,

$$\begin{aligned} &\left\| \int_{B(d, h)_{\lambda, w}} f_{d|b}^{\mathcal{K}}(s) ds - \int_{B(d, h)_{\lambda, w}} f_{d|b}^{\mathcal{K}}(d) ds \right\| \\ &= \left\| \int_{B(d, h)_{\lambda, w}} \left\{ f_{d|b}^{\mathcal{K}}(s) - f_{d|b}^{\mathcal{K}}(d) \right\} ds \right\| \end{aligned}$$

$$\begin{aligned}
&\leq \int_{B(d,h)_{\lambda,w}} \sup_{\|s-d\|_{\lambda,w} \leq \delta_\epsilon} \left| f_{d|b}^{\mathcal{K}}(s) - f_{d|b}^{\mathcal{K}}(d) \right| ds \\
&\leq \int_{B(d,h)_{\lambda,w}} \epsilon ds \\
&= \underbrace{\epsilon \cdot \int_{B(d,h)_{\lambda,w}} ds}_{V_{\lambda,w}(N;h)}
\end{aligned}$$

where $B(d,h)_{\lambda,w}$ is the N -dimensional ball of radius h , centered on d , at the sense of norm $\|\cdot\|_{\lambda,w}$ and $V_{\lambda,w}(N;h)$ is the volume of this ball. Hence, we have

$$\forall \epsilon > 0, \forall h < \min(\delta_\epsilon, \min(w_k)),$$

$$\left| \underbrace{\int_{B(d,h)_{\lambda,w}} f_{d|b}^{\mathcal{K}}(s) ds}_{P(\|\mathbf{d}-d\|_{\lambda,w} \leq h | \mathbf{b}=b)} - f_{|b}(b,d) \cdot \underbrace{\int_{B(d,h)_{\lambda,w}} ds}_{\equiv V_{\lambda,w}(N;h)} \right| \leq \epsilon V_{\lambda,w}(N;h).$$

As $V_{\lambda,w}(N;h) f_{d|b}^{\mathcal{K}}(d) > 0$, this means $\forall \epsilon > 0, \forall h < \min(\delta_\epsilon, \min(w_k))$,

$$\left\| \frac{P(\|\mathbf{d}-d\|_{\lambda,w} \leq h | \mathbf{b}=b)}{V_{\lambda,w}(N;h) f_{d|b}^{\mathcal{K}}(d)} - 1 \right\| \leq \frac{\epsilon}{f_{|b}(b,d)} \leq C,$$

with $C > 0$. Consequently, and because $V_{\lambda,w}(N;h) \sim C' . h^K$ with $C' > 0$, we have

$$P(\|\mathbf{d}-d\|_{\lambda,w} \leq h | \forall k \in \mathcal{K}, \mathbf{b}^k = b^k) = \underbrace{V_{\lambda,w}(N;h)}_{\sim C' . h^K} \underbrace{f_{d|b}^{\mathcal{K}}(d)}_{>0} \sim C'' . h^K,$$

when h tends to 0 with $C'' > 0$. Going back to equality (B.1.3), it results in

$$\begin{aligned}
\phi_{d^B, \chi}(h) &= P(d^B(\chi, \chi) \leq h) \\
&= \underbrace{P(\forall k \in \mathcal{K} \mathbf{b}^k = b^k)}_{>0} \underbrace{P(\|\mathbf{d}-d\|_{\lambda,w} \leq h | \forall k \in \mathcal{K} \mathbf{b}^k = b^k)}_{\sim C'' . h^K} \\
&\sim C'' . h^K,
\end{aligned}$$

for h small enough. □

Lemma 4. *Assuming that there exists a finite constant $C > 0$ such that*

$$\phi_{d,\chi}(\epsilon) \sim C\epsilon^\tau \quad \text{as } \epsilon \rightarrow 0,$$

and under the conditions (2.2.9), (2.2.8), (2.2.11), (2.2.10) and (2.2.3), the functional kernel regression estimate can reach the rate of convergence:

$$\hat{r}(\chi) - r(\chi) = \mathcal{O}_{a.co.}\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+\tau}}\right)$$

Proof. This is proved in (Ferraty and Vieu 2006). □

The proof of Theorem 1 is a direct consequence of Lemmas 3 and 4, for the semimetric $d = d^B$.

B.2 Proof of Theorem 2

In order to prove Theorem 2, we consider the following Lemmas.

Lemma 5. *Under conditions of Theorem 2, we have*

$$\exists C > 0, \text{ such that } \frac{C}{n\Phi_H} + \frac{n-1}{n}b_H \leq MISE^*(H).$$

Proof. We start with the following decomposition of $MISE^*(H)$, that holds because of condition (2.2.15):

$$MISE^*(H) = \int \mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^n \delta_{i\chi}\right)^2\right) W(\chi) dP_\chi(\chi) \quad (\text{B.2.1})$$

$$= \int \mathbb{E}\left(\frac{1}{n^2}\sum_{i=1}^n \delta_{i\chi}^2\right) W(\chi) dP_\chi(\chi) \quad (\text{B.2.2})$$

$$\begin{aligned} &+ \int \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^n \mathbb{E}(\delta_{i\chi}\delta_{j\chi}|\chi) W(\chi) dP_\chi(\chi) \\ &= \frac{1}{n} \int \mathbb{E}(\delta_{i\chi}^2) W(\chi) dP_\chi(\chi) + \frac{n-1}{n} \int \mathbb{E}^2(\delta_{i\chi}|\chi) W(\chi) dP_\chi(\chi) \\ &= R_3(H) + \frac{n-1}{n}b_H, \end{aligned}$$

with

$$R_3(H) = \frac{1}{n} \int \mathbb{E}(\delta_{i\chi}^2) W(\chi) dP_\chi(\chi). \quad (\text{B.2.3})$$

It remains to bound $R_3(H)$ from below. We consider

$$\begin{aligned} \mathbb{E}(\delta_{i\chi}^2) &= \mathbb{E}\left(\left((Y_i - r(\chi_i) + r(\chi_i) - r(\chi))K_H(\chi, \chi_i)\right)^2\right) \\ &= \mathbb{E}\left(\left(\epsilon_i K_H(\chi, \chi_i) + (r(\chi_i) - r(\chi))K_H(\chi, \chi_i)\right)^2\right) \\ &= \mathbb{E}\left(\epsilon^2 K_H^2(\chi, \chi_i)\right) + \mathbb{E}\left((r(\chi_i) - r(\chi))^2 K_H^2(\chi, \chi_i)\right) \\ &\quad + 2\mathbb{E}\left(\epsilon_i (r(\chi_i) - r(\chi))K_H(\chi, \chi_i)\right). \end{aligned}$$

The last term of this equation is null, because of condition (2.2.21), and the second term is positive. Thus, using conditions (2.2.22) and (2.2.16), we have

$$\begin{aligned} \mathbb{E}(\delta_{i\chi}^2) &\geq \mathbb{E}(\epsilon_i^2 K_H^2(\chi, \chi_i)) = \mathbb{E}\left(\mathbb{E}(\epsilon_i^2 | \chi_i) \mathbb{E}(K_H^2(\chi, \chi_i) | \chi_i)\right) \\ &\geq \sigma_0^2 \mathbb{E}\left(\frac{K^2\left(\frac{d(\chi, \chi_i)}{h}\right)}{\mathbb{E}^2\left(K\left(\frac{d(\chi, \chi_i)}{h}\right)\right)}\right) \\ &\geq \frac{\sigma_0^2}{\mathbb{E}^2\left(K\left(\frac{d(\chi, \chi_i)}{h}\right)\right)} \mathbb{E}\left(K\left(\frac{d(\chi, \chi_i)}{h}\right)\right) \\ &\geq C' \frac{\sigma_0^2}{\Phi_H}, \end{aligned}$$

for a certain $C' > 0$. Then, Lemma 5 follows from

$$R_3(H) \geq \frac{1}{n} \int C' \frac{\sigma_0^2}{\Phi_H} W(\chi) dP_x \geq \frac{C}{n\Phi_H},$$

where the last inequality holds because of condition (2.2.19). □

Lemma 6. *Under conditions of Theorem 2, we have*

$$\sup_{\substack{H \in \mathcal{H}_n \\ \chi \in \mathcal{W}}} \left| \hat{r}_{1H}(\chi) - 1 \right| \longrightarrow 0 \quad a.s.$$

and

$$\sup_{\substack{H \in \mathcal{H}_n \\ \chi \subset \mathcal{W}}} \left| \hat{r}_{1H}^{-j}(\chi) - 1 \right| \longrightarrow 0 \quad a.s.$$

Proof. We denote by c_χ center that is the closest to χ in condition (2.2.20). Using condition (2.2.20), we observe that $\forall \epsilon > 0$,

$$P\left(\sup_{\substack{H \in \mathcal{H}_n \\ \chi \subset \mathcal{W}}} |\hat{r}_{1H}(\chi) - 1| > \epsilon\right) \quad (\text{B.2.4})$$

$$\leq P\left(\sup_{\substack{H \in \mathcal{H}_n \\ \chi \subset \mathcal{W}}} |\hat{r}_{1H}(\chi) - \hat{r}_{1H}(c_\chi)| > \frac{\epsilon}{2}\right) \quad (\text{B.2.5})$$

$$+ \#\mathcal{H}_n d_n \sup_{\substack{H \in \mathcal{H}_n \\ \chi \subset \mathcal{W} \\ 1 \leq k \leq d_n}} \left(P(|\hat{r}_{1H}(c_k) - 1| > \frac{\epsilon}{2})\right). \quad (\text{B.2.6})$$

We first consider the first term on the right side of the inequality. We have

$$\begin{aligned} & |\hat{r}_{1H}(\chi) - \hat{r}_{1H}(c_\chi)| \quad (\text{B.2.7}) \\ &= \left| \frac{\sum_{i=1}^n \Delta_i(\chi)}{n\mathbb{E}(\Delta_i(\chi))} - \frac{\sum_{i=1}^n \Delta_i(c_\chi)}{n\mathbb{E}(\Delta_i(c_\chi))} \right| \\ &= \left| \frac{\sum_{i=1}^n \Delta_i(\chi)\mathbb{E}(\Delta_i(c_\chi)) - \sum_{i=1}^n \Delta_i(c_\chi)\mathbb{E}(\Delta_i(\chi))}{n\mathbb{E}(\Delta_i(\chi))\mathbb{E}(\Delta_i(c_\chi))} \right| \\ &= \left| \frac{\sum_{i=1}^n \left((\Delta_i(\chi) - \Delta_i(c_\chi))\mathbb{E}(\Delta_i(c_\chi)) + \Delta_i(c_\chi)(\mathbb{E}(\Delta_i(c_\chi)) - \mathbb{E}(\Delta_i(\chi))) \right)}{n\mathbb{E}(\Delta_i(\chi))\mathbb{E}(\Delta_i(c_\chi))} \right|. \end{aligned}$$

Then, because the kernel is Lipschitz on $\mathbb{R}^+$ by condition (2.2.16), we know that

$$\left| \Delta_i(\chi) - \Delta_i(c_\chi) \right| \leq C \frac{d(\chi, c_\chi)}{h} \leq C \frac{r_n}{h},$$

and

$$\left| \mathbb{E}(\Delta_i(c_\chi) - \Delta_i(\chi)) \right| \leq \mathbb{E}\left(\left| \Delta_i(c_\chi) - \Delta_i(\chi) \right| \mathbf{1}_{d(\chi_i, \chi) \leq h \cup d(\chi_i, c_\chi) \leq h}\right) \leq C \frac{r_n}{h} \Phi_H$$

Thus, equation (B.2.7) gives

$$\sup_{\substack{H \in \mathcal{H}_n \\ \chi \subset \mathcal{W}}} |\hat{r}_{1H}(\chi) - \hat{r}_{1H}(c_\chi)| \leq C \frac{r_n}{\inf_{\substack{H \in \mathcal{H}_n \\ \chi \subset \mathcal{W}}} n\Phi_H},$$

which tends to 0 because of condition (2.2.18), so that

$$P\left(\sup_{\substack{H \in \mathcal{H}_n \\ \chi \subset \mathcal{W}}} |\hat{r}_{1H}(\chi) - \hat{r}_{1H}(c_\chi)| > \frac{\epsilon}{2}\right) = 0 \quad (\text{B.2.8})$$

for n large enough. We now consider the second term on the right side of inequality (B.2.6). We have

$$\hat{r}_{1H}(c_k) - 1 = \frac{1}{n} \sum_{j=1}^n (K_H(c_k, \chi_j) - 1) = \frac{1}{n\mathbb{E}(\Delta_j(c_k))} \sum_{j=1}^n U_j,$$

with

$$U_j = \Delta_j(c_k) - \mathbb{E}(\Delta_j(c_k)).$$

We would like to make use of a Bernstein inequality (see Van der Vaart and Wellner (1996, Lemma 2.2.11), for instance), for the U_j . By construction, we have $\mathbb{E}(U_j) = 0$, and $|U_j|$ is bounded, as the kernel is bounded. Moreover, using condition (2.2.16) and the fact that $\Phi_H \leq 1$,

$$\mathbb{E}\left(\left(\Delta_j(c_k) - \mathbb{E}(\Delta_j(c_k))\right)^2\right) = \mathbb{E}(\Delta_j^2(c_k)) - \mathbb{E}^2(\Delta_j(c_k)) \quad (\text{B.2.9})$$

$$\leq C_{2,2}\Phi_H + C_{2,1}^2\Phi_H^2 \leq C\Phi_H. \quad (\text{B.2.10})$$

Finally, as U_j is bounded and for $m \geq 2$, we have

$$\mathbb{E}(|U_j|^m) = \mathbb{E}(|U_j|^2 |U_j|^{m-2}) \leq \mathbb{E}(|U_j|^2 C^{m-2}) \leq C\Phi_H \leq \frac{m!}{2} C' \Phi_H.$$

This tells us that we are in the conditions of the Bernstein inequality, with $M = 1$, $v_i = C' \Phi_H$ and $v = \sum_{i=1}^n C' \Phi_H = nC' \Phi_H$. This ensures that, for all ϵ positive,

$$\begin{aligned}
P(|\hat{r}_{1H}(c_k) - 1| > \epsilon) &= P\left(\left|\frac{1}{n\mathbb{E}(\Delta_j(c_k))} \sum_{j=1}^n U_j\right| > \epsilon\right) \\
&= P\left(\left|\sum_{j=1}^n U_j\right| > \epsilon n\mathbb{E}(\Delta_j(c_k))\right) \\
&\leq P\left(\left|\sum_{j=1}^n U_j\right| > \epsilon nC_{1,1}\Phi_H\right) \\
&\leq 2 \exp\left(-\frac{(\epsilon nC_{1,1}\Phi_H)^2}{2(nC' \Phi_H + \epsilon nC_{1,1}\Phi_H)}\right) \\
&\leq 2 \exp\left(-\frac{1}{2} \frac{\epsilon^2 C_{1,1}^2 n \Phi_H}{C' + \epsilon C_{1,1}}\right) \\
&= 2 \exp(-C_\epsilon n \Phi_H)
\end{aligned}$$

where condition (2.2.16) is used. From this expression and from equations (B.2.6) and (B.2.8), and by using conditions (2.2.18), (2.2.20) and (2.2.24), it follows that $\forall \epsilon > 0$,

$$\begin{aligned}
0 &\leq \sum_{n=1}^{\infty} P\left(\sup_{\substack{H \in \mathcal{H}_n \\ \chi \subset \mathcal{W}}} |\hat{r}_{1H}(\chi) - 1| > \epsilon\right) \\
&\leq \sum_{n=1}^{\infty} P\left(\sup_{\substack{H \in \mathcal{H}_n \\ \chi \subset \mathcal{W}}} |\hat{r}_{1H}(\chi) - \hat{r}_{1H}(c_\chi)| > \frac{\epsilon}{2}\right) \tag{B.2.11}
\end{aligned}$$

$$\begin{aligned}
&+ \sum_{n=1}^{\infty} \#\mathcal{H}_n d_n \sum_{n=1}^{\infty} \sup_{\substack{H \in \mathcal{H}_n \\ \chi \subset \mathcal{W} \\ 1 \leq k \leq d_n}} P(|\hat{r}_{1H}(c_k) - 1| > \frac{\epsilon}{2}) \\
&\leq \sum_{n=1}^{\infty} P\left(C \frac{r_n}{\inf_{\substack{H \in \mathcal{H}_n \\ \chi \subset \mathcal{W}}} n \Phi_H} > \frac{\epsilon}{2}\right) + \sum_{n=1}^{\infty} n^{\alpha+\eta} \sup_{\substack{H \in \mathcal{H}_n \\ \chi \subset \mathcal{W} \\ 1 \leq k \leq d_n}} 2 \exp(-C_\epsilon n \Phi_H) \\
&\leq C + \sum_{n=1}^{\infty} n^{\alpha+\eta} 2 \exp(-C_\epsilon \inf_{\substack{H \in \mathcal{H}_n \\ \chi \subset \mathcal{W} \\ 1 \leq k \leq d_n}} (n \Phi_H)) \\
&\leq C + \sum_{n=1}^{\infty} n^{\alpha+\eta} 2 \exp(-C' n^\delta) \leq C'' \tag{B.2.12}
\end{aligned}$$

This ensures the uniform almost complete convergence of $\hat{r}_{1H}(\chi)$ to 1, itself implying the required almost sure convergence. The convergence to 1 of $\hat{r}_{1H}^{-j}(\chi)$ is shown using very similar ideas. In that case, (B.2.6) is replaced by

$$\begin{aligned}
& P\left(\sup_{\substack{H \in \mathcal{H}_n \\ j=1 \dots n \\ \chi \in \mathcal{W}}} |\hat{r}_{1H}^{-j}(\chi) - 1| > \epsilon\right) \\
& \leq P\left(\sup_{\substack{H \in \mathcal{H}_n \\ j=1 \dots n \\ \chi \in \mathcal{W}}} |\hat{r}_{1H}^{-j}(\chi) - \hat{r}_{1H}^{-j}(c_\chi)| > \frac{\epsilon}{2}\right) \\
& \quad + \#\mathcal{H}_n d_n n \sup_{\substack{H \in \mathcal{H}_n \\ j=1 \dots n \\ \chi \in \mathcal{W} \\ 1 \leq k \leq d_n}} \left(P(|\hat{r}_{1H}^{-j}(c_k) - 1| > \frac{\epsilon}{2})\right). \tag{B.2.13}
\end{aligned}$$

□

Lemma 7. We consider the quantity $\hat{g}_H(\chi) = \hat{r}_{2H}(\chi) - \hat{r}(\chi)\hat{r}_{1H}(\chi)$, and we note that $B(\chi) = \mathbb{E}(\hat{g}_H(\chi))$. Under conditions of Theorem 2, we have

1. The estimator $\hat{g}_H(\chi)$ has the form of a delta sequence:

$$\hat{g}_H(\chi) = \frac{1}{n} \sum_{i=1}^n \delta_{i\chi}, \tag{B.2.14}$$

with $\delta_{i\chi} = (Y_i - r(\chi))K_H(\chi, \chi_i)$.

2. For $k = 1, 2, \dots$ there is a constant C_k so that for any $m = 2 \dots 2k$, we have

$$\left| \int \dots \int \left[\prod_{i,i'=1}^m \delta_{i'i}^{\alpha_{i'i}} \right] \left[\prod_{i=1}^m W(\chi_i)^{\beta_i} \right] dP_\chi(\chi_1) dP_\chi(\chi_m) \right| \leq C_k \left(\frac{1}{\Phi_H} \right)^{k - \frac{m}{2}}, \tag{B.2.15}$$

where $\alpha_{i'i} = 0 \dots k$ are subject to

$$\sum_{i,i'=1}^m \alpha_{i'i} = k, \tag{B.2.16}$$

and the restriction that

for each $i = 1, \dots, m$, there is an $i' \neq i$ so that either $\alpha_{i'i}$ or $\alpha_{ii'}$ is non zero, (B.2.17)

and where $\beta_i = 0, 1$, with $\beta_i = 1$ any time an $\alpha_{i'i} \geq 1$, and with $W(\chi_i)^{\beta_i}$ taken to be 1 when $W(\chi_i) = \beta_i = 0$.

3. The quantity $\tilde{\delta}_{ij} = \int \delta_{j\chi} \delta_{i\chi} W(\chi) dP(\chi)$ is such that for $k = 1, 2, \dots$ there is a constant C'_k so that for any $m = 2 \dots 2k$, we have

$$|\int \dots \int \left[\prod_{i,i'=1}^m \tilde{\delta}_{i'i}^{\alpha_{i'i}} \right] dP_\chi(\chi_1) dP_\chi(\chi_m)| \leq C_k \left(\frac{1}{\Phi_H} \right)^{k - \frac{m}{2}}, \quad (\text{B.2.18})$$

where $\alpha_{i'i} = 0 \dots k$ are subject to

$$\sum_{i,i'=1}^m \alpha_{i'i} = k, \quad (\text{B.2.19})$$

and the restriction that

for each $i = 1 \dots m$, there is an $i' \neq i$ so that either $\alpha_{ii'}$ or $\alpha_{ii'}$ is non zero. (B.2.20)

4. There exists a constant $C > 0$ such that

$$\int \int \tilde{\delta}_{ij} dP_\chi(\chi_i) dP_\chi(\chi_j) \leq C. \quad (\text{B.2.21})$$

5. There exists a constant $C' > 0$ such that

$$\int \tilde{\delta}_{\chi\chi} dP_\chi(\chi) \geq \frac{C}{\Phi_H}. \quad (\text{B.2.22})$$

6. There exists $\xi > 0$ so that for $k = 1, 2, \dots$ there is a constant $C'_k > 0$ such that

$$\int B(\chi)^{2k} W(\chi) dP_\chi(\chi) \leq C_k b_H. \quad (\text{B.2.23})$$

Proof. The proof of (B.2.14) is trivial by definition of $\hat{g}_H(\chi)$. The proof of (B.2.15) is as follows. Because of condition (2.2.16), we have

$$\prod_{i,i'=1}^m \mathbb{E}^{\alpha_{i'i}} \left(K \left(\frac{d(\chi, \chi_i)}{h} \right) \right) = \mathbb{E}^{\sum \alpha_{i'i}} \left(K \left(\frac{d(\chi, \chi_i)}{h} \right) \right) \quad (\text{B.2.24})$$

$$= \mathbb{E}^k \left(K \left(\frac{d(\chi, \chi_i)}{h} \right) \right) \geq C_{1,1}^k \Phi_H^k. \quad (\text{B.2.25})$$

Then, using successively the definition of $K_H(\chi, \chi_i)$ in $\delta_{i\chi}$, the fact that r is bounded, the fact that W is bounded, the Newton binome, condition (2.2.25), and condition (2.2.16), we have

$$\begin{aligned}
& \left| \mathbb{E} \left(\left[\prod_{i,i'=1}^m \delta_{i'i}^{\alpha_{i'i}} \right] \left[\prod_{i=1}^m W(\chi_i)^{\beta_i} \right] \right) \right| \\
& \leq \frac{1}{C_{1,1}^k \Phi_H^k} \left| \mathbb{E} \left(\left[\prod_{i,i'=1}^m \left((Y_{i'} - r(\chi_i)) K \left(\frac{d(\chi_i, \chi_{i'})}{h} \right) \right)^{\alpha_{i'i}} \right] \left[\prod_{i=1}^m W(\chi_i)^{\beta_i} \right] \right) \right| \\
& \leq \frac{1}{C_{1,1}^k \Phi_H^k} \left| \mathbb{E} \left(\left[\prod_{i,i'=1}^m \left((|Y_{i'}| + C_r) K \left(\frac{d(\chi_i, \chi_{i'})}{h} \right) \right)^{\alpha_{i'i}} \right] \left[\prod_{i=1}^m W(\chi_i)^{\beta_i} \right] \right) \right| \\
& \leq \frac{1}{C_{1,1}^k \Phi_H^k} \left| \mathbb{E} \left(\left(\mathbb{E} \left[\prod_{i,i'=1}^m \left((|Y_{i'}| + C_r) K \left(\frac{d(\chi_i, \chi_{i'})}{h} \right) \right)^{\alpha_{i'i}} \right] \right. \right. \right. \\
& \quad \left. \left. \left. \left[\prod_{i=1}^m W(\chi_i)^{\beta_i} \right] |\chi_1 \dots \chi_m \right) \right) \right| \\
& \leq \frac{1}{C_{1,1}^k \Phi_H^k} \left| \mathbb{E} \left(C_W \prod_{i,i'=1}^m \mathbb{E} \left((|Y_{i'}| + C_r)^{\alpha_{i'i}} |\chi_1 \dots \chi_m| K^{\alpha_{i'i}} \left(\frac{d(\chi_i, \chi_{i'})}{h} \right) \right) \right) \right| \\
& \leq \frac{1}{C_{1,1}^k \Phi_H^k} \left| \mathbb{E} \left(C_W \prod_{i,i'=1}^m 2^{\alpha_{i'i}} \left(\mathbb{E} (|Y_{i'}|^{\alpha_{i'i}} |\chi_1 \dots \chi_m|) + C_r^{\alpha_{i'i}} \right) \right. \right. \\
& \quad \left. \left. K^{\alpha_{i'i}} \left(\frac{d(\chi_i, \chi_{i'})}{h} \right) \right) \right| \\
& \leq \frac{1}{C_{1,1}^k \Phi_H^k} \left| \mathbb{E} \left(C_W \prod_{i,i'=1}^m \sup_{s \in 1, 2, \dots, k} \left(2^s (C_k^s + C_r^s) \right) K^{\alpha_{i'i}} \left(\frac{d(\chi_i, \chi_{i'})}{h} \right) \right) \right| \\
& \leq \frac{1}{C_{1,1}^k \Phi_H^k} \left| \mathbb{E} \left(C_W \prod_{i,i'=1}^m C_k' K^{\alpha_{i'i}} \left(\frac{d(\chi_i, \chi_{i'})}{h} \right) \right) \right| \\
& = \frac{C_k''}{\Phi_H^k} \left| \mathbb{E} \left(\prod_{i,i'=1}^m K \left(\frac{d(\chi_i, \chi_{i'})}{h} \right)^{\alpha_{i'i}} \right) \right| \leq C_k'' \Phi_H^{-k + \frac{m}{2}}. \tag{B.2.26}
\end{aligned}$$

The last inequality comes from condition (2.2.16) and from the fact that it is always possible to find $\text{floor}(\frac{m+1}{2})$ pairs (i_l, j_l) such that for all $l = 1 \dots \text{floor}(\frac{m+1}{2})$, $i_l \neq j_l$ and i_l or j_l is unique among the set of pairs $(i_l, j_l)_{l=1 \dots \text{floor}(\frac{m+1}{2})}$. There are thus at

least $\frac{m}{2}$ pairs i, i' such that $\alpha_{i'i}$ is non zero. As a consequence, (B.2.26) is valid and it proves statement (B.2.15). The proof of statement (B.2.18) follows nearly the same steps, and we have

$$\begin{aligned} & \left| \mathbb{E} \left(\left[\prod_{i,i'=1}^m \tilde{\delta}_{i'i}^{\alpha_{i'i}} \right] \right) \right| \\ & \leq \frac{1}{C_{1,2}^{2k} \Phi_H^{2k}} \left| \mathbb{E} \left(\prod_{i,i'=1}^m \left(\int K \left(\frac{d(\chi_i, \chi)}{h} \right) K \left(\frac{d(\chi_{i'}, \chi)}{h} \right) dP_\chi(\chi) \right)^{\alpha_{i'i}} \right) \right|. \end{aligned}$$

Again, we know that there exists at least $\frac{m}{2}$ pairs i, i' such that $\alpha_{i'i}$ is non zero and i or i' is unique. For those pairs, we use Holder's inequality and condition (2.2.16) to see that

$$\mathbb{E} \left(\left(\int K \left(\frac{d(\chi_i, \chi)}{h} \right) K \left(\frac{d(\chi_{i'}, \chi)}{h} \right) dP_\chi(\chi) \right)^{\alpha_{i'i}} \middle| \chi_i, \chi_{i'}, i \neq i' \right) \leq C \Phi_H^{\alpha_{i'i}}.$$

As the kernel is bounded, this integral is bounded for all other pairs. Thus, we have

$$\left| \mathbb{E} \left(\prod_{i,i'=1}^m \tilde{\delta}_{i'i}^{\alpha_{i'i}} \right) \right| \leq \frac{1}{C_{1,2}^{2k} \Phi_H^{2k}} C \Phi_H^{k+m/2} = \left(\frac{1}{\Phi_H} \right)^{(k-\frac{m}{2})},$$

which proves statement (B.2.18). We now consider statement (B.2.21). Using the Theorem of Fubini to permute the integrals, we have

$$\begin{aligned} & \left| \int \int \int \delta_{j\chi} \delta_{i\chi} W(\chi) dP_\chi(\chi) dP_\chi(\chi_i) dP_\chi(\chi_j) \right| \\ & = \left| \int \int \delta_{j\chi} dP_\chi(\chi_j) \int \delta_{i\chi} dP_\chi(\chi_i) W(\chi) dP_\chi(\chi) \right| \\ & = \mathbb{E}(\mathbb{E}(\delta_{j\chi}) \mathbb{E}(\delta_{i\chi}) W(\chi)). \end{aligned}$$

Then, using condition (2.2.25) and the fact that r is bounded, we have for $j = i, i'$,

$$\mathbb{E}(|\delta_{j\chi}|) \leq \mathbb{E} \left(\left| (Y_j - r(\chi)) \frac{K \left(\frac{d(\chi, \chi_j)}{h} \right)}{\mathbb{E} \left(K \left(\frac{d(\chi, \chi_j)}{h} \right) \right)} \right| \right)$$

$$\begin{aligned}
&\leq \mathbb{E} \left(\mathbb{E}(|Y_j| | \chi_j) \frac{K \left(\frac{d(\chi, \chi_j)}{h} \right)}{\mathbb{E} \left(K \left(\frac{d(\chi, \chi_j)}{h} \right) \right)} \right) + \mathbb{E} \left(C_r \frac{K \left(\frac{d(\chi, \chi_j)}{h} \right)}{\mathbb{E} \left(K \left(\frac{d(\chi, \chi_j)}{h} \right) \right)} \right) \\
&\leq (C_1 + C_r) \mathbb{E} \left(\frac{K \left(\frac{d(\chi, \chi_j)}{h} \right)}{\mathbb{E} \left(K \left(\frac{d(\chi, \chi_j)}{h} \right) \right)} \right) = C_1 + C_r.
\end{aligned}$$

Combining this result and the fact that W is bounded, we have

$$\mathbb{E} \left(\mathbb{E}(|\delta_{j\chi}|) \mathbb{E}(|\delta_{i\chi}|) W(\chi) \right) \leq C,$$

which proves statement (B.2.21). In a view to prove statement (B.2.22), we note that

$$\begin{aligned}
(Y_i - r(\chi))^2 &= ((Y_i - r(\chi_i) + r(\chi_i) - r(\chi))^2 = (\epsilon_i + r(\chi_i) - r(\chi))^2 \\
&= \epsilon_i^2 + 2\epsilon_i(r(\chi_i) - r(\chi)) + (r(\chi_i) - r(\chi))^2,
\end{aligned}$$

and

$$\mathbb{E}((Y_i - r(\chi))^2 | \chi_i) \geq \mathbb{E}(\epsilon_i^2 | \chi_i)$$

because of condition (2.2.21). Then, using conditions (2.2.22) and (2.2.16) successively, we have

$$\begin{aligned}
\int \tilde{\delta}_{ii} dP_\chi(\chi_i) &= \int \int \delta_{i\chi} \delta_{i\chi} W(\chi) dP_\chi(\chi) dP_\chi(i) \\
&\geq C \mathbb{E} \left((Y_i - r(\chi))^2 \frac{K^2 \left(\frac{d(\chi, \chi_i)}{h} \right)}{\mathbb{E}^2 \left(K \left(\frac{d(\chi, \chi_i)}{h} \right) \right)} \right) \\
&\geq C \mathbb{E} \left(\mathbb{E}(\epsilon_i^2 | \chi_i) \frac{K^2 \left(\frac{d(\chi, \chi_i)}{h} \right)}{\mathbb{E}^2 \left(K \left(\frac{d(\chi, \chi_i)}{h} \right) \right)} \right) \\
&\geq C \sigma_0^2 \mathbb{E} \left(\frac{K^2 \left(\frac{d(\chi, \chi_i)}{h} \right)}{\mathbb{E}^2 \left(K \left(\frac{d(\chi, \chi_i)}{h} \right) \right)} \right) \geq \frac{C}{\Phi_H}.
\end{aligned}$$

Finally, statement (B.2.23) is shown true by taking $C_k = C^{2k-2}$.

□

Lemma 8. *Under conditions of Theorem 2, we have*

$$\sup_{H \in \mathcal{H}_n} \left| \frac{MISE^*(H) - ISE^*(H)}{MISE^*(H)} \right| \longrightarrow 0 \quad a.s.,$$

with

$$ISE^*(H) = \frac{1}{n^2} \int \sum_{j,k=1}^n \delta_{jx} \delta_{kx} W(\chi) dP_\chi(\chi). \quad (\text{B.2.27})$$

Proof. The proof relies on the following decomposition:

$$ISE^*(H) = R_1(H) + R_2(H) + R_3(H) + 2\left(1 - \frac{1}{n}\right)S(H) + b_H\left(1 - \frac{1}{n}\right),$$

with $R_3(H)$ defined by (B.2.3) and

$$\begin{aligned} R_1(H) &= \frac{1}{n^2} \int \sum_{\substack{i,j=1 \\ i \neq j}}^n \left(\delta_{ix} - \mathbb{E}(\delta_{ix}) \right) \left(\delta_{jx} - \mathbb{E}(\delta_{jx}) \right) W(\chi) dP_\chi(\chi), \\ R_2(H) &= \frac{1}{n^2} \int \sum_{i=1}^n \left(\delta_{ix}^2 - \mathbb{E}(\delta_{ix}^2) \right) W(\chi) dP_\chi(\chi), \end{aligned} \quad (\text{B.2.28})$$

$$S(H) = \frac{1}{n} \int \mathbb{E}(\delta_{jx}) \sum_{i=1}^n \left(\delta_{ix} - \mathbb{E}(\delta_{ix}) \right) W(\chi) dP_\chi(\chi) \quad (\text{B.2.29})$$

Those definitions meet the one of Marron and Härdle (1986). Then, because of this decomposition, and because of decomposition (B.2.1), we have

$$\begin{aligned} \sup_{H \in \mathcal{H}_n} \left| \frac{MISE^*(H) - ISE^*(H)}{MISE^*(H)} \right| &= \sup_{H \in \mathcal{H}_n} \left| \frac{R_1(H) + R_2(H) + 2\left(1 - \frac{1}{n}\right)S(H)}{MISE^*(H)} \right| \\ &\leq \sup_{H \in \mathcal{H}_n} \left| \frac{R_1(H)}{MISE^*(H)} \right| + \sup_{H \in \mathcal{H}_n} \left| \frac{R_2(H)}{MISE^*(H)} \right| \\ &\quad + 2 \frac{n-1}{n} \sup_{H \in \mathcal{H}_n} \left| \frac{S(H)}{MISE^*(H)} \right|. \end{aligned}$$

We then use exactly the same steps as Marron and Härdle (1986, p.105), with conditions (3.1) and (3.3) in that paper replaced by conditions (B.2.14), (2.2.24), and (2.2.17) and (2.2.18) respectively, and where the conditions equivalent to (3.4) to (3.7) in that paper are shown valid by Lemma 7. This allows to show that for all

$k = 1, 2, \dots$, we have

$$\begin{aligned} \mathbb{E} \left(\left(\frac{R_1(H)}{MISE^*(H)} \right)^{2k} \right) &\leq C_k n^{-\gamma k} \\ \mathbb{E} \left(\left(\frac{R_2(H)}{MISE^*(H)} \right)^{2k} \right) &\leq C_k n^{-\gamma k} \\ \mathbb{E} \left(\left(\frac{S(H)}{MISE^*(H)} \right)^{2k} \right) &\leq C_k n^{-\gamma k}. \end{aligned} \quad (\text{B.2.30})$$

Then, given this, given condition (2.2.24), and using Markov inequality, we have

$$\begin{aligned} &P \left(\sup_{H \in \mathcal{H}_n} \left| \frac{MISE^*(H) - ISE^*(H)}{MISE^*(H)} \right| > \epsilon \right) \\ &\leq \#\mathcal{H}_n \sup_{H \in \mathcal{H}_n} P \left(\left| \frac{MISE^*(H) - ISE^*(H)}{MISE^*(H)} \right| > \epsilon \right) \\ &\leq C_n n^\alpha \frac{1}{\epsilon^{2k}} \sup_{H \in \mathcal{H}_n} \mathbb{E} \left(\left(\frac{MISE^*(H) - ISE^*(H)}{MISE^*(H)} \right)^{2k} \right). \\ &\leq C_n n^\alpha \frac{1}{\epsilon^{2k}} C_k n^{-\gamma k} = C n^{-\gamma k + \alpha} \end{aligned}$$

As $\alpha, \gamma > 0$, the validity of equation (B.2.31) for all $k = 1, 2, \dots$ tells us that, for all $\epsilon > 0$, one can always chose a k such that $\gamma k - \alpha > 1$ and a N large enough so that for all $n > N$, the above probability is smaller than ϵ . This indicates the almost sure convergence of $\sup_{H \in \mathcal{H}_n} \left| \frac{MISE^*(H) - ISE^*(H)}{MISE^*(H)} \right|$, which proves the Lemma. $\square$

Lemma 9. *Under conditions of Theorem 2, we have*

$$\sup_{H \in \mathcal{H}_n} \left| \frac{MISE^*(H) - ASE^*(H)}{MISE^*(H)} \right| \longrightarrow 0 \quad a.s.,$$

with

$$ASE^*(H) = \frac{1}{n} \sum_{i=1}^n (\hat{r}_{2H}(\chi_i) - r(\chi_i) \hat{r}_{1H}(\chi_i))^2 W(\chi_i) = \frac{1}{n^3} \sum_{i,j,k=1}^n \delta_{ji} \delta_{ki} W(\chi_i). \quad (\text{B.2.31})$$

Proof. The proof of this Lemma follows the same steps as Theorem 4 in Marron and Härdle (1986). First, long but simple calculations lead to the following decomposition of

$ASE^*(H)$:

$$\begin{aligned}
ASE^*(H) &= \frac{n-2}{n} ISE^*(H) \\
&+ T_1(H) + T_2(H) + \frac{n-1}{n} T_3(H) + 2T_4(H) \\
&+ 2T_5(H) + T_6(H) + T_7(H) \\
&+ 2\frac{n-2}{n} U_1(H) + 2\frac{n-1}{n} (U_2(H) + U_3(H)) \\
&+ \frac{(n-2)(n-1)}{n^2} V(H) \\
&+ \frac{1}{n} R_2(H) + \frac{1}{n} R_3(H), \tag{B.2.32}
\end{aligned}$$

with $ISE^*(H)$, $R_2(H)$ and $R_3(H)$ defined by equations (B.2.27), (B.2.28) and (B.2.3), and with the following definitions:

$$\begin{aligned}
T_1(H) &= \frac{1}{n^3} \sum_{\substack{i,j,k=1 \\ i \neq j \neq k \neq i}}^n (\delta_{ij}\delta_{ik}W(\chi_i) + \mathbb{E}(\delta_{ij}\delta_{ik}W(\chi_i)) \\
&\quad - \mathbb{E}(\delta_{ij}\delta_{ik}W(\chi_i)|\chi_j, \chi_k) - \mathbb{E}(\delta_{ij}\delta_{ik}W(\chi_i)|\chi_i, \chi_k) \\
&\quad - \mathbb{E}(\delta_{ij}\delta_{ik}W(\chi_i)|\chi_i, \chi_j) + \mathbb{E}(\delta_{ij}\delta_{ik}W(\chi_i)|\chi_k) \\
&\quad + \mathbb{E}(\delta_{ij}\delta_{ik}W(\chi_i)|\chi_j) - \mathbb{E}(\delta_{ij}\delta_{ik}W(\chi_i)|\chi_i)), \\
T_2(H) &= \frac{1}{n^3} \sum_{\substack{i,j=1 \\ i \neq j}}^n (\delta_{ij}^2 W(\chi_i) - \mathbb{E}(\delta_{ij}^2 W(\chi_i)|\chi_i) \\
&\quad - \mathbb{E}(\delta_{ij}^2 W(\chi_i)|\chi_j) + \mathbb{E}(\delta_{ij}^2 W(\chi_i))), \\
T_3(H) &= \frac{1}{n^2} \sum_{i=1}^n (\mathbb{E}(\delta_{ii}^2 W(\chi_i)|\chi_i) - \mathbb{E}(\delta_{ii}^2 W(\chi_i))), \\
T_4(H) &= \frac{1}{n^3} \sum_{\substack{i,j=1 \\ i \neq j}}^n (\delta_{ii}\delta_{ij}W(\chi_i) - \mathbb{E}(\delta_{ii}\delta_{ij}W(\chi_i)|\chi_i) \\
&\quad - \mathbb{E}(\delta_{ii}\delta_{ij}W(\chi_i)|\chi_j) + \mathbb{E}(\delta_{ii}\delta_{ij}W(\chi_i))), \\
T_5(H) &= \frac{1}{n^3} \sum_{\substack{i,j=1 \\ i \neq j}}^n (\mathbb{E}(\delta_{ii}\delta_{ij}W(\chi_i)|\chi_j) - \mathbb{E}(\delta_{ii}\delta_{ij}W(\chi_i))), \\
T_6(H) &= \frac{1}{n^3} \sum_{i=1}^n (\delta_{ii}^2 W(\chi_i) - \mathbb{E}(\delta_{ii}^2 W(\chi_i))),
\end{aligned}$$

$$\begin{aligned}
T_7(H) &= \frac{1}{n^2} \mathbb{E}(\delta_{ii}^2 W(\chi_i)), \\
U_1(H) &= \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^n (\delta_{ij} W(\chi_i) B(\chi_i) - \mathbb{E}(B(\chi_i) W(\chi_i) \delta_{ij} | \chi_j)), \\
U_2(H) + U_3(H) &= \frac{1}{n^2} \sum_{i=1}^n \int \delta_{ii} \delta_{ix} W(\chi_i) dP_\chi(\chi), \\
V(H) &= \frac{1}{n} \sum_{i=1}^n (W(\chi_i) B^2(\chi_i) - \mathbb{E}(B^2(\chi_i) W(\chi_i))).
\end{aligned}$$

Those terms are the same as the ones identified by the same name in Marron and Härdle (1986). As a consequence of this decomposition, we have

$$\begin{aligned}
& \frac{ASE^*(H) - MISE^*(H)}{MISE^*(H)} \tag{B.2.33} \\
&= \left(\frac{\frac{n-2}{n} ISE^*(H) + \frac{1}{n} R_3(H) - MISE^*(H)}{MISE^*(H)} \right) \\
&+ \left(\frac{T_1(H)}{MISE^*(H)} \right) + \left(\frac{T_2(H)}{MISE^*(H)} \right) \\
&+ \frac{n-1}{n} \left(\frac{T_3(H)}{MISE^*(H)} \right) + 2 \left(\frac{T_4(H)}{MISE^*(H)} \right) \\
&+ 2 \left(\frac{T_5(H)}{MISE^*(H)} \right) + \left(\frac{T_6(H)}{MISE^*(H)} \right) + \left(\frac{T_7(H)}{MISE^*(H)} \right) \\
&+ \frac{2(n-2)}{n} \left(\frac{U_1(H)}{MISE^*(H)} \right) + 2 \frac{n-1}{n} \left(\frac{U_2(H) + U_3(H)}{MISE^*(H)} \right) \\
&+ \frac{(n-2)(n-1)}{n^2} \left(\frac{V(H)}{MISE^*(H)} \right) \\
&+ \frac{1}{n} \left(\frac{R_2(H)}{MISE^*(H)} \right)
\end{aligned}$$

In this expression, the eleven last terms on the right hand side are shown to converge to 0 almost surely in exactly the same manner as in Marron and Härdle (1986), with conditions (3.1) and (3.3) in that paper replaced by conditions (B.2.14), (2.2.24), and (2.2.17) and (2.2.18) respectively, and where the conditions equivalent to (3.4) to (3.7) in that paper are shown valid by Lemma 7. Now, going back to equation (B.2.33), it

remains to prove that

$$\frac{\frac{n-2}{n}ISE^*(H) + \frac{1}{n}R_3(H) - MISE^*(H)}{MISE^*(H)} \longrightarrow 0 \quad a.s. \quad (B.2.34)$$

Therefore, we consider the following decomposition:

$$\begin{aligned} & \frac{\frac{n-2}{n}ISE^*(H) + \frac{1}{n}R_3(H) - MISE^*(H)}{MISE^*(H)} \\ &= \frac{n-2}{n} \frac{ISE^*(H) - MISE^*(H)}{MISE^*(H)} \\ &+ \frac{1}{n} \frac{R_3(H) - MISE^*(H)}{MISE^*(H)} + \frac{1}{n}. \end{aligned}$$

Using (B.2.1) and Lemma 5, we observe that

$$\left| \frac{1}{n} \frac{R_3(H) - MISE^*(H)}{MISE^*(H)} \right| = \left| \frac{n-1}{n^2} \frac{b_H}{MISE^*(H)} \right| \leq \frac{1}{n}$$

which converges to 0 as n goes to ∞ . We then make use of Lemma 8, and equation (B.2.34) is proved, which proves the Lemma. $\square$

Lemma 10. *Under conditions of Theorem 2, we have*

$$\sup_{H \in \mathcal{H}_n} \left| \frac{MISE^*(H) - ASE(H)}{MISE^*(H)} \right| \longrightarrow 0 \quad a.s.$$

Proof. We first state that

$$\sup_{H \in \mathcal{H}_n} \left| \frac{ASE^*(H) - ASE(H)}{ASE^*(H)} \right| \longrightarrow 0 \quad a.s., \quad (B.2.35)$$

with $ASE^*(H)$ defined by (B.2.31). We have, for all $H \in \mathcal{H}_n$,

$$\begin{aligned} ASE(H) &= \frac{1}{n} \sum_{i=1}^n (\hat{r}_H(\chi_i) - r(\chi_i))^2 W(\chi_i) \\ &= \frac{1}{n} \sum_{i=1}^n \left(\frac{\hat{r}_{2H}(\chi_i) - r(\chi_i)\hat{r}_{1H}(\chi_i)}{\hat{r}_{1H}(\chi_i)} \right)^2 W(\chi_i) \\ &= \frac{1}{n} \sum_{i=1}^n \left(\hat{r}_{2H}(\chi_i) - r(\chi_i)\hat{r}_{1H}(\chi_i) \right)^2 W(\chi_i) \left(1 + \frac{1 - \hat{r}_{1H}(\chi_i)}{\hat{r}_{1H}(\chi_i)} \right)^2. \end{aligned}$$

Expression (B.2.35) follows from this decomposition, because of Lemma 6. Given this, we can write

$$\begin{aligned}
& \sup_{H \in \mathcal{H}_n} \left| \frac{MISE^*(H) - ASE(H)}{MISE^*(H)} \right| \\
& \leq \sup_{H \in \mathcal{H}_n} \left| \frac{MISE^*(H) - ASE^*(H)}{MISE^*(H)} \right| + \sup_{H \in \mathcal{H}_n} \left| \frac{ASE^*(H) - ASE(H)}{MISE^*(H)} \right| \\
& = \sup_{H \in \mathcal{H}_n} \left| \frac{MISE^*(H) - ASE^*(H)}{MISE^*(H)} \right| \\
& + \sup_{H \in \mathcal{H}_n} \left(\left| \frac{ASE^*(H) - ASE(H)}{ASE^*(H)} \right| \left| \frac{ASE^*(H) - MISE^*(H)}{MISE^*(H)} + 1 \right| \right)
\end{aligned}$$

The proof is then completed because of (B.2.35) and of Lemma 9. $\square$

Lemma 11. *Under conditions of Theorem 2, we have*

$$\sup_{H \in \mathcal{H}_n} \left| \frac{\widetilde{ASE}(H) - ASE(H)}{MISE^*(H)} \right| \longrightarrow 0 \quad a.s.$$

Proof. For all $H \in \mathcal{H}_n$, we consider the result (B.2.35) for $ASE^*(H)$, and the following decomposition of $\widetilde{ASE}(H)$:

$$\begin{aligned}
\widetilde{ASE}(H) &= \frac{1}{n} \sum_{j=1}^n \left(\hat{r}_H^{-j}(\chi_j) - r(\chi_j) \right)^2 W(\chi_j), \\
&= \frac{1}{n} \sum_{j=1}^n \left(\frac{\hat{r}_{2H}^{-j}(\chi_j) - r(\chi_j) \hat{r}_{1H}^{-j}(\chi_j)}{\hat{r}_{1H}^{-j}(\chi_j)} \right)^2 W(\chi_j) \\
&= \frac{1}{n} \sum_{j=1}^n \left(\hat{r}_{2H}^{-j}(\chi_j) - r(\chi_j) \hat{r}_{1H}^{-j}(\chi_j) \right)^2 W(\chi_j) \left(1 + \frac{1 - \hat{r}_{1H}^{-j}(\chi_j)}{\hat{r}_{1H}^{-j}(\chi_j)} \right)^2.
\end{aligned}$$

By the same arguments as in (B.2.35), it results in

$$\sup_{H \in \mathcal{H}_n} \left| \frac{\widetilde{ASE}(H) - \widetilde{ASE}^*(H)}{\widetilde{ASE}^*(H)} \right| \longrightarrow 0 \quad a.s. \tag{B.2.36}$$

with

$$\widetilde{ASE}^*(H) = \frac{1}{n} \sum_{j=1}^n \left(\hat{r}_{2H}^{-j}(\chi_j) - r(\chi_j) \hat{r}_{1H}^{-j}(\chi_j) \right)^2 W(\chi_j).$$

Consequently, we have

$$\begin{aligned}
& \sup_{H \in \mathcal{H}_n} \left| \frac{\widetilde{ASE}(H) - ASE(H)}{MISE^*(H)} \right| \\
& \leq \sup_{H \in \mathcal{H}_n} \left| \frac{\widetilde{ASE}^*(H) - \widetilde{ASE}(H)}{MISE^*(H)} \right| \\
& + \sup_{H \in \mathcal{H}_n} \left| \frac{\widetilde{ASE}^*(H) - ASE(H)}{MISE^*(H)} \right| \\
& \leq \sup_{H \in \mathcal{H}_n} \left(\left| \frac{\widetilde{ASE}^*(H) - \widetilde{ASE}(H)}{\widetilde{ASE}^*(H)} \right| \left(\left| \frac{\widetilde{ASE}^*(H) - MISE^*(H)}{MISE^*(H)} \right| + 1 \right) \right) \\
& + \sup_{H \in \mathcal{H}_n} \left| \frac{\widetilde{ASE}^*(H) - ASE^*(H)}{MISE^*(H)} \right| \\
& + \sup_{H \in \mathcal{H}_n} \left(\left| \frac{ASE^*(H) - ASE(H)}{ASE^*(H)} \right| \left(\left| \frac{ASE^*(H) - MISE^*(H)}{MISE^*(H)} \right| + 1 \right) \right) \\
& \leq \sup_{H \in \mathcal{H}_n} \left(\left| \frac{\widetilde{ASE}^*(H) - \widetilde{ASE}(H)}{\widetilde{ASE}^*(H)} \right| \right. \\
& \quad \cdot \left(\left| \frac{\widetilde{ASE}^*(H) - ASE^*(H)}{MISE^*(H)} \right| + \left| \frac{ASE^*(H) - ASE(H)}{MISE^*(H)} \right| + 1 \right) \Bigg) \\
& + \sup_{H \in \mathcal{H}_n} \left| \frac{\widetilde{ASE}^*(H) - ASE^*(H)}{MISE^*(H)} \right| \\
& + \sup_{H \in \mathcal{H}_n} \left(\left| \frac{ASE^*(H) - ASE(H)}{ASE^*(H)} \right| \left(\left| \frac{ASE^*(H) - MISE^*(H)}{MISE^*(H)} \right| + 1 \right) \right) \\
& \leq \sup_{H \in \mathcal{H}_n} \left(\left| \frac{\widetilde{ASE}^*(H) - \widetilde{ASE}(H)}{\widetilde{ASE}^*(H)} \right| \left(\left| \frac{\widetilde{ASE}^*(H) - ASE^*(H)}{MISE^*(H)} \right| + 1 \right) \right) \\
& + \sup_{H \in \mathcal{H}_n} \left(\left| \frac{\widetilde{ASE}^*(H) - \widetilde{ASE}(H)}{\widetilde{ASE}^*(H)} \right| \left(\left| \frac{ASE^*(H) - MISE^*(H)}{MISE^*(H)} \right| \right) \right) \\
& + \sup_{H \in \mathcal{H}_n} \left| \frac{\widetilde{ASE}^*(H) - ASE^*(H)}{MISE^*(H)} \right| \\
& + \sup_{H \in \mathcal{H}_n} \left(\left| \frac{ASE^*(H) - ASE(H)}{ASE^*(H)} \right| \left(\left| \frac{ASE^*(H) - MISE^*(H)}{MISE^*(H)} \right| + 1 \right) \right)
\end{aligned}$$

Thus, the Lemma will be proved by using (B.2.35), (B.2.36) and Lemma 9, plus the fact that

$$\sup_{H \in \mathcal{H}_n} \left| \frac{\widetilde{ASE}^*(H) - ASE^*(H)}{MISE^*(H)} \right| \rightarrow 0 \quad a.s.$$

This last result is shown as follows. Because of decomposition (B.2.32) and of the following decomposition

$$\begin{aligned} \widetilde{ASE}^*(H) &= \frac{n-2}{n} ISE^*(H) + T_1(H) + T_2(H) + \frac{n-1}{n} T_3(H) \\ &\quad + \frac{1}{n} (R_2(H) + R_3(H)) + 2 \frac{n-2}{n} U_1(H) + \frac{(n-2)(n-1)}{n^2} V(H), \end{aligned}$$

we have, for all $H \in \mathcal{H}_n$,

$$\begin{aligned} \left| \frac{\widetilde{ASE}^*(H) - ASE^*(H)}{MISE^*(H)} \right| &\leq 2 \left| \frac{T_4(H)}{MISE^*(H)} \right| + 2 \left| \frac{T_5(H)}{MISE^*(H)} \right| + \left| \frac{T_6(H)}{MISE^*(H)} \right| \\ &\quad + \left| \frac{T_7(H)}{MISE^*(H)} \right| + 2 \frac{n-1}{n} \left| \frac{U_2(H) + U_3(H)}{MISE^*(H)} \right|. \end{aligned}$$

As discussed for equation (B.2.33) that contains the same terms, all the terms on the right hand side converge to zero almost surely. This is enough to proof Lemma 11. $\square$

Lemma 12. *Under conditions of Theorem 2, we have*

$$\sup_{H \in \mathcal{H}_n} \left| \frac{CT(H)}{MISE^*(H)} \right| \rightarrow 0 \quad a.s.$$

Proof. Using the definitions of $CT(H)$, $\hat{r}_{1H}(\chi)$ and $\hat{r}_{2H}(\chi)$, one gets:

$$\begin{aligned} &\sup_{H \in \mathcal{H}_n} \frac{|CT(H)|}{MISE^*(H)} \\ &\leq \sup_{H \in \mathcal{H}_n} \left| (MISE^*(H))^{-1} \frac{1}{n} \sum_{j=1}^n \epsilon_j (\hat{r}^{-j}(\chi_j) - r(\chi_j)) W(\chi_j) \right| \end{aligned}$$

$$\begin{aligned}
&\leq \sup_{H \in \mathcal{H}_n} \left| (MISE^*(H))^{-1} \frac{1}{n} \sum_{j=1}^n \epsilon_j \frac{(\hat{r}_{2H}^{-j}(\chi_j) - r(\chi_j) r_{1H}^{-j}(\chi_j))}{r_{1H}^{-j}(\chi_j)} W(\chi_j) \right| \\
&\leq \sup_{H \in \mathcal{H}_n} \left| (MISE^*(H))^{-1} \frac{1}{n} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \epsilon_j \epsilon_i \Delta_i(\chi_j) \frac{W(\chi_j)}{(n-1) \mathbb{E}(\Delta_i(\chi_j)) r_{1H}^{-j}(\chi_j)} \right| \\
&\quad + \sup_{H \in \mathcal{H}_n} \left| (MISE^*(H))^{-1} \frac{1}{n} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \epsilon_j (r(\chi_i) - r(\chi_j)) \Delta_i(\chi_j) \frac{W(\chi_j)}{(n-1) \mathbb{E}(\Delta_i(\chi_j)) r_{1H}^{-j}(\chi_j)} \right|
\end{aligned}$$

We note hereafter

$$U_{i,j} = (MISE^*(H))^{-1} \epsilon_j \epsilon_i \Delta_i(\chi_j) \frac{W(\chi_j)}{(n-1) \mathbb{E}(\Delta_i(\chi_j)) r_{1H}^{-j}(\chi_j)}$$

and

$$V_{i,j} = (MISE^*(H))^{-1} \epsilon_j (r(\chi_i) - r(\chi_j)) \Delta_i(\chi_j) \frac{W(\chi_j)}{(n-1) \mathbb{E}(\Delta_i(\chi_j)) r_{1H}^{-j}(\chi_j)}$$

Now the aim is to show

$$\sup_{H \in \mathcal{H}_n} \left| \frac{1}{n} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n U_{i,j} \right| = o_{a.s.}(1) \quad (\text{B.2.37})$$

and

$$\sup_{H \in \mathcal{H}_n} \left| \frac{1}{n} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n V_{i,j} \right| = o_{a.s.}(1) \quad (\text{B.2.38})$$

To state (B.2.37) and (B.2.38) we are going to show that for any $\epsilon > 0$ the series

$$\sum_{n=1}^{\infty} P\left(\sup_{H \in \mathcal{H}_n} \left| \frac{1}{n} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n U_{i,j} \right| > \epsilon \right) \quad \text{and} \quad \sum_{n=1}^{\infty} P\left(\sup_{H \in \mathcal{H}_n} \left| \frac{1}{n} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n V_{i,j} \right| > \epsilon \right)$$

are convergent. We consider the set

$$A_n = \left\{ \inf_{H \in \mathcal{H}_n, \chi \in \mathcal{W}, 1 \leq j \leq n} r_{1H}^{-j}(\chi) < \frac{1}{2} \right\}.$$

From results obtained in the proof of Lemma 6 it is fairly easy to get for any $s > 0$:

$$\begin{aligned}
 P(A_n) &\leq n \# \mathcal{H}_n \sup_{H \in \mathcal{H}_n, 1 \leq j \leq n} P \left(\inf_{\chi \in \mathcal{W}} r_{1H}^{-j}(\chi) < \frac{1}{2} \right) \\
 &\leq C n^{\alpha+1} \sup_{H \in \mathcal{H}_n, 1 \leq j \leq n} P \left(\sup_{\chi \in \mathcal{W}} |r_{1H}^{-j}(\chi) - 1| > \frac{1}{2} \right) \\
 &\leq C_s n^{-1-s} \text{ for } n \text{ large enough (i.e. } n \geq n_s \text{)}. \tag{B.2.39}
 \end{aligned}$$

On the other hand, for any $\epsilon > 0$ and any positive integer k , it comes from Markov inequality and condition (2.2.24),

$$\begin{aligned}
 &P \left(\sup_{H \in \mathcal{H}_n} \left| \frac{1}{n} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n U_{i,j} \right| > \epsilon \cap \overline{A_n} \right) \\
 &\leq \# \mathcal{H}_n \sup_{H \in \mathcal{H}_n} P \left(\left| \frac{1}{n} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n U_{i,j} \mathbf{1}_{\overline{A_n}} \right| > \epsilon \right) \\
 &\leq \epsilon^{-2k} \# \mathcal{H}_n \sup_{H \in \mathcal{H}_n} \mathbb{E} \left(\left| \frac{1}{n} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n U_{i,j} \mathbf{1}_{\overline{A_n}} \right|^{2k} \right) \tag{B.2.40} \\
 &\leq C n^\alpha \sup_{H \in \mathcal{H}_n} \mathbb{E} \left(\left| \frac{1}{n} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n U_{i,j} \mathbf{1}_{\overline{A_n}} \right|^{2k} \right)
 \end{aligned}$$

Hence, it is enough to show that for some positive integer k

$$\sum_{n=1}^{\infty} n^\alpha \sup_{H \in \mathcal{H}_n} \mathbb{E} \left(\left| \frac{1}{n} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n U_{i,j} \mathbf{1}_{\overline{A_n}} \right|^{2k} \right) < +\infty \tag{B.2.41}$$

Now, we define

$$I_{2k} = \{(i_1, \dots, i_{2k}, j_1, \dots, j_{2k}) \in \{1, \dots, n\}^{2k} | i_l \neq j_l \forall 1 \leq l \leq 2k\}.$$

It comes

$$\mathbb{E} \left(\frac{1}{n^{2k}} \left(\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n U_{i,j} \right)^{2k} \mathbf{1}_{\overline{A_n}} \right)$$

$$\begin{aligned}
&= n^{-2k} \sum_{(i_l, j_l)_{1 \leq l \leq 2k} \in I_{2k}} \left| \mathbb{E} \left(\Pi_{l=1}^{2k} U_{i_l, j_l} \mathbf{1}_{\bar{A}_n} \right) \right| \\
&\leq n^{-2k} \sum_{m=2}^{4k} \sum_{(i_l, j_l)_{1 \leq l \leq 2k} \in J_m} \left| \mathbb{E} \left(\Pi_{l=1}^{2k} (MISE^*(H))^{-1} \epsilon_{j_l} \epsilon_{i_l} \Delta_{i_l}(\chi_{j_l}) \right. \right. \\
&\quad \left. \left. \cdot \frac{W(\chi_{j_l})}{(n-1) \mathbb{E}(\Delta_{i_l}(\chi_{j_l})) r_{1H}^{-j_l}(\chi_{j_l})} \mathbf{1}_{\bar{A}_n} \right) \right| \\
&\leq n^{-2k} \sum_{m=2}^{2k} \sum_{(i_l, j_l)_{1 \leq l \leq 2k} \in J_m} \left| \mathbb{E} \left(\Pi_{l=1}^{2k} (MISE^*(H))^{-1} \right. \right. \\
&\quad \left. \left. \frac{\epsilon_{j_l} \epsilon_{i_l} \Delta_{i_l}(\chi_{j_l}) W(\chi_{j_l})}{(n-1) \mathbb{E}(\Delta_{i_l}(\chi_{j_l})) r_{1H}^{-j_l}(\chi_{j_l})} \mathbf{1}_{\bar{A}_n} \right) \right|, \tag{B.2.42}
\end{aligned}$$

where J_m contains elements of I_{2k} that involve exactly m different indices. The last line comes from the fact that when $m > 2k$ at least one of the ϵ 's appears with exponent 1 and hence the mean equals 0 (conditioning w.r.t. $\chi_1, \dots, \chi_n$). Using the fact that W is bounded, using conditions (2.2.16) and (2.2.25), it comes directly from (B.2.42)

$$\begin{aligned}
&\mathbb{E} \left(\frac{1}{n^{2k}} \left(\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n U_{i,j} \right)^{2k} \mathbf{1}_{\bar{A}_n} \right) \\
&\leq C'_k n^{-2k} (MISE^*(H) n \Phi_H)^{-2k} \sum_{m=2}^{2k} \sum_{(i_l, j_l)_{1 \leq l \leq 2k} \in J_m} \mathbb{E} \left(\Pi_{l=1}^{2k} \Delta_{i_l}(\chi_{j_l}) W(\chi_{j_l}) \right) \\
&\leq C'_k n^{-2k} \sum_{m=2}^{2k} \#J_m \Phi_H^{\frac{m}{2}} \\
&\leq C'_k \Phi_H^k \sum_{m=2}^{2k} (n \Phi_H)^{m-2k} \Phi_H^{k-\frac{m}{2}} \\
&\leq C'_k \Phi_H^k \tag{B.2.43}
\end{aligned}$$

where the second inequality comes from the fact that one can extract at least $\frac{m}{2}$ pairs (i_{l_p}, j_{l_p}) in which for each p either i_{l_p} or j_{l_p} appears uniquely once. Now, (B.2.43) and (2.2.17) are enough to state (B.2.41) for k large enough. The proof of (B.2.38) is

very similar.

$$\begin{aligned} \mathbb{E} \left(\frac{1}{n^{2k}} \left(\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n V_{i,j} \right)^{2k} \mathbf{1}_{\overline{A}_n} \right) &\leq \\ n^{-2k} \sum_{m=2}^{3k} \sum_{(i_l, j_l)_{1 \leq l \leq 2k} \in J_m} &\left| \mathbb{E} \left(\prod_{l=1}^{2k} \frac{(r(\chi_{i_l}) - r(\chi_{j_l})) \epsilon_{i_l} \Delta_{i_l}(\chi_{j_l}) w(\chi_{j_l})}{(n-1) \text{MISE}^*(H) \mathbb{E}(\Delta(\chi_{j_l})) r_{1H}^{-j_l}(\chi_{j_l})} \mathbf{1}_{\overline{A}_n} \right) \right|, \end{aligned} \quad (\text{B.2.44})$$

The last line comes from the fact that when $m > 3k$ at least one of the ϵ 's appears with exponent 1 and hence the mean equals 0 (conditioning w.r.t. $\chi_1, \dots, \chi_n$). Now conditions (2.2.23), (2.2.17) and (2.2.18), as well as Lemma 5 are used to get

$$\begin{aligned} &\mathbb{E} \left(\frac{1}{n^{2k}} \left(\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n V_{i,j} \right)^{2k} \mathbf{1}_{\overline{A}_n} \right) \\ &\leq C'_k \sum_{m=2}^{3k} n^{m-2k} \left(\text{MISE}^*(H) (n\Phi_H)^2 \right)^{-k} \Phi_H^{\frac{m}{2}} \\ &\leq C'_k \Phi_H^{\frac{k}{2}} \sum_{m=2}^{3k} (n\Phi_H)^{m-3k} \Phi_H^{\frac{3k-m}{2}} \\ &\leq C'_k \Phi_H^{\frac{k}{2}}, \end{aligned}$$

what is enough to get (B.2.38). □

Lemma 13. *Under conditions of Theorem 2, we have*

$$\sup_{H \in \mathcal{H}_n} \left| \frac{\text{MISE}(H) - \text{MISE}^*(H)}{\text{MISE}^*(H)} \right| \longrightarrow 0 \quad a.s$$

Proof. We consider the following decomposition of $MISE(H)$:

$$\begin{aligned}
 MISE(H) &= \int \mathbb{E} \left(\frac{(\hat{r}_{2H}(\chi) - r(\chi)\hat{r}_{1H}(\chi))^2}{\hat{r}_{1H}(\chi)^2} \right) W(\chi) dP_\chi(\chi) \\
 &= \int \mathbb{E} \left((\hat{r}_{2H}(\chi) - r(\chi)\hat{r}_{1H}(\chi))^2 \right. \\
 &\quad \left. \cdot (1 + (1 - \hat{r}_{1H}(\chi)) + \frac{(1 - \hat{r}_{1H}(\chi))^2}{\hat{r}_{1H}(\chi)})^2 \right) W(\chi) dP_\chi(\chi) \\
 &= MISE^*(H) + A_1(H) + A_2(H) + 2A_3(H) + 2A_4(H) + 2A_5(H),
 \end{aligned}$$

with

$$\begin{aligned}
 A_1(H) &= \int \mathbb{E} \left((\hat{r}_{2H}(\chi) - r(\chi)\hat{r}_{1H}(\chi))^2 (1 - \hat{r}_{1H}(\chi))^2 \right) W(\chi) dP_\chi(\chi) \\
 A_2(H) &= \int \mathbb{E} \left((\hat{r}(\chi) - r(\chi))^2 (1 - \hat{r}_{1H}(\chi))^4 \right) W(\chi) dP_\chi(\chi) \\
 A_3(H) &= \int \mathbb{E} \left((\hat{r}_{2H}(\chi) - r(\chi)\hat{r}_{1H}(\chi))^2 (1 - \hat{r}_{1H}(\chi)) \right) W(\chi) dP_\chi(\chi) \\
 A_4(H) &= \int \mathbb{E} \left((\hat{r}_{2H}(\chi) - r(\chi)\hat{r}_{1H}(\chi)) (\hat{r}(\chi) - r(\chi)) (1 - \hat{r}_{1H}(\chi))^2 \right) W(\chi) dP_\chi(\chi) \\
 A_5(H) &= \int \mathbb{E} \left((\hat{r}_{2H}(\chi) - r(\chi)\hat{r}_{1H}(\chi)) (\hat{r}(\chi) - r(\chi)) (1 - \hat{r}_{1H}(\chi))^3 \right) W(\chi) dP_\chi(\chi).
 \end{aligned}$$

We consider the term $A_1(H)$. Long but simple calculations leads to the following decomposition:

$$A_1(H) = A_{11} + A_{12} + A_{13} + 2A_{14} + 2A_{15} + 2A_{16},$$

with

$$\begin{aligned}
 A_{11} &= \int \mathbb{E} \left((\hat{r}_{2H}(\chi) - E(\hat{r}_{2H}(\chi)))^2 (1 - \hat{r}_{1H}(\chi))^2 \right) W(\chi) dP_\chi(\chi) \\
 A_{12} &= \int B^2(\chi) \mathbb{E} \left((1 - \hat{r}_{1H}(\chi))^2 \right) W(\chi) dP_\chi(\chi) \\
 A_{13} &= \int r^2(\chi) \mathbb{E} \left((1 - \hat{r}_{1H}(\chi))^4 \right) W(\chi) dP_\chi(\chi)
 \end{aligned}$$

$$\begin{aligned}
A_{14} &= \int B(\chi) \mathbb{E} \left((\hat{r}_{2H}(\chi) - \mathbb{E}(\hat{r}_{2H}(\chi))) (1 - \hat{r}_{1H}(\chi))^2 \right) W(\chi) dP_\chi(\chi) \\
A_{15} &= \int r(\chi) \mathbb{E} \left((\hat{r}_{2H}(\chi) - \mathbb{E}(\hat{r}_{2H}(\chi))) (1 - \hat{r}_{1H}(\chi))^3 \right) W(\chi) dP_\chi(\chi) \\
A_{16} &= \int B(\chi) r(\chi) \mathbb{E} \left((1 - \hat{r}_{1H}(\chi))^3 \right) W(\chi) dP_\chi(\chi).
\end{aligned}$$

For dealing with term A_{11} , we first observe that

$$\mathbb{E} \left(\left(\hat{r}_{2H}(\chi) - \mathbb{E}(\hat{r}_{2H}(\chi)) \right)^2 \left(1 - \hat{r}_{1H}(\chi) \right)^2 \right) \quad (\text{B.2.45})$$

$$\begin{aligned}
&= \frac{1}{n^4 \mathbb{E}^4(\Delta_1(\chi))} \left[\sum_{\substack{i,j=1 \\ i \neq j}}^n \mathbb{E} \left(\left(Y_i \Delta_i(\chi) - \mathbb{E}(Y_i \Delta_i(\chi)) \right)^2 \right) \right. \\
&\quad \cdot \mathbb{E} \left(\left(\Delta_j(\chi) - \mathbb{E}(\Delta_j(\chi)) \right)^2 \right) \quad (\text{B.2.46})
\end{aligned}$$

$$\begin{aligned}
&+ 2 \sum_{\substack{i,j=1 \\ i \neq j=1}}^n \mathbb{E} \left(\left(Y_i \Delta_i(\chi) - \mathbb{E}(Y_i \Delta_i(\chi)) \right) \left(\Delta_i(\chi) - \mathbb{E}(\Delta_i(\chi)) \right) \right. \\
&\quad \cdot \left(Y_j \Delta_j(\chi) - \mathbb{E}(Y_j \Delta_j(\chi)) \right) \left(\Delta_j(\chi) - \mathbb{E}(\Delta_j(\chi)) \right) \Big) \\
&+ \sum_{i=1}^n \mathbb{E} \left(\left(Y_i \Delta_i(\chi) - \mathbb{E}(Y_i \Delta_i(\chi)) \right)^2 \left(\Delta_i(\chi) - \mathbb{E}(\Delta_i(\chi)) \right)^2 \right) \Big] \\
&= \frac{1}{n^4 \mathbb{E}^4(\Delta_1(\chi))} \left[n(n-1) \text{Var}(Y_1 \Delta_1(\chi)) \text{Var}(\Delta_1(\chi)) \right. \\
&\quad + 2n(n-1) \mathbb{E}^2 \left(\left(Y_1 \Delta_1(\chi) - \mathbb{E}(Y_1 \Delta_1(\chi)) \right) \left(\Delta_1(\chi) - \mathbb{E}(\Delta_1(\chi)) \right) \right) \\
&\quad \left. + n \mathbb{E} \left(\left(Y_1 \Delta_1(\chi) - \mathbb{E}(Y_1 \Delta_1(\chi)) \right)^2 \left(\Delta_1(\chi) - \mathbb{E}(\Delta_1(\chi)) \right)^2 \right) \right],
\end{aligned}$$

which relies on condition (2.2.15). Because of condition (2.2.16), we have

$$\text{Var}(\Delta_1(\chi)) = \mathbb{E}(\Delta_1^2(\chi)) - \mathbb{E}^2(\Delta_1(\chi)) \leq \mathbb{E}(\Delta_1^2(\chi)) \leq C_{2,1} \Phi_H. \quad (\text{B.2.47})$$

Similarly, because of condition (2.2.26), we have

$$\begin{aligned} \text{Var}\left(Y_1 \Delta_1(\chi)\right) &= \mathbb{E}\left(Y_1^2 \Delta_1^2(\chi)\right) - \mathbb{E}^2\left(Y_1 \Delta_1(\chi)\right) \\ &\leq \mathbb{E}\left(Y_1^2 \Delta_1^2(\chi)\right) \leq C\Phi_H. \end{aligned} \quad (\text{B.2.48})$$

Because of conditions (2.2.16) and (2.2.26), and because $\Phi_H \leq 1$, we have

$$\begin{aligned} &\mathbb{E}^2\left(\left(Y_1 \Delta_1(\chi) - \mathbb{E}\left(Y_1 \Delta_1(\chi)\right)\right)\left(\Delta_1(\chi) - \mathbb{E}\left(\Delta_1(\chi)\right)\right)\right) \\ &= \left(2\mathbb{E}\left(Y_1 \Delta_1(\chi)\right)\mathbb{E}\left(\Delta_1(\chi)\right) - 2\mathbb{E}\left(Y_1 \Delta_1^2(\chi)\right)\right)^2 \\ &= 4\left(\mathbb{E}^2\left(Y_1 \Delta_1(\chi)\right)\mathbb{E}^2\left(\Delta_1(\chi)\right) + \mathbb{E}^2\left(Y_1 \Delta_1^2(\chi)\right) \right. \\ &\quad \left. - 2\mathbb{E}\left(Y_1 \Delta_1^2(\chi)\right)\mathbb{E}\left(Y_1 \Delta_1(\chi)\right)\mathbb{E}\left(\Delta_1(\chi)\right)\right) \\ &\leq C\left(\Phi_H^4 + \Phi_H^2 + \Phi_H^3\right) \leq C'\Phi_H^2. \end{aligned} \quad (\text{B.2.49})$$

Using the same conditions, we also have

$$\begin{aligned} &\mathbb{E}\left(\left(Y_1 \Delta_1(\chi) - \mathbb{E}\left(Y_1 \Delta_1(\chi)\right)\right)^2\left(\Delta_1(\chi) - \mathbb{E}\left(\Delta_1(\chi)\right)\right)^2\right) \\ &= \mathbb{E}\left(Y_1^2 \Delta_1^2(\chi)\right)\mathbb{E}^2\left(\Delta_1(\chi)\right) + \mathbb{E}\left(Y_1^2 \Delta_1^4(\chi)\right) - 2\mathbb{E}\left(Y_1^2 \Delta_1^3(\chi)\right)\mathbb{E}\left(\Delta_1(\chi)\right) \\ &\quad + \mathbb{E}^4\left(Y_1 \Delta_1(\chi)\right) + \mathbb{E}\left(\Delta_1^2(\chi)\right)\mathbb{E}^2\left(Y_1 \Delta_1(\chi)\right) \\ &\quad - 4\mathbb{E}^2\left(\Delta_1(\chi)\right)\mathbb{E}^2\left(Y_1 \Delta_1(\chi)\right) \\ &\quad - 2\mathbb{E}\left(Y_1 \Delta_1^3(\chi)\right)\mathbb{E}\left(Y_1 \Delta_1(\chi)\right) \\ &\quad + 4\mathbb{E}\left(Y_1 \Delta_1^2(\chi)\right)\mathbb{E}\left(\Delta_1(\chi)\right)\mathbb{E}\left(Y_1 \Delta_1(\chi)\right) \\ &\leq C(\Phi_H + \Phi_H^2 + \Phi_H^3 + \Phi_H^4) \leq C'\Phi_H. \end{aligned} \quad (\text{B.2.50})$$

Now, given equations (B.2.45), (B.2.47), (B.2.48), (B.2.49) and (B.2.50), and using condition (2.2.16) and the fact that W is bounded, we obtain

$$\begin{aligned} A_{11} &= \int \mathbb{E} \left((\hat{r}_{2H}(\chi) - E(\hat{r}_{2H}(\chi)))^2 (1 - \hat{r}_{1H}(\chi))^2 \right) W(\chi) dP_\chi(\chi) \\ &\leq \frac{1}{n^4 \Phi_H^4} (n(n-1)C\Phi_H^2 + 2n(n-1)C'\Phi_H^2 + nC''\Phi_H) = \mathcal{O} \left(\frac{1}{n^2 \Phi_H^2} \right). \end{aligned}$$

As a result, because of Lemma 5 and using condition (2.2.18), we have

$$\sup_{H \in \mathcal{H}_n} \frac{|A_{11}|}{MISE^*(H)} \leq \sup_{H \in \mathcal{H}_n} C \frac{n\Phi_H}{n^2 \Phi_H^2} = \sup_{H \in \mathcal{H}_n} C \frac{1}{n\Phi_H} = \mathcal{O} \left(\frac{1}{\inf_{H \in \mathcal{H}_n} n\Phi_H} \right) = o(1).$$

We then consider the term A_{12} . Using condition (2.2.15) and equation (B.2.47), we have

$$\begin{aligned} \mathbb{E} \left((1 - \hat{r}_{1H}(\chi))^2 \right) &= \frac{1}{n^2 \mathbb{E}^2(\Delta_1(\chi))} \sum_{i=1}^n \mathbb{E} \left(\mathbb{E}(\Delta_i(\chi)) - \Delta_i(\chi) \right)^2 \\ &= \frac{n \text{Var}(\Delta_1(\chi))}{n^2 \mathbb{E}^2(\Delta_1(\chi))} \leq \frac{C}{n\Phi_H} \end{aligned}$$

Consequently, and because W is bounded,

$$A_{12} \leq \frac{C}{n\Phi_H} \int B^2(\chi) W(\chi) dP_\chi(\chi) \leq \frac{C'}{n\Phi_H} b_H,$$

so that

$$\sup_{H \in \mathcal{H}_n} \frac{|A_{12}|}{MISE^*(H)} \leq C \frac{b_H}{n\Phi_H} \frac{n}{(n-1)b_H} = \mathcal{O} \left(\frac{1}{\inf_{H \in \mathcal{H}_n} n\Phi_H} \right) = o(1),$$

because of condition (2.2.18), and because of Lemma 5. In view of dealing with term A_{13} , and because of condition (2.2.15), we note that

$$\begin{aligned}
 & \mathbb{E}\left(1 - \hat{r}_{1H}(\chi)\right)^4 \tag{B.2.51} \\
 &= \frac{1}{n^4 \mathbb{E}^4(\Delta_1(\chi))} \left(3 \sum_{\substack{i,j=1 \\ i \neq j}}^n \mathbb{E} \left(\left(\mathbb{E}(\Delta_i(\chi)) - \Delta_i(\chi) \right)^2 \left(\mathbb{E}(\Delta_j(\chi)) - \Delta_j(\chi) \right)^2 \right) \right. \\
 & \quad \left. + \sum_{i=1}^n \mathbb{E} \left(\left(\mathbb{E}(\Delta_i(\chi)) - \Delta_i(\chi) \right)^4 \right) \right) \\
 &= \frac{3n(n-1) \text{Var}^2(\Delta_1(\chi))}{n^4 \mathbb{E}^4(\Delta_1(\chi))} + \frac{1}{n^3 \mathbb{E}^4(\Delta_1(\chi))} \mathbb{E} \left(\left(\mathbb{E}(\Delta_1(\chi)) - \Delta_1(\chi) \right)^4 \right).
 \end{aligned}$$

Because of condition (2.2.16), and because $\Phi_H < 1$, we have

$$\begin{aligned}
 & \mathbb{E} \left(\left(\mathbb{E}(\Delta_1(\chi)) - \Delta_1(\chi) \right)^4 \right) \\
 &= \mathbb{E}(\Delta_1^4(\chi)) - 4\mathbb{E}(\Delta_1(\chi)) \mathbb{E}(\Delta_1^3(\chi)) + 6\mathbb{E}^2(\Delta_1(\chi)) \mathbb{E}(\Delta_1^2(\chi)) \\
 & \quad - 4\mathbb{E}^4(\Delta_1(\chi)) + \mathbb{E}^4(\Delta_1(\chi)) \\
 &\leq C(\Phi + \Phi_H^2 + \Phi_H^3 + \Phi_H^4) \leq C' \Phi_H.
 \end{aligned}$$

Moreover, due to equation (B.2.47), condition (2.2.16), and because $\Phi_H < 1$, we have

$$\mathbb{E}\left(1 - \hat{r}_{1H}(\chi)\right)^4 \leq C \frac{1}{n^4 \Phi_H^4} n^2 \Phi_H^2 + C' \frac{1}{n^3 \Phi_H^4} \Phi_H \leq C \frac{1}{n^2 \Phi_H^2}$$

Thus, as r and W are bounded, we obtain

$$A_{13} \leq \frac{C}{n^2 \Phi_H^2} \int r^2(\chi) W(\chi) dP_\chi(\chi) \leq \frac{C'}{n^2 \Phi_H^2}$$

Consequently, using (2.2.18) and Lemma 5, we have

$$\sup_{H \in \mathcal{H}_n} \frac{|A_{13}|}{MISE^*(H)} \leq \sup_{H \in \mathcal{H}_n} C \left(\frac{1}{n^2 \Phi_H^2} + \frac{1}{n^3 \Phi_H^3} \right) C n \phi_H \leq \mathcal{O} \left(\frac{1}{\inf_{H \in \mathcal{H}_n} n \Phi_H} \right) = o(1).$$

Now, using Cauchy-Schwarz inequality, we show that

$$\begin{aligned} |A_{14}| &\leq \sqrt{|A_{11}|} \sqrt{|A_{12}|} \\ |A_{15}| &\leq \sqrt{|A_{11}|} \sqrt{|A_{13}|} \quad \text{and} \\ |A_{16}| &\leq \sqrt{|A_{12}|} \sqrt{|A_{13}|}. \end{aligned}$$

It results from this that

$$\begin{aligned} \sup_{H \in \mathcal{H}_n} \frac{|A_{14}|}{MISE^*(H)} &= o(1), \\ \sup_{H \in \mathcal{H}_n} \frac{|A_{15}|}{MISE^*(H)} &= o(1) \quad \text{and} \\ \sup_{H \in \mathcal{H}_n} \frac{|A_{16}|}{MISE^*(H)} &= o(1). \end{aligned}$$

Thus, $\sup_{H \in \mathcal{H}_n} \frac{|A_1|}{MISE^*(H)} = o(1)$. In a view to deal with term A_2 , we observe that

$$\mathbb{E}\left((\hat{r}(\chi) - r(\chi))^2 | \chi_1 \dots \chi_n\right) \leq 2\left(\mathbb{E}(\hat{r}^2(\chi) | \chi_1 \dots \chi_n) + r^2(\chi)\right) \leq C,$$

the last inequality deriving from conditions (2.2.25) and (2.2.16), and from the fact that r is bounded. Consequently, we have

$$A_2 \leq C \int \mathbb{E}\left((1 - \hat{r}_{1H}(\chi))^4\right) W(\chi) dP_\chi(\chi),$$

where the integral is dealt with in the same way as for term A_{13} , so that we obtain

$\sup_{H \in \mathcal{H}_n} \frac{|A_2|}{MISE^*(H)} = o(1)$. Finally, using Cauchy-Schwarz inequality, we observe that

$$\begin{aligned} |A_3| &\leq \sqrt{MISE^*(H)} \sqrt{|A_1|}, \\ |A_4| &\leq \sqrt{MISE^*(H)} \sqrt{|A_2|} \quad \text{and} \\ |A_5| &\leq \sqrt{|A_1|} \sqrt{|A_2|}. \end{aligned}$$

It results from this that

$$\begin{aligned} \sup_{H \in \mathcal{H}_n} \frac{|A_3|}{MISE^*(H)} &= o(1), \\ \sup_{H \in \mathcal{H}_n} \frac{|A_4|}{MISE^*(H)} &= o(1) \quad \text{and} \\ \sup_{H \in \mathcal{H}_n} \frac{|A_5|}{MISE^*(H)} &= o(1). \end{aligned}$$

The proof of our Lemma is then completed. □

Lemma 14. *Under conditions of Theorem 2, we have*

$$\frac{MISE^*(H^{CV})}{MISE^*(H^*)} \rightarrow 1 \quad a.s.$$

Proof. The proof of this Lemma is completed as soon as we can prove that

$$\left| \frac{MISE^*(H^{CV}) - MISE^*(H^*)}{MISE^*(H^*)} \right| \rightarrow 0 \quad a.s.$$

In order to show this convergence, we make use of the following upper bound:

$$\begin{aligned} &\left| MISE^*(H^{CV}) - MISE^*(H^*) \right| \tag{B.2.52} \\ &\leq \left| MISE^*(H^{CV}) - ASE(H^{CV}) \right| + \left| ASE(H^{CV}) - ASE(H^n) \right| \\ &\quad + \left| ASE(H^n) - ASE(H^*) \right| + \left| ASE(H^*) - MISE^*(H^*) \right|, \end{aligned}$$

with $H^n = \arg \min_{H \in \mathcal{H}_n} ASE(H)$. Given (B.2.52), we can further major

$$|MISE^*(H^{CV}) - MISE^*(H^*)|$$

using the inequality

$$\begin{aligned}
& \left| ASE(H^{CV}) - ASE(H^n) \right| \\
& \leq \left| ASE(H^{CV}) - ASE(H^n) - CV(H^{CV}) + CV(H^n) \right| \\
& \leq \left| \widetilde{ASE}(H^{CV}) - \widetilde{ASE}(H^n) - CV(H^{CV}) + CV(H^n) \right| \\
& \quad + \left| \widetilde{ASE}(H^{CV}) - ASE(H^{CV}) \right| + \left| \widetilde{ASE}(H^n) - ASE(H^n) \right| \\
& \leq 2 \left| CT(H^{CV}) \right| + 2 \left| CT(H^n) \right| + \left| \widetilde{ASE}(H^{CV}) - ASE(H^{CV}) \right| \\
& \quad + \left| \widetilde{ASE}(H^n) - ASE(H^n) \right| \tag{B.2.53}
\end{aligned}$$

which is valid because $CV(H^n) \geq CV(H^{CV})$ and $ASE(H^n) \leq ASE(H^{CV})$, by construction, and because of the decomposition

$$CV(H) = \widetilde{ASE}(H) - 2CT(H) + R,$$

with $R = \frac{1}{n} \sum_{j=1}^n (Y_j - r(\chi_j))^2 W(\chi_j)$. Similarly, because

$$MISE^*(H^*) \leq MISE^*(H^n),$$

one gets:

$$\left| ASE(H^n) - ASE(H^*) \right| \tag{B.2.54}$$

$$\begin{aligned}
& \leq \left| ASE(H^n) - ASE(H^*) - MISE^*(H^n) + MISE^*(H^*) \right| \\
& \leq \left| ASE(H^n) - MISE^*(H^n) \right| + \left| ASE(H^*) - MISE^*(H^*) \right| \tag{B.2.55}
\end{aligned}$$

Then, inequalities (B.2.52), (B.2.53) and (B.2.55) lead to

$$\begin{aligned}
& \left| \frac{MISE^*(H^{CV}) - MISE^*(H^*)}{MISE^*(H^*)} \right| \\
& \leq 2 \left| \frac{ASE(H^*) - MISE^*(H^*)}{MISE^*(H^*)} \right| \\
& \quad + \frac{MISE^*(H^{CV})}{MISE^*(H^*)} \left(\left| \frac{MISE^*(H^{CV}) - ASE(H^{CV})}{MISE^*(H^{CV})} \right| \right. \\
& \quad \left. + \left| \frac{\widetilde{ASE}(H^{CV}) - ASE(H^{CV})}{MISE^*(H^{CV})} \right| + 2 \left| \frac{CT(H^{CV})}{MISE^*(H^{CV})} \right| \right) \\
& \quad + \frac{MISE^*(H^n)}{MISE^*(H^*)} \left(\left| \frac{MISE^*(H^n) - ASE(H^n)}{MISE^*(H^n)} \right| + \left| \frac{\widetilde{ASE}(H^n) - ASE(H^n)}{MISE^*(H^n)} \right| \right. \\
& \quad \left. + 2 \left| \frac{CT(H^n)}{MISE^*(H^n)} \right| \right).
\end{aligned}$$

We also have

$$\frac{MISE^*(H^{CV})}{MISE^*(H^*)} = \frac{MISE^*(H^{CV}) - MISE^*(H^*)}{MISE^*(H^*)} + 1,$$

and

$$\begin{aligned}
\frac{MISE^*(H^n)}{MISE^*(H^*)} & \leq \frac{MISE^*(H^n)}{ASE(H^n)} \frac{ASE(H^*)}{MISE^*(H^*)} \\
& = \frac{1}{1 + \frac{|ASE(H^n) - MISE^*(H^n)|}{MISE^*(H^n)}} \left(\frac{|ASE(H^*) - MISE^*(H^*)|}{MISE^*(H^*)} + 1 \right),
\end{aligned}$$

where the second inequality holds because $ASE(H^n) \leq ASE(H^*)$. Combining all those results, and using notations (2.2.31), (2.2.32) and (2.2.33), we finally get

$$\left| \frac{MISE^*(H^{CV}) - MISE^*(H^*)}{MISE^*(H^*)} \right| (1 - T_\alpha - T_\beta - T_\gamma) \leq 2T_\alpha + (T_\alpha + T_\beta + T_\gamma) \frac{2}{1 - T_\alpha}.$$

Consequently, the proof of Lemma 14 is completed because of Lemmas 10, 11 and 12 which ensure the convergence to zero of T_α , T_β , and T_γ respectively. $\square$

Given those Lemmas, the proof of Theorem 2 is as follows. We consider the following decomposition, which holds because of Lemma 13:

$$\begin{aligned}
& \left| \frac{MISE(H^{CV})}{MISE(H^*)} - 1 \right| \\
= & \left| \frac{MISE(H^{CV}) - MISE(H^*)}{MISE^*(H^*)} \right| \frac{MISE^*(H^*)}{MISE(H^*)} \\
\leq & \frac{MISE^*(H^*)}{MISE(H^*)} \left(\frac{|MISE(H^{CV}) - MISE^*(H^{CV})|}{MISE^*(H^*)} \right. \\
& + \frac{|MISE^*(H^{CV}) - MISE^*(H^*)|}{MISE^*(H^*)} \\
& \left. + \frac{|MISE(H^*) - MISE^*(H^*)|}{MISE^*(H^*)} \right) \\
\leq & \left(\frac{1}{1 - \sup_{H \in \mathcal{H}_n} \frac{|MISE(H) - MISE^*(H)|}{MISE^*(H)}} \right) \\
& \cdot \left(\frac{MISE^*(H^{CV})}{MISE^*(H^*)} \sup_{H \in \mathcal{H}_n} \frac{|MISE(H^{CV}) - MISE^*(H^{CV})|}{MISE^*(H^{CV})} \right. \\
& \left. + \frac{|MISE^*(H^{CV}) - MISE^*(H^*)|}{MISE^*(H^*)} + \sup_{H \in \mathcal{H}_n} \frac{|MISE(H^*) - MISE^*(H^*)|}{MISE^*(H^*)} \right)
\end{aligned}$$

This tends to zero almost surely because of Lemmas 13 and 14. The proof of Theorem 2 is then completed.

Appendix C

Some proofs and procedures used in Chapter 4

C.1 Proof of Property 1

The proof relies on the following alternative formulation of the algorithm.

Alternative formulation of the G-BUHW algorithm

Input $\text{IN} = \{\mathcal{D}^{(p)}, \mathcal{E}^{\text{IN}}, {}_e X\}$,
 with ${}_e X = \mathbf{e} \cdot {}_e X = \sum_{j=1}^p X_j e_j$,
 where $\mathbf{e}$ is the canonical basis of $\mathbb{R}^p$,
 and ${}_e X = (X_1 \dots X_p)^t$ defines the coordinates of a vector X in $\mathbb{R}^p$ in the canonical basis.

Output: $\text{NET-ID} = \{\mathcal{D}^{(p)}, {}_\psi (\mathbf{1})_{e,\psi} X\}$ where
 ${}_\psi (\mathbf{1})_e$ is an orthonormal matrix of dimension $p \times p$ allowing the passage from the canonical basis $\mathbf{e}$ to a basis $\psi = \{\psi_k\}_{k=0 \dots p-1}$ of $\mathbb{R}^p$.
 and ${}_\psi X = (d_0, \dots, d_{p-1})^t$ defines the coordinates of the vector X in the basis ψ :
 ${}_\psi X = {}_\psi (\mathbf{1})_e {}_e X = \sum_{k=0}^{p-1} d_k \psi_k$.

Notations: Index i designates the iteration of the algorithm;
 $\mathbf{f}^{(i)}$ is the basis of $\mathbb{R}^p$ on which X is projected at iteration i .
 $\mathbf{f}'$ is a basis of $\mathbb{R}^p$.
 $\boldsymbol{\omega} = (\omega_1 \dots \omega_p)$ is a vector of weights in $\mathbb{R}^p$.
 $X^{(i)}$ is the value of the nodes remaining in the network at iteration i .
 $\langle f_1, f_2 \rangle$ designates the vectorial subspace generated by the pair of vectors f_1, f_2 .
 V^l is a vectorial subspace of $\mathbb{R}^p$.
 M is a matrix of basis change.

Initialization: $i := 1;$
 $\mathbf{f}^{(1)} := \mathbf{e}.$
 $\boldsymbol{\omega} = (1 \dots 1)^t.$
 $\mathcal{E}^{(i)} = \mathcal{E}^{\text{IN}}$
 $M := \mathbf{e}$

Iteration #i:

1. Each edge $\epsilon_l = (j, k)$ in $\mathcal{E}^{(i)}$ is associated to a vectorial subspace $V^l \subset \mathbb{R}^p$ in the following way:
 $\epsilon_l = (j, k) \leftrightarrow V^l = \langle f_j^{(i)}, f_k^{(i)} \rangle.$
2. In each subspace V^l define $\begin{pmatrix} \chi^l \\ \psi^l \end{pmatrix} := \begin{pmatrix} \sin \phi^l & \cos \phi^l \\ \cos \phi^l & -\sin \phi^l \end{pmatrix} \begin{pmatrix} f_j \\ f_k \end{pmatrix}$ with $\phi^l := \arctan(\frac{\omega_j}{\omega_k});$
 $\phi^l \in [0; \frac{\pi}{2}].$
3. Define $d_l := \psi^l.X$
4. Select an edge $\epsilon_{l^*} = (j^*, k^*)$ such that:
 $l^* := \arg \min_l |d_l|.$
 In case of multiple equal minimum values $|d_l|$, select the smallest index l .
5. Define a matrix of change of basis from $f^{(i)}$ to a basis f' as follows:

$$f'(\mathbf{1})_{f^{(i)}} := \begin{matrix} & j^* & & k^* & \\ \begin{matrix} j^* \\ \\ k^* \end{matrix} & \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & \sin \phi^{l^*} & & \cos \phi^{l^*} \\ & & & 1 & \\ & & & & \ddots \\ & & \cos \phi^{l^*} & & -\sin \phi^{l^*} \\ & & & & & 1 \\ & & & & & & \ddots \end{pmatrix} \end{matrix}$$

This is an orthonormal rotation matrix, and we have $f'X = f'(\mathbf{1})_{f^{(i)}} f^{(i)}X$.

6. Define a permutation matrix from changing from basis f' to $f^{(i+1)}$ as follows:

$$f^{(i+1)}(\mathbf{1})_{f'} := \begin{matrix} & k^* & & p-i+1 & \\ \begin{matrix} k^* \\ \\ p-i+1 \end{matrix} & \begin{pmatrix} 1 & 0 & & & \\ & \ddots & & & \\ & & 1 & 0 & \\ & & 0 & 1 & \\ & & & \ddots & \ddots \\ & & & & 0 & 1 \\ & & 1 & & 0 & 0 \\ & & & & & 1 & \ddots \\ & & & & & & \ddots & 0 \\ & & & & & & & & 1 \end{pmatrix} \end{matrix}$$

This is an orthonormal permutation matrix.

7. Encode the partial value of the NET-ID:

$$M := {}_{f^{(i+1)}}(\mathbf{1})_{f' f'}(\mathbf{1})_{f^{(i)}} M.$$

M is an orthonormal matrix as it is the product of a permutation matrix, a rotation matrix and an orthonormal matrix.

8. Prepare next iteration:

Update $\mathcal{E}^i$ by replacing all indexes k^* by j^* .

Discard duplicated edges in $\mathcal{E}^i$, keeping only the first occurrence of each edge.

This defines $\mathcal{E}^{(i+1)}$.

$${}_{f^{(i+1)}}X := {}_{f^{(i+1)}}(\mathbf{1})_{f' f'}(\mathbf{1})_{f^{(i)}} X$$

$$\omega = {}_{f^{(i+1)}}(\mathbf{1})_{f' f'}(\mathbf{1})_{f^{(i)}} \omega$$

$$i := i + 1.$$

9. Back to Step 1, until $\text{length}(\mathcal{E}^{(i)}) == 1$.

Final-step:

$${}_{\psi}(\mathbf{1})_e := M$$

$${}_{\psi}X = {}_{\psi}(\mathbf{1})_e X$$

By construction ${}_{\psi}(\mathbf{1})_e$ is a data-driven orthonormal matrix for changing from basis in $\mathbb{R}^p$ and this change is applied to the vector X . The proof of Proposition 1 is then completed as soon as we show the equivalence of this algorithm with the G-BUHWHT algorithm. This equivalence is direct as soon as we see that:

1. we have the following equivalence:

$$\frac{\omega_k}{\sqrt{\omega_j^2 + \omega_k^2}} = \sin \phi ; \quad \frac{\omega_j}{\sqrt{\omega_j^2 + \omega_k^2}} = \cos \phi \Leftrightarrow \phi = \arctan \frac{\omega_k}{\omega_j} ; \quad \phi \in [0; \frac{\pi}{2}].$$

2. the knowledge of $\mathcal{E}^{\text{OUT}}$ is equivalent to the knowledge of ${}_{\psi}(\mathbf{1})_e$. This is shown as follows.

- (a) Knowing ${}_{\psi}(\mathbf{1})_e$, $\mathcal{E}^{\text{OUT}}$ is determined by

$$\mathcal{E}_k^{\text{OUT}} = (\min_j(\psi_{kj} > 0), \min_j(\psi_{kj} < 0)), \quad \text{for } k = 1 \dots p-1,$$

$$\text{and } \mathcal{E}_0^{\text{OUT}} = (1, 1).$$

- (b) Knowing $\mathcal{E}^{\text{OUT}}$, the form of ${}_{\psi}(\mathbf{1})_e$ is obtained by iteratively building the vectors ψ_k , starting with $k = p-1$ and going bottom-up to $k = 0$, as in the following example.

$$\begin{array}{cccccccccccccccc}
& & & & & & \epsilon_{4,j} & & & & & & & & & \\
& & & & & & \epsilon_{3,j} & & & & & & & & & \\
& & & & & & & & & & & & & & & \\
& & & & \epsilon_{2,j} & \epsilon_{2,k} & \epsilon_{1,j} & & & \epsilon_{1,k} & \epsilon_{3,k} & \epsilon_{4,k} & & & & \\
\psi_0 & \left(\begin{array}{cccccccccccccccc}
\alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0 & \alpha_0
\end{array} \right) \\
\vdots & & & & & & & & & & & & & & & \\
\psi_{p-4} & \begin{array}{cccccccccccccccc}
0 & & \dots & & 0 & \alpha_{p-4} & 0 & \dots & 0 & \alpha_{p-4} & \alpha_{p-4} & \beta_{p-4} & 0 & 0 & \\
0 & & \dots & & 0 & \alpha_{p-3} & 0 & \dots & 0 & \alpha_{p-3} & \beta_{p-3} & 0 & \dots & 0 & \\
\dots & 0 & \alpha_{p-2} & \beta_{p-2} & 0 & & & & \dots & & & & & 0 & \\
0 & & \dots & & 0 & \alpha_{p-1} & 0 & \dots & 0 & \beta_{p-1} & 0 & \dots & & 0 &
\end{array} \\
\psi_{p-1} &
\end{array}$$

with the positions of α_k , $k = 0 \dots p - 1$, and β_k , $k = 1 \dots p - 1$, being determined from the the labels of the nodes j and k limiting the oriented edges ϵ_k . The $2p - 1$ unknown values in this matrix are determined by the following $2p - 1$ constraints:

- $\|\psi_k\| = 1$ and $\alpha_k > 0$, for $k = 0 \dots p - 1$.
- $\psi_0 \cdot \psi_k = 0$ for $k = 1 \dots p - 1$

The solution is unique and it is easy to verify that the resulting matrix is orthogonal. $\square$

C.2 Inversion of the G-BUHWHT

The key of the inverse transform is to reconstruct the matrix $\psi(\mathbf{1})_e$ from $\mathcal{D}^{(p)}, \mathcal{E}^{\text{OUT}}, d\}$. We do it from bottom row to top row. We index the columns of this matrix by $\ell_l, l = 1, 2, \dots, p$ according to the codebook. Imagine we have reconstructed the rows $p, p-1, \dots, k+1$ and we want to reconstruct the k^{th} row of the matrix (with $k \neq 1$). Let (A, B) denote the k^{th} element of $\mathcal{E}^{\text{OUT}}$. We know that the k^{th} row of $\psi(\mathbf{1})_e$ has elements valued either 0 or a positive value α_k or a negative value β_k . The support of the positive part consists in the column labeled A as well as all the columns that are part of the support of another row below the k^{th} one for which column A is part of the support. We note L_A the length of this support. The support of the positive part consists in the column labeled B as well as all the columns that are part of the support of another row below the i^{th} one for which column B is part of the support. We note L_B the length of this support. We then have the following constraints on α_k and β_k :

$$L_A \alpha_k + L_B \beta_k = 0, \quad \alpha_k > 0, \quad \beta_k < 0,$$

and

$$L_A \alpha_k^2 + L_B \beta_k^2 = 1.$$

Hence, we have

$$\alpha_k = \sqrt{\frac{L_B}{L_A L_B + L_A^2}}$$

$$\beta_k = -\sqrt{\frac{L_A}{L_A L_B + L_B^2}}.$$

We proceed this way to determine the rows p to 2 in $\psi(\mathbf{1})_e$. The first row is made of constant elements $\frac{1}{\sqrt{p}}$. This process determines the basis $\{\psi_k\}_{k=0\dots p-1}$. Using the codebook, we can then map those basis vectors towards the basis elements $\{\Psi_k\}_{k=0\dots p-1}$ encoding the same information in the geometry of the image. Finally, the initial image is obtained as the sum of each of those elements multiplied by the related value of the detail coefficient.

C.3 Proof of Property 4.

This property directly derives from the bottom-up building process of the G-BUHWWT. Hierarchy is induced by the minimization constraint on $\{|d_l|\}$ at each iteration of the algorithm under the constraint of the structure imposed by the graph. This may be emphasized by rephrasing iteration i in the second formulation of the algorithm, in the above proof of Property 1, as follows.

- **General idea.** At iteration i , X is decomposed as the sum of its projection on an approximation space $A^{(i)}$ and its complementary space $A^{(i)\perp}$ which encodes the missing information. The G-BUHWWT might be seen as an iterative dimension reduction process, with the dimensionality of the approximation space $A^{(i)}$ being reduced by 1 at each iteration of the algorithm.
- **Steps 1 and 2.** One reduces $A^{(i)}$ to a subspace $A^{(i+1)}$ by replacing a pair of generating vectors of $A^{(i)}$ by its weighted bisector. This means that we reduce a plane of $A^{(i)}$ to its projection onto an axis, this one being chosen so as to maximize the weight of the information kept (as measured by $|\omega \cdot \chi|$). The set of candidate planes for dimension reduction is determined by $\mathcal{E}^{(i)}$. This can be interpreted as the constraint to reduce dimensionality by reducing a pair of axes that support *neighbor* zones in the image. It can therefore be seen as a *functional constraint* on the reduction process as it means that the reduction process has to be in accordance with the geometric structure of the image, encoded by $\mathcal{E}^{(i)}$.
- **Steps 3 to 4.** The selection of the actual plane in which the dimension reduction does occur is data-driven, with the aim to lose as little information as possible

about the image at each iteration, given the functional constraint limiting the possibilities for reduction. Minimizing the set of candidate detail $\{|d^l|\}$ means that the selection is done so as to minimize the power $(d^l)^2$ of the lost information about the image.

- **Steps 5 to 8.** The change of basis at iteration i induces the following expansion of the image.

$$X = \underbrace{\sum_{j=1}^{j^*-1} (X \cdot f_j^{(i)}) f_j^{(i)} + (X \cdot \chi^{l^*}) \chi^{l^*} + \sum_{j=j^*+1}^{p-i-2} (X \cdot f_j^{(i)}) f_j^{(i)}}_{\substack{\text{projection on } A^{(i)} \\ = \text{smoothed image at iteration } i}} \quad (\text{C.3.1})$$

$$+ \underbrace{(X \cdot \psi^{l^*}) \psi^{l^*} + \sum_{j>p-i}^p (X \cdot f_j^{(i)}) f_j^{(i)}}_{\substack{\text{projection on } A^{(i)\perp} \\ = \text{loss at iteration } i}} \quad (\text{C.3.2})$$

At the last iteration, $A^{(i)}$ reduces to one single axis supporting the mean value of the image and the whole information about the structure of the image lies in $A^{(i)\perp}$.

The ordering of the vectors of $\{\psi_k\}$, and hence of $\{\Psi_k\}$, implies the hierarchy. This property might thus be seen as an ordering according to the power of the supported pattern in the image, under a functional constraint. $\square$