I N S T I T U T D E S T A T I S T I Q U E
B I O S T A T I S T I Q U E E T
S C I E N C E S A C T U A R I E L L E S
(I S B A)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

2013/31

BAGIDIS: Statistically investigating curves with sharp local patterns using a new functional measure of dissimilarity

TIMMERMANS, C. and R. VON SACHS

BAGIDIS: Statistically investigating curves with sharp local patterns using a new functional measure of dissimilarity

Catherine TIMMERMANS and Rainer VON SACHS *

30 juin 2013

Abstract

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 ^{1}H-NMR spectrometric curves and solar irradiance time series.

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

*Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université catholique de Louvain, Voie du Roman Pays, 20, B-1348 Louvain-la-Neuve, Belgium. Email: catherine.timmermans@uclouvain.be and rvs@uclouvain.be. Financial support from the IAP research network grants P06/03 and P07/06 of the Belgian government (Belgian Science Policy) is gratefully acknowledged.

1 INTRODUCTION

With this article we define, investigate and exploit an efficient measure of the dissimilarity between curves that show sharp local features. Examples for such data typically arise in numerous scientific fields, including medicine (e.g. H-NMR spectroscopic data for metabonomic analyses, EEG or ECG spectral analysis), geophysics (e.g. earth quake data) or astronomy (e.g. solar irradiance time series). A given peak in a set of such curves might be affected, from one curve to the other, by a vertical amplification, a horizontal shift or both simultaneously. The challenge in this problem comes from the observation that when trying to compare a large number of curves, for subsequent functional classification or prediction purposes, for instance, commonly used dissimilarity measures do not return coherent results as soon as there is a horizontal component of variation from one curve to another. This work proposes therefore a new dissimilarity measure which has the ability to capture both horizontal and vertical variations of the peaks, in a unified framework, i.e. in a coherent way within an integrated procedure (avoiding any preprocessing, e.g. in case of misalignment). This dissimilarity measure is embedded within a complete algorithmic procedure, which we call the Bagidis methodology, and which as such is our new proposal 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 2 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*: these bases 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 3 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 a series of 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. Section 4 is devoted to the new statistical tests and tools we propose to use within the BAGIDIS framework. 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 Timmermans et al. (2013). This later work of ours, restricted to the functional prediction set-up of Ferraty and Vieu (2006), proves rates of convergence and also treats questions of how to adaptively choose some of the nuisance parameters of our method more specifically in a supervised context. On the contrary, our present work is focused on the definition of the new dissimilarity measure and discusses its properties, without restricting ourselves to a particular kind of dissimilarity-based method.

The performances of our methodology are assessed against classical competitors on several simulated and real data examples. This is the purpose of Section 5. It shows the superiority of our method as soon as variations of the significant patterns in the curves have a 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 consist in ^{1}H NMR spectrometric curves and in solar irradiance time series.

We end this introduction by briefly repeating the threefold original contribution of our work. 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 new and relatively simple way to efficiently quantify the closeness of curves with significant sharp local features that might be simultaneously vertically amplified and horizontally shifted - without needing to borrow strength from pre-processing methods such as aligning (Bigot and Charlier 2011), or dynamically time warping (Giorgino 2009). Extensions of this genuinely univariate version of BAGIDIS to higher dimensions can be found in Timmermans and Fryzlewicz (2012) where we treat in particular image analysis and classification.

2 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 $\boldsymbol{x}$ in this basis as

$$\boldsymbol{x} = \sum_{k=0}^{N-1} d_k \boldsymbol{\psi}_k, \tag{1}$$

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

2.1 The BUUHWT Algorithm

The BUUHWT algorithm is a transform that associates a given series $\boldsymbol{x} \in \mathbb{R}^N$ to its BSUHWB expansion (eq. (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 $\{d_k\}_{k=0\ldots N-1}$ in the BSUHWB, we also explicitly need to build the basis vectors $\{\boldsymbol{\psi}_k\}_{k=0\ldots N-1}$.

A series of length N encodes the expansion coefficients of a vector $\boldsymbol{x}$ in the canonical basis $\{\boldsymbol{e}_j\}_{j=1\ldots N}$ 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, $j = 1 \ldots N$.

The BUUHWT 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 $\boldsymbol{x}$. This reduction is achieved by projecting one well-chosen plane, defined by two consecutive basis vectors, onto the single normed axis which supports 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 $\{\boldsymbol{e}_j, \boldsymbol{e}_{j+1}\}$, with $j = 1 \ldots N-1$. We note that we do not consider reducing the plane generated by $\{\boldsymbol{e}_N, \boldsymbol{e}_1\}$ or any plane generated by $\{\boldsymbol{e}_j, \boldsymbol{e}_l\}$ 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

$$\boldsymbol{\Psi}^p_{j,j+1} = \frac{\omega_j}{\sqrt{\omega_j^2 + \omega_{j+1}^2}}\,\boldsymbol{e}_j + \frac{\omega_{j+1}}{\sqrt{\omega_j^2 + \omega_{j+1}^2}}\,\boldsymbol{e}_{j+1}.$$

  It will be referred to as the *projection-axis*. Conform to the intuition, if the axes $\boldsymbol{e}_j$ and $\boldsymbol{e}_{j+1}$ have been given the same weight, the projection-axis is oriented along their bisector and supports 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 $\{\boldsymbol{e}_j, \boldsymbol{e}_{j+1}\}$ is defined as

$$L_{j,j+1} := (\boldsymbol{x}.\boldsymbol{\Psi}^d_{j,j+1})^2,$$

  with (.) indicating the scalar product, and with

$$\boldsymbol{\Psi}^d_{j,j+1} := \frac{\omega_{j+1}}{\sqrt{\omega_j^2 + \omega_{j+1}^2}}\boldsymbol{e}_j - \frac{\omega_j}{\sqrt{\omega_j^2 + \omega_{j+1}^2}}\boldsymbol{e}_{j+1},$$

  being perpendicular to $\boldsymbol{\Psi}^p_{j,j+1}$ and referred to as the *detail-axis*. Consequently, the reduction takes place in the plane generated by $\{\boldsymbol{e}_{j^*}, \boldsymbol{e}_{j^*+1}\}$ 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 $\boldsymbol{x}$ is now generated by the $N-1$ following vectors:

$$\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{j^*-1}, \boldsymbol{\Psi}^p_{j^*,j^*+1}, \boldsymbol{e}_{j^*+2}, \ldots, \boldsymbol{e}_N$$

and $\boldsymbol{x}$ is approximated by

$$\boldsymbol{x}^{[N-1]} := \sum_{\substack{j=1 \\ j\neq j^*,j^*+1}}^{N} (\boldsymbol{x}.\boldsymbol{e}_j)\,\boldsymbol{e}_j + (\boldsymbol{x}.\boldsymbol{\Psi}^p_{j^*,j^*+1})\,\boldsymbol{\Psi}^p_{j^*,j^*+1}.$$

The weight associated with $\boldsymbol{\Psi}^p_{j^*,j^*+1}$ is defined as

$$((\omega_1, \ldots, \omega_{j^*-1}, \omega_{j^*}, \omega_{j^*+1}, \omega_{j^*+2} \ldots, \omega_N).\boldsymbol{\Psi}^p_{j^*,j^*+1}),$$

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

- The discarded vector $\boldsymbol{\Psi}^d_{j^*,j^*+1}$ is encoded as the basis vector ψ_{N-1} of the Bsuhwb, and $(\boldsymbol{x}.\boldsymbol{\Psi}^d_{j^*,j^*+1})$ is encoded as the detail coefficient d_{N-1}.

We then reduce by one additional unit the dimension of the vectorial subspace in which $\boldsymbol{x}$ is projected, and we do so until the series approximation reduces to the projection of $\boldsymbol{x}$ onto one single axis. Thereby, we successively build the Bsuhwb vectors $\boldsymbol{\psi}_{N-2}, \boldsymbol{\psi}_{N-3} \ldots \boldsymbol{\psi}_1$ and the associated detail coefficients $d_{N-2}, d_{N-3} \ldots d_1$, in exactly the same manner. The remaining basis vector $\boldsymbol{\psi}_0$ is a constant normed vector and $d_0 = (\boldsymbol{x}.\boldsymbol{\psi}_0)$. 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 Buuhwt algorithm

Input

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

Output:

$\boldsymbol{x} \in \mathbb{R}^N$ is encoded in its Bsuhwb $\{\boldsymbol{\psi}_k\}_{k=0\ldots N-1}$: $\boldsymbol{x} = \sum_{k=0}^{N-1} d_k \boldsymbol{\psi}_k$.

Notations:

Index i designates the iteration of the algorithm;
$\boldsymbol{x}^{[N-i]}$ is the approximation of x in dimension $N-i$ at iteration i;
$\{\boldsymbol{f}_j\}_{j=1\ldots N-i}$ is a complete set of generating vectors for the vectorial subspace $V^{[N-i]}$ in which $\boldsymbol{x}^{[N-i]}$ is defined;
$\{\omega_j\}_{j=1\ldots N-i}$ is a set of weights associated to the $N-i$ components of $\boldsymbol{x}^{[N-i]}$;
$\boldsymbol{\Psi}^p_{j,j+1}$ is the candidate projection-axis in the plane generated by $\{\boldsymbol{f}_j, \boldsymbol{f}_{j+1}\}$;
$\boldsymbol{\Psi}^d_{j,j+1}$ is the associated detail-axis.;
$L_{j,j+1}$ is the information loss associated to a projection of the plane $\{\boldsymbol{f}_j, \boldsymbol{f}_{j+1}\}$ onto the axis $\boldsymbol{\Psi}^p_{j,j+1}$.

Initialization:

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

Iteration #i:

1. Definition of candidate pairs of generating vectors for dimension reduction:
   $\{\boldsymbol{f}_j, \boldsymbol{f}_{j+1}\}$, $j = 1 \ldots N-i$

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

Final-step:

$\boldsymbol{\psi}_0 := \boldsymbol{f}_1$;
$d_0 := \boldsymbol{x}^{[1]}.\boldsymbol{f}_1$.

In the above implementation, the computational complexity of the BUUHWT 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 $\mathcal{O}(N \log N)$.

2.2 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 $\boldsymbol{e}_j$ and $\boldsymbol{e}_{j+1}$, which supports the measurements at abscissas t_j and t_{j+1}. Assuming $\omega_j = \omega_{j+1} = 1$, the detail-vector $\boldsymbol{\Psi}^d_{j,j+1}$ is obtained as $\frac{1}{\sqrt{2}}\boldsymbol{e}_j - \frac{1}{\sqrt{2}}\boldsymbol{e}_{j+1}$. When plotted as a function of time, $\boldsymbol{\Psi}^d_{j,j+1}$ 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 $\boldsymbol{e}_j$ and by one axis that has already been obtained as the average of two basis vectors $\boldsymbol{e}_{j+1}$ and $\boldsymbol{e}_{j+2}$, the resulting detail-axis is defined as

$$\begin{aligned} \boldsymbol{\Psi}^d_{j,j+1,j+2} &= \frac{\omega_{j+1,j+2}}{\sqrt{\omega^2_j + \omega^2_{j+1,j+2}}}\boldsymbol{e}_j - \frac{\omega_j}{\sqrt{\omega^2_j + \omega^2_{j+1,j+2}}}(\frac{1}{\sqrt{2}}(\boldsymbol{e}_{j+1} + \boldsymbol{e}_{j+2})) \\ &= \frac{2}{\sqrt{6}}\boldsymbol{e}_j - \frac{1}{\sqrt{6}}(\boldsymbol{e}_{j+1} + \boldsymbol{e}_{j+2}). \end{aligned}$$

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 BUUHWT algorithm. Furthermore, the so-built basis vectors of the

BSUHWB are orthonormal, as the vectors $\mathbf{\Psi}^p$ and $\mathbf{\Psi}^d$ are always defined through a plane rotation in an orthonormal basis, so that orthonormality is preserved. For instance, we have

$$\begin{pmatrix} \mathbf{\Psi}^p_{j,j+1} \\ \mathbf{\Psi}^d_{j,j+1} \end{pmatrix} = \begin{pmatrix} \sin\phi & \cos\phi \\ \cos\phi & -\sin\phi \end{pmatrix} \begin{pmatrix} \boldsymbol{e}_j \\ \boldsymbol{e}_{j+1} \end{pmatrix}$$

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

$$\boldsymbol{\psi}_k(t) = (\frac{1}{b_k - s_k + 1} - \frac{1}{e_k - s_k + 1})^{1/2} . \mathbf{1}_{s_k \leq t \leq b_k} \quad (2)$$

$$-(\frac{1}{e_k - b_k} - \frac{1}{e_k - s_k + 1})^{1/2} . \mathbf{1}_{b_k+1 \leq t \leq e_k}, \quad (3)$$

where $t = 1 \ldots N$ 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 $\boldsymbol{\psi}_k$ (s_k, b_k and e_k stand for *start*, *breakpoint* and *end* respectively). Figure 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 $\boldsymbol{\psi}_k$ 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 $\boldsymbol{x}$ on the wavelet $\boldsymbol{\psi}_k$ encodes the importance of the related level change in the series $\boldsymbol{x}$. 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 BUUHWT 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 2, *left*.

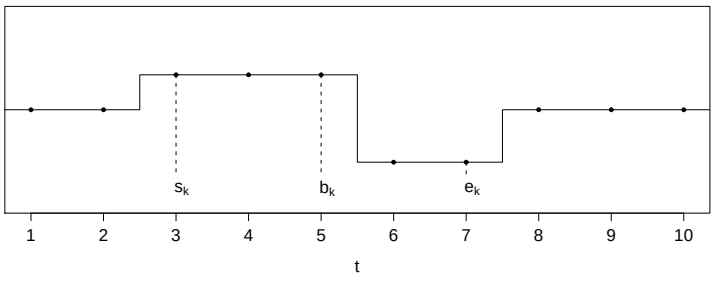

Figure 1: **An unbalanced Haar wavelet.** The points s_k, b_k and e_k (eq. 2) are identified.

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 $\{\boldsymbol{\psi}_k\}_{k=0\ldots N-1}$ is uniquely determined by the ordered set of breakpoints $\{b_k\}_{k=1\ldots N-1}$ of its wavelet basis vectors.

As a consequence, and because expansion (1) is invertible and hence unique, the set of points $\{\boldsymbol{y}_k\}_{k=1\ldots N-1} = \{(b_k, d_k)\}_{k=1\ldots N-1}$ determines the *shape* of the series $\boldsymbol{x}$ uniquely, the complete determination of $\boldsymbol{x}$ 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* $\boldsymbol{x}$ *in the b-d plane.* An example of such a representation is presented in Figure 2, *right.*

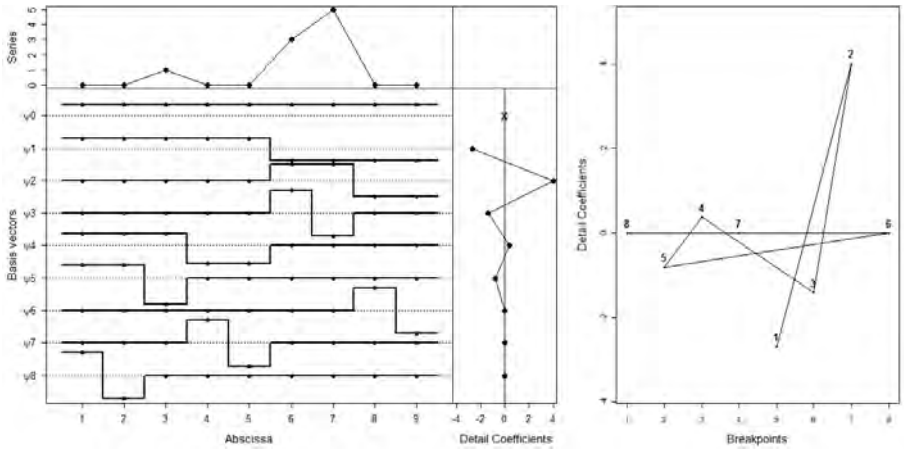

Figure 2: *Left:* **Illustration of the BSUHWB expansion of a series.** The series is plotted in the upper part of the Figure. The corresponding abscissa axis is common for that graph and for the main graph, so that it is located at the bottom of that one. The main part of the Figure shows the basis 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 $\boldsymbol{\psi}_0$ is not indicated. *Right:* **Representation of the same series in the $b-d$ plane.** Points $(b_k, d_k)_{k=1\ldots N-1}$ are numbered according to their rank k in the hierarchical BSUHWB and linked to each other.

The Hierarchy in the BUUHWT Expansion. Finally, the BUUHWT induces an interesting property of hierarchy in the resulting BSUHWB 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 BSUHWB 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 BUUHWT algorithm itself : by construction, our series $\boldsymbol{x}$ is encoded in its BSUHWB 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 BUUHWT 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 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.

3 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 breakpoint and detail coefficients.

3.1 Definition

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

$$d_p^{\text{BAGIDIS}}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \sum_{k=1}^{N-1} w_k \delta_p^k(\boldsymbol{y}_k^{(1)}, \boldsymbol{y}_k^{(2)}), \tag{4}$$

with

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

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 $\left| b_k^{(1)} - b_k^{(2)} \right|$ 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 $\left| d_k^{(1)} - d_k^{(2)} \right|$ 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 (4), 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 (4). 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. For further details on how to solve this optimisation problem (in a supervised context), we refer to our companion paper (Timmermans et al. 2013) on embedding our BAGIDIS methodology into the more theoretical framework of nonparametric

functional regression. Some perspectives are given in Subsection 3.3.2, however. In the examples of this section, we make use of two different approaches.

- The following weight function

$$w_k = \frac{\log(N+1-k)}{\log(N+1)}, \quad k = 1 \dots N-1, \tag{5}$$

  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 $\delta_p^k = \left\| \boldsymbol{y}_k^{(1)} - \boldsymbol{y}_k^{(2)} \right\|_p$ 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.

3.2 Properties

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

A Visual Motivation. The motivation for equation (4) 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

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

It has thus the same properties as a distance, except that $d(\boldsymbol{x}^i, \boldsymbol{x}^j) = 0 \not\Longrightarrow \boldsymbol{x}^i = \boldsymbol{x}^j$. 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 $\|.\|_p$, and because the weights w_k are non negative, we have immediately:

$$d_p^{\text{Bagidis}}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \geq 0,$$

$$d_p^{\text{Bagidis}}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = d_p^{\text{Bagidis}}(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}),$$

and

$$
\begin{aligned}
d_p^{\text{BAGIDIS}}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) &= \sum_{k=1}^{N-1} w_k \left\| \boldsymbol{y}_k^{(1)} - \boldsymbol{y}_k^{(2)} \right\|_p \qquad (6) \\
&\leq \sum_{k=1}^{N-1} w_k \left\| \boldsymbol{y}_k^{(1)} - \boldsymbol{y}_k^{(3)} \right\|_p + \sum_{k=1}^{N-1} w_k \left\| \boldsymbol{y}_k^{(3)} - \boldsymbol{y}_k^{(2)} \right\|_p \\
&= d_p^{\text{BAGIDIS}}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(3)}) + d_p^{\text{BAGIDIS}}(\boldsymbol{x}^{(3)}, \boldsymbol{x}^{(2)}).
\end{aligned}
$$

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 3.3.1. 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 (4).

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 3.3.1. It is thus a major strength of the BAGIDIS semi-distance.

An example is provided with the curves in Figure 3 *top left*. A series $\boldsymbol{x}$ is defined. It is compared with 5 different series. A series is defined as $1.125 * \boldsymbol{x}$, this amplification factor being chosen so that the upper plate of the series rises by 1 unit. It is also compared with the series $\boldsymbol{x}$ *shifted by one unit to the left*. Those three series are expected to be measured close to each other. The series defined as $1.125 * \boldsymbol{x}$ *shifted by one unit to the left* is expected to be close to them too, although slightly more distant of $\boldsymbol{x}$ 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 $\boldsymbol{x}$. Is 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 3 *top right*, *middle left*, *middle right*, *bottom left* and *bottom right*, respectively, as compared with the signature of $\boldsymbol{x}$. All those series have a sparse encoding in the BSUHWB: only three 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. 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) 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 3.3.1.

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: **Dissimilarity matrices between the 6 series defined in Figure 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.

3.3 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

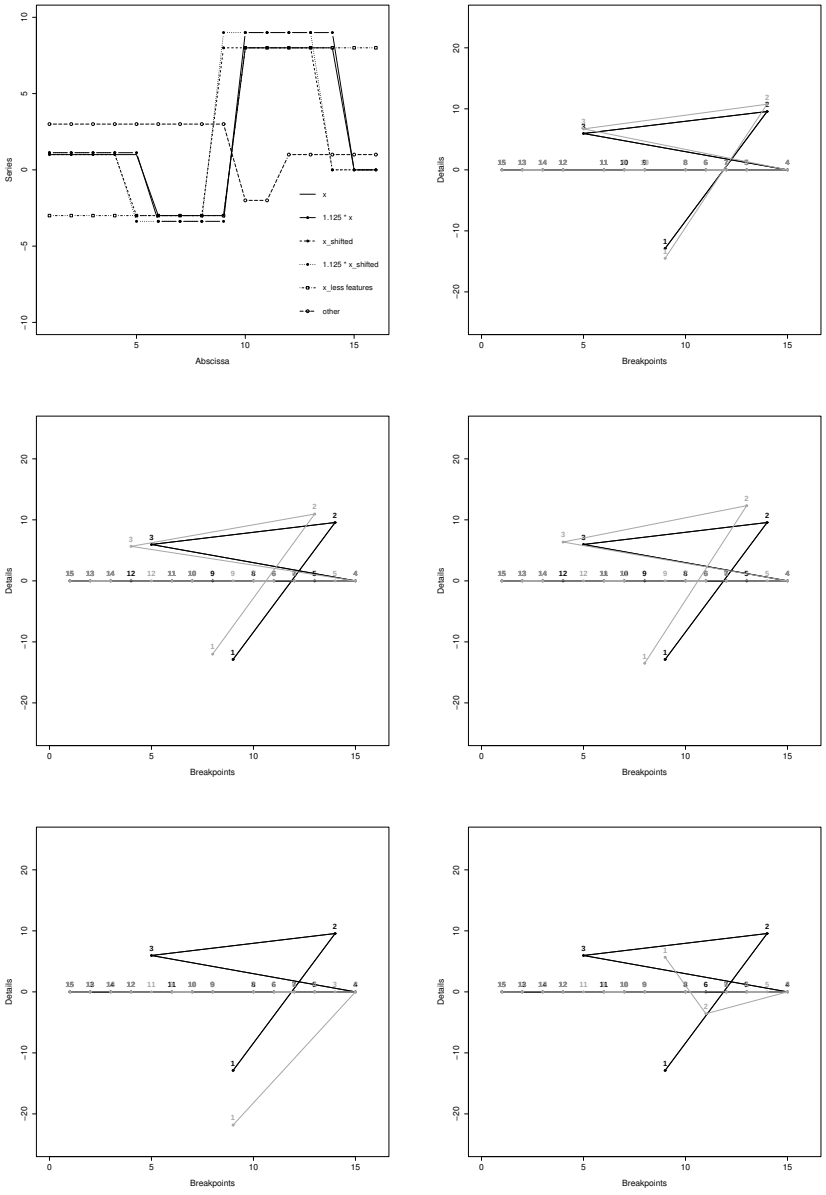

Figure 3: **Comparison of 6 series in the $b-d$ plane.** *Top left:* Representation of the 6 series. *From top right to bottom left:* Representation in the b-d plane of the series $\boldsymbol{x}$ (in black) and $1.125 * \boldsymbol{x}$, $\boldsymbol{x}$ *shifted to the left*, $1.125 * \boldsymbol{x}$ *shifted to the left*, a simplified version of $\boldsymbol{x}$ and an other series, successively (in grey).

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.

3.3.1 Comparison With Existing Approaches

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.

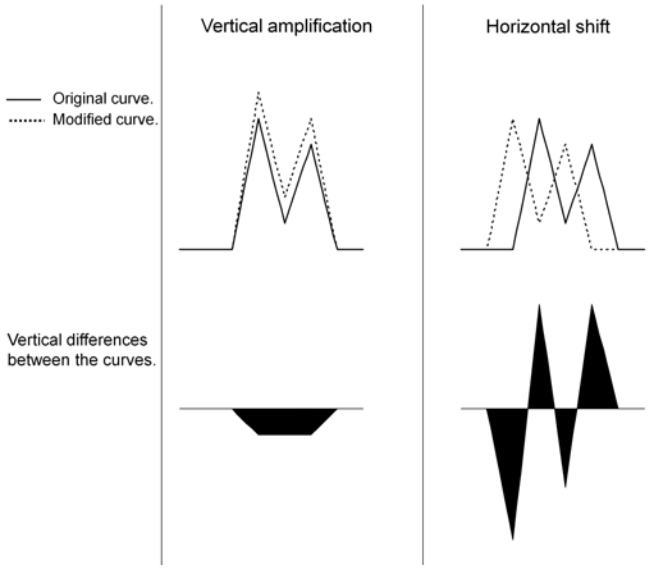

Figure 4: Schematic illustration of the difficulty for point-to-point distances to take into account horizontal variations of curves.

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 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 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.

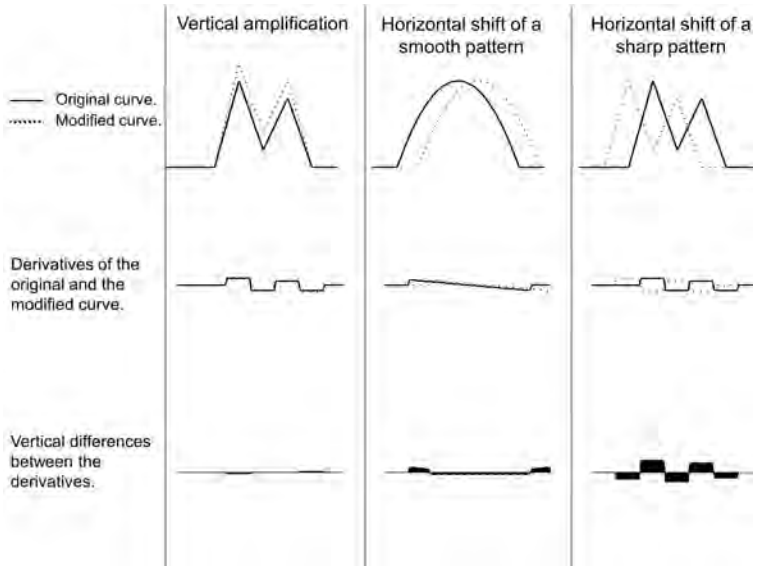

Figure 5: Schematic illustration of the difficulty for derivative-based functional semi-distances to take into account horizontal variations of curves with sharp features.

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 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 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 multiscale 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 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

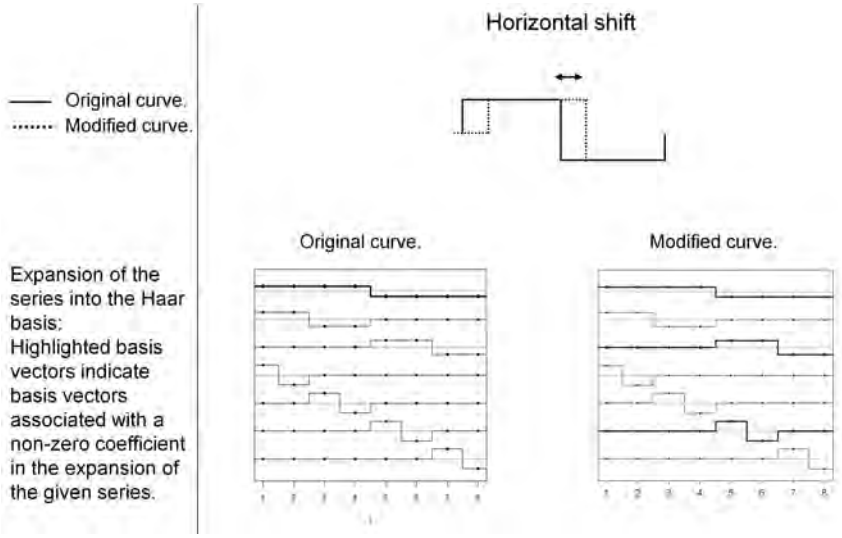

Figure 6: Schematic illustration of the difficulty for wavelet-based methods to detect the similarity of curves shifted relative to each other.

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 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. In 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 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.

3.3.2 Parametrization and Extensions

Scaling. Unavoidably, the dissimilarities we measure using p-norm partial distances (as in expression (4)) are directly dependent on the scales of the 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

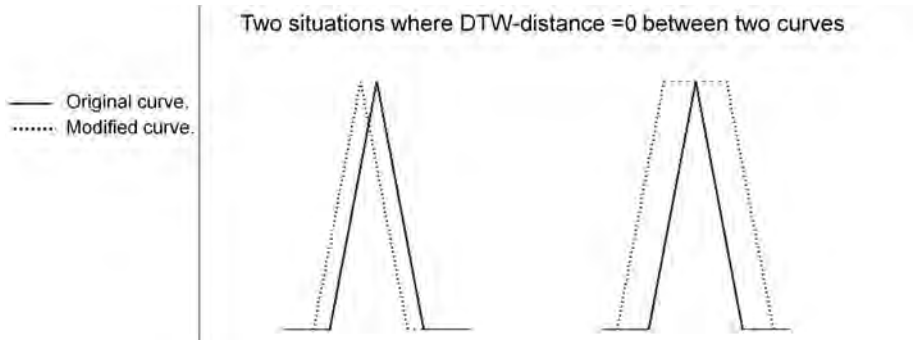

Figure 7: Schematic illustration of situations where the *dynamic time warping* distance returns a distance zero between curves of which a pattern is affected by an horizontal shift or shape variation.

following variant of the BAGIDIS dissimilarity between $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$:

$$\begin{aligned} d_{p\lambda}^{\text{BAGIDIS}}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) &= \sum_{k=1}^{N-1} w_k \left\| \boldsymbol{y}_k^{(1)} - \boldsymbol{y}_k^{(2)} \right\|_{p,\lambda} \\ &= \sum_{k=1}^{N-1} w_k \left(\lambda \left| b_k^{(1)} - b_k^{(2)} \right|^p + (1-\lambda) \left| d_k^{(1)} - d_k^{(2)} \right|^p \right)^{1/p}. \end{aligned} \tag{7}$$

This dissimilarity measure is shown to be a semi-distance in exactly the same manner as d_p^{BAGIDIS}, by observing that $\left\| \boldsymbol{y}_k^{(1)} - \boldsymbol{y}_k^{(2)} \right\|_{p\lambda} = \left\| \tilde{\boldsymbol{y}}_k^{(1)} - \tilde{\boldsymbol{y}}_k^{(2)} \right\|_p$, with $\tilde{\boldsymbol{y}}_k^{(i)} = (\lambda^{1/p}\, b_k^{(i)},\, \lambda^{1/p}\, d_k^{(i)})$, $i = 1, 2$. 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 section, 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 Timmermans et al. (2013) in the framework of kernel nonparametric functional regression as defined by Ferraty and Vieu (2006). In this companion work, we propose to optimize the weight function using a cross-validation procedure, and we prove the optimality of the selected weights. Moreover, we show 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 $\boldsymbol{y}_k = (b_k, d_k)$ obtained for Gaussian noises and the empirical distribution of the $\tilde{\boldsymbol{y}}_k = (\tilde{b_k}, \tilde{d_k})$ 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 $\boldsymbol{x}^{(i)}$, 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 $\boldsymbol{x}^{(i)}$. 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 2.2). This observation was used for defining the *prior* weight function used in this section (equation 5).

- 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 BUUHWT 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 treated in Timmermans and Fryzlewicz (2012) .

Regarding the selection of an unbalanced Haar basis that is well-suited to a series, an alternative to BUUHWT 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 BUUHWT transform has an extension to images, as defined in Timmermans and Fryzlewicz (2012) 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 BUUHWT 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 et al. 2004). On the contrary, the BUUHWT 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.

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 datas, 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 BUUHWT expansion of the so-defined subseries and we compute the BAGIDIS semi-distance defined by expression (7). 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 5.4 and 5.5, 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 5.4 and 5.5.

4 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 5.

4.1 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

$$\bar{\boldsymbol{x}}^{(b-d)} \equiv (\bar{\boldsymbol{b}}, \bar{\boldsymbol{d}}) \quad \text{where} \quad \bar{b}_k = \frac{1}{n}\sum_{i=1}^{n} b_k^{(i)} \quad \text{and} \quad \bar{d}_k = \frac{1}{n}\sum_{i=1}^{n} d_k^{(i)}. \tag{8}$$

At each rank k, $\bar{\boldsymbol{x}}_k^{(b-d)}$ 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 $\bar{\boldsymbol{x}}^{(b-d)}$ 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 $\bar{\boldsymbol{x}}(t) = (1/n)\sum_{i=1}^{n} \boldsymbol{x}^{(i)}(t)$ would miss, for the reasons discussed in Subsection 3.3.1.
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 $\bar{\boldsymbol{b}}$. Nevertheless, definition (8) 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:

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

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

4.2 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 $\bar{\boldsymbol{x}}^{(d-d,g)} \equiv (\boldsymbol{b}^{(g)}, \boldsymbol{d}^{(g)})$, $g = 1 \dots G$, and suppose the global empirical mean curve has been estimated as $\bar{\boldsymbol{x}}^{(b-d,0)} \equiv (\boldsymbol{b}^{(0)}, \boldsymbol{d}^{(0)})$. 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

$$F = \frac{\frac{1}{G-1}\sum_{g=1}^{G} n_g \, d_{p\lambda}^{\text{BAGIDIS}^2}(\bar{\boldsymbol{x}}^{(b-d,g)}, \bar{\boldsymbol{x}}^{(b-d,0)})}{\frac{1}{n-G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} d_{p\lambda}^{\text{BAGIDIS}^2}(\boldsymbol{x}^{(g,i)}, \bar{\boldsymbol{x}}^{(b-d,g)})}, \tag{9}$$

where n_g is the number of curves in group g, $g = 1 \dots G$, and $\boldsymbol{x}^{(g,i)}$ 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 paper. The point here is to emphasize the fact that such a test for differences between groups is possible in the BAGIDIS framework.

4.3 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:

$$\begin{cases} \textbf{Test 1}: \ H_0: \ d(G1,G1) = d(G1,G2); \ H_1: d(G1,G1) < d(G1,G2). \\ \textbf{Test 2}: \ H_0: \ d(G2,G2) = d(G1,G2); \ H_1: d(G2,G2) < d(G1,G2). \end{cases} \tag{10}$$

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 8. 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.

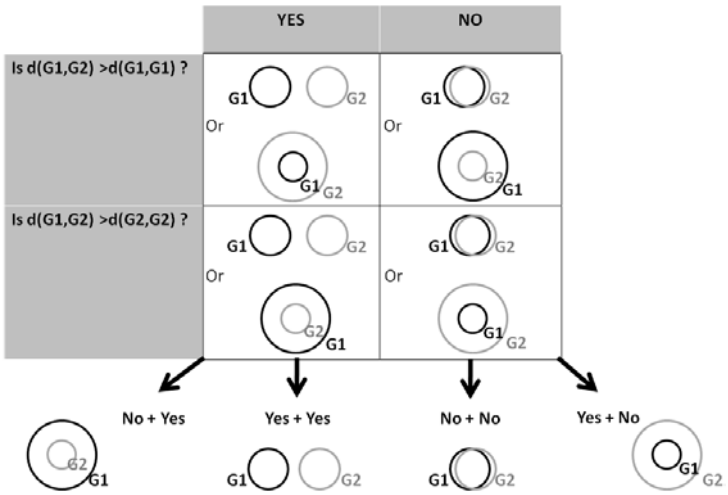

Figure 8: *Left:* **Summary table for the interpretation of the double t-test for the relative position in space of two groups of curves** $G1$ **and** $G2$**,** according to a given distance or semi-distance.

5 DATA ANALYSES USING BAGIDIS

Real and simulated applications are proposed in this section. A simple simulated example of supervised classification is proposed in Subsection 5.1. 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 the noise level. Another example involving sparse curves is then provided in Subsection 5.2. A simulated example of Anova F-test within the Bagidis framework is presented in Subsection 5.3. Two real datasets are then investigated. Both involve data with important sharp features. The first dataset, in Subsection 5.4, 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 5.5, we study the similarity of the dynamics of solar irradiance time series. This dataset is highly noisy, with measurements that are collected with large time steps. 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.

5.1 A Simulated Example of Supervised Classification

The following test is proposed: 10 mother series are defined (Figure 9, *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: 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 noisefree 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^{Eucl} (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 9, *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^{EUCL} and d^{HAAR} are essentially superimposed.

Table 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^{EUCL} are provided, those rates being essentially similar to those obtained using $d^{\text{BAGIDIS}}_{2,0.5}$ and d^{HAAR} respectively. Most misclassification rates of $d^{\text{BAGIDIS}}_{1,0.5}$ 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 3.3.1. $d^{\text{BAGIDIS}}_{1,0.5}$ 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 BUUHWT 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 asymetric. Conversely, the asymetry of the oscillating patterns of the x10-family is correctly captured so that this family is well-identified.

Percentages of misclassification

	$\sigma = 0.01$		$\sigma = 3$		$\sigma = 6$	
Family	$d^{\text{BAGIDIS}}_{1,0.5}$	d^{EUCL}	$d^{\text{BAGIDIS}}_{1,0.5}$	d^{EUCL}	$d^{\text{BAGIDIS}}_{1,0.5}$	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 2: Individual misclassification rates for the simulated example of Subsection 5.1. Percentages of misclassification for each family of series are presented for $d^{\text{BAGIDIS}}_{1,0.5}$ 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.

5.2 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 problem 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 10 shows example series out of the datasets generated for each level of the noise. We compute the dissimilarity matrix for the datasets using

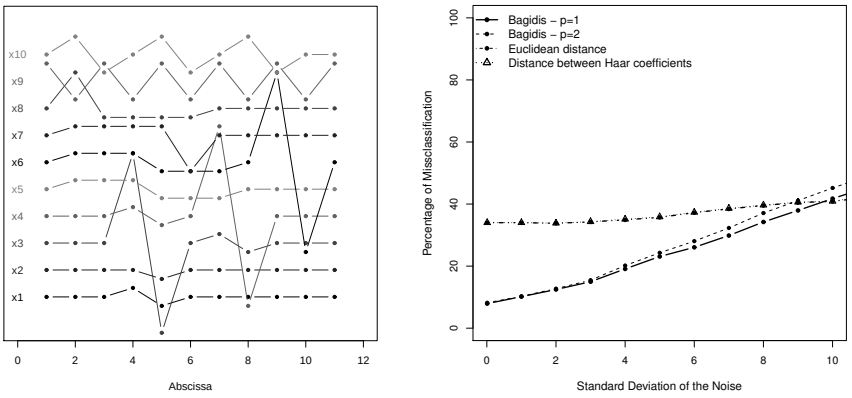

Figure 9: *Left*: The 10 *mother* series of the supervised classification example. *Right*: Percentages of misclassification as a function of the standard deviation of the Gaussian noise affecting the simulated series. Results for d^{Eucl} and d^{Haar} are essentially superimposed. Results for $d^{\text{Bagidis}}_{1,0.5}$ are slightly better than for $d^{\text{Bagidis}}_{2,0.5}$, however close.

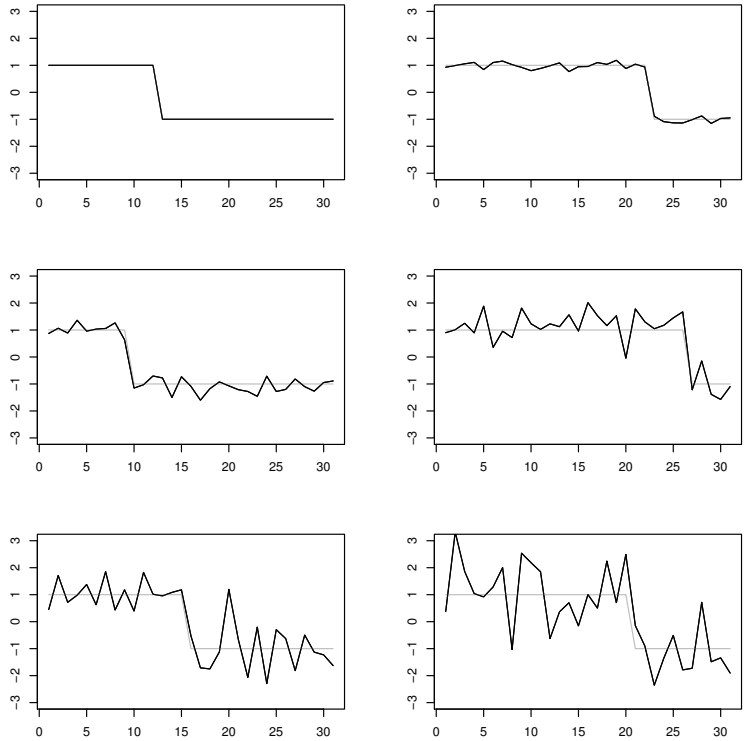

Figure 10: **Illustration of the simulated data used in Subsection 5.2**. *From left to right and top to bottom*: simulated series with standard deviation fixed at $\sigma = 0, 0.1, 0.25, 0.5, 0.75$ and 1 respectively, and different location of the jump. The true underlying signal is plotted in gray.

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 11. We observe that the BAGIDIS semi-distance is the only one which essentially preserves the empirically expected diagonal structure of the dissimilarity matrix up to the maximal level of noise tested.

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

A BAGIDIS-ANOVA F-test (equation (9)) 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 12, *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 12, *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 12, *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 assumption of equal variance in each group (Bishop and Dudewicz 1978; Weerahandi 1995), which could easily be applied here.

5.4 Searching for Factors Effects on a Set of Spectrometric Curves

Dataset and Experimental Design. The dataset we consider is made of 24 spectra of biological serum (Figure 13). 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 ^{1}H NMR data in the field of metabonomics (Dubois 2009; Rousseau 2011). The spectra have been acquired using ^{1}H 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 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 ^{1}H 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

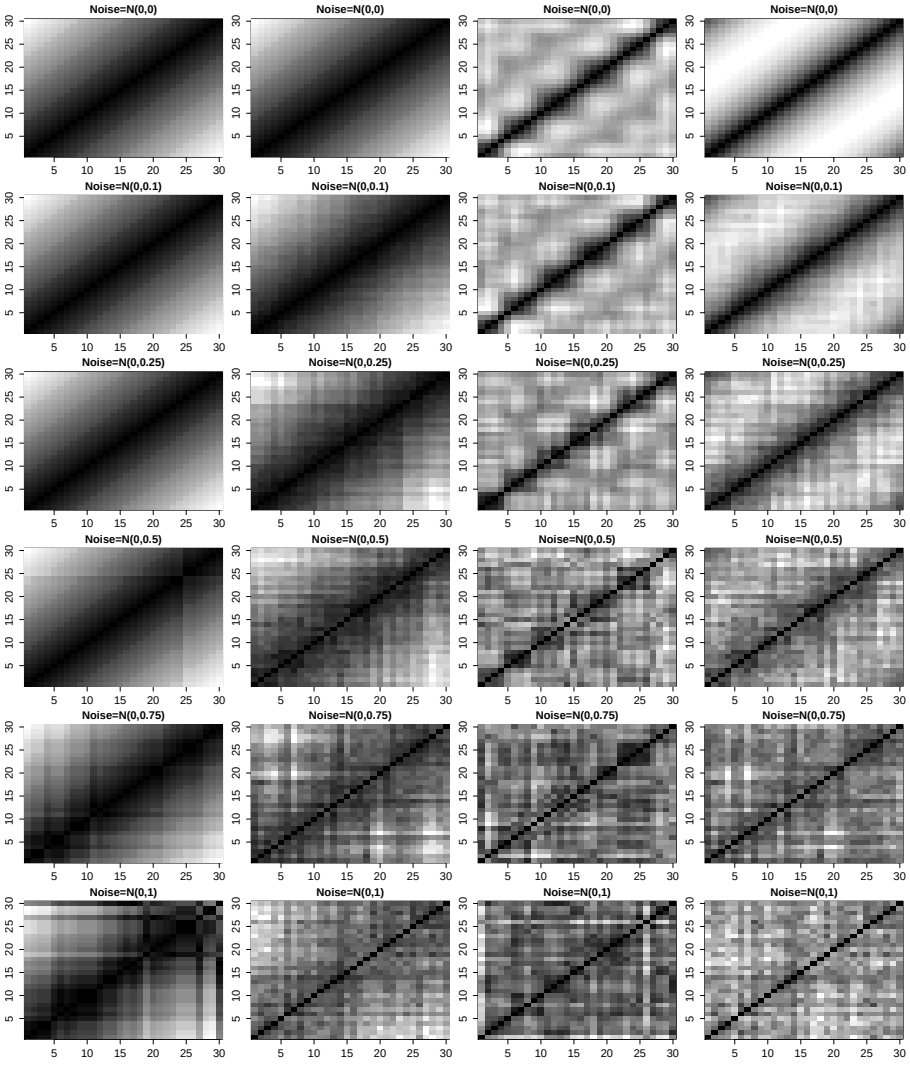

Figure 11: **Images of the dissimilarity matrices between possibly noisy sparse curves with a moving jump from level** 1 **to level** -1**.** The indices identifying the curves in the representation of the dissimilarity matrices encode the location of the jump. Black indicates a distance 0, white indicates 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.

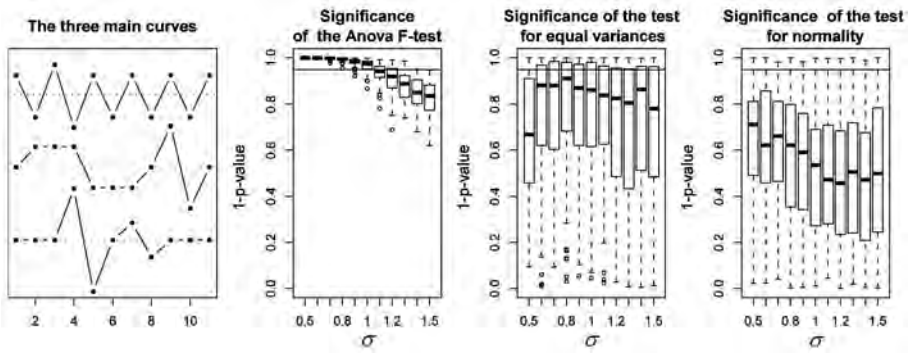

Figure 12: *From left to right:* The three mother curves for the simulation study. Boxplots of the significances (= 1− *p-values*) of the Bagidis Anova F-test for each tested value of the noise. Boxplots of the significances of the Bartlett test for equal variances amongst the groups. Boxplots of the significances of the Jarque-Bera normality test for the residuals.

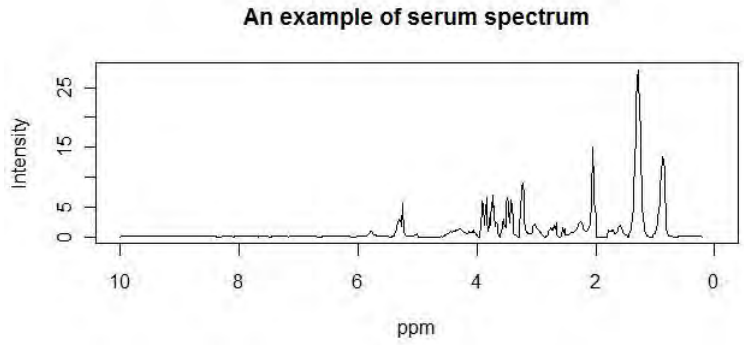

Figure 13: An example of preprocessed serum spectrum, obtained using ^{1}H NMR spectroscopy, with a pulse sequence called CPMG that leads to the suppression of the proteinic markers in the spectra - otherwise, proteins mask the signals of other molecules of interest: the metabolites (Laboratoire de Chimie Pharmaceutique, Université de Liège).

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 14, *top left*, shows the result we obtain using the Bagidis semi-distance $d^{\text{Bagidis}}_{1,0.5}$ (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 unfrozing. Principal component analysis (Figure 14, *top right*) and a functional semi-distance based either on the comparison of the first derivatives (Figure 14, *center left*) or on the first functional principal components (Figure 14, *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 14, *bottom left*, and Figure 14, *bottom right*, respectively. The projection we obtain with $\lambda = 1$ is clearly similar to our first result using Bagidis (Figure 14, *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

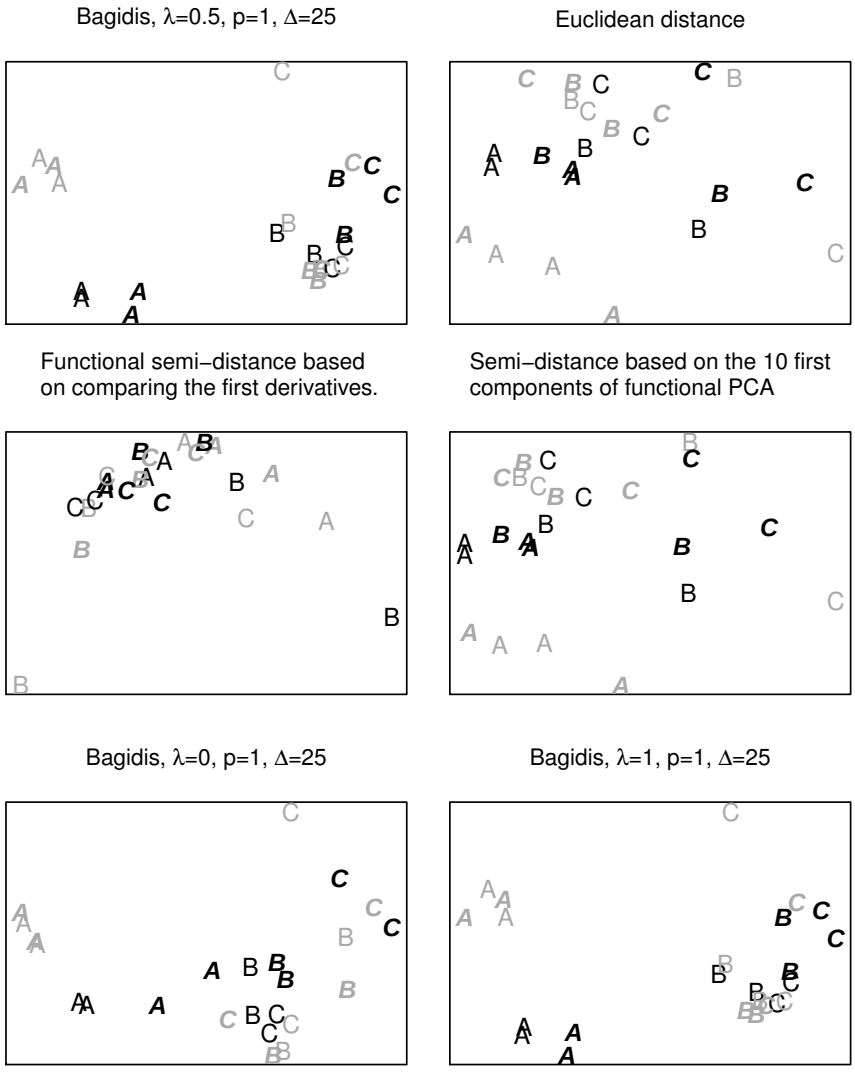

Figure 14: Multidimensional scaling representations of the spectrometric dataset, based on several matrices of distances or semi-distances. The projected spectra are labeled according to the levels of the three factors of the experimental design: they are denoted A, B or C depending on the day of collection 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.

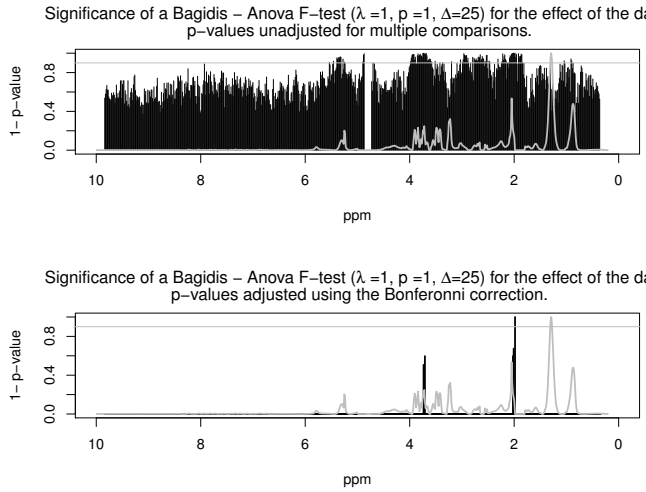

Figure 15: *Top:* Significances ($= 1-$*p-values*) of each individual Bagidis Anova F-test for the effect of the day on the spectra. *Bottom:* Significances of those Bagidis Anova F-test corrected for 726 multiple simultaneous comparisons (Bonferroni correction). In both cases, one typical spectrum is superimposed to the graphs.

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 (9)) 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 15, *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 15, *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 significances of the F-tests in Figure 15, *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 16), 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 (10)) 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

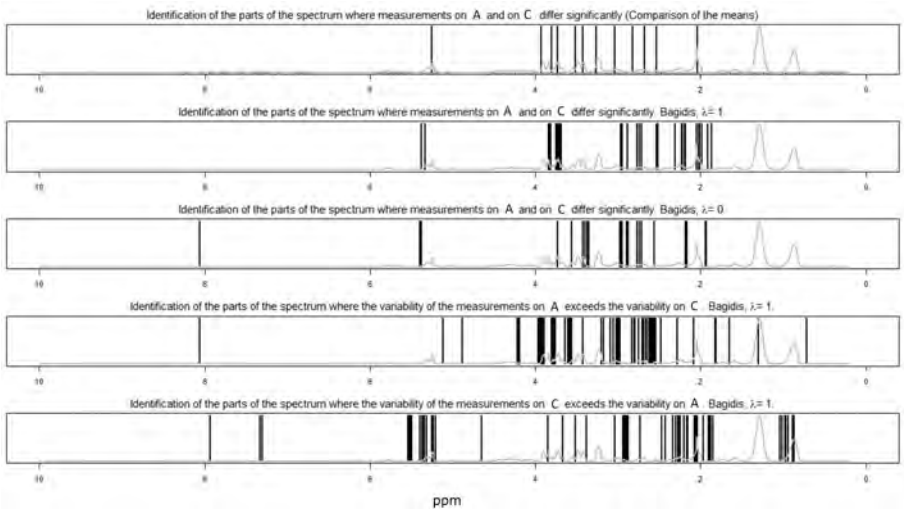

Figure 16: Results of multiple geometrical t-tests for the relative geometrical positions of the groups of spectra measured on *day A* and on *day C*. On *row 1, row 2 and row 3*, vertical lines identify the parts of the spectrum that are detected significantly different (with a confidence level of 95%, 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.

design. Figure 16, *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 16, *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 highlighted 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 16, 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.

5.5 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 18). 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 17, 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 groups corresponding to the temperature - i.e. to the altitude of emission in the solar atmosphere.

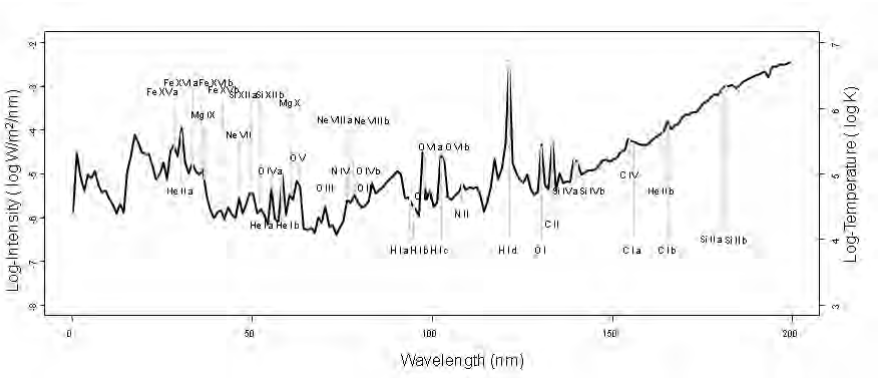

Figure 17: **Representation of the solar EUV spectrum**, as measured by the instrument Timed/SEE (dataset SEE-level3A-v.10, (Woodraska and Woods 2009)) and **identification of the 38 emission lines** we study. The black curve represents the log intensity of the irradiance as a function of the wavelength. It is measured along the left Y-axis. The lines are identified by the ions that generate them. As some ions emit multiple lines, we sometimes add a letter at the end of the names of the ions so as to distinguish between the different lines. Those names are placed above or below the corresponding lines, at a height that corresponds to its log-temperature, as indicated on the right Y-axis. The hotter lines are emitted by the more external layer of the solar atmosphere, the corona, and the colder ones are emitted by the inner layer called chromosphere.

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.

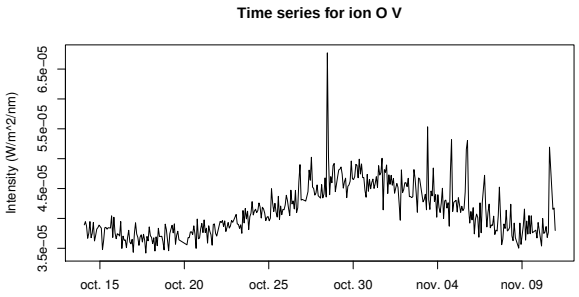

Figure 18: Irradiance time series for the spectral line O V at 62.97 nm from October 14 to Novembre 10, 2003, as measured by the instrument Timed/SEE (dataset SEE-level3A-v.10, (Woodraska and Woods 2009)).

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 19, *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 19, *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.

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.

6 CONCLUSION

In this paper, 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

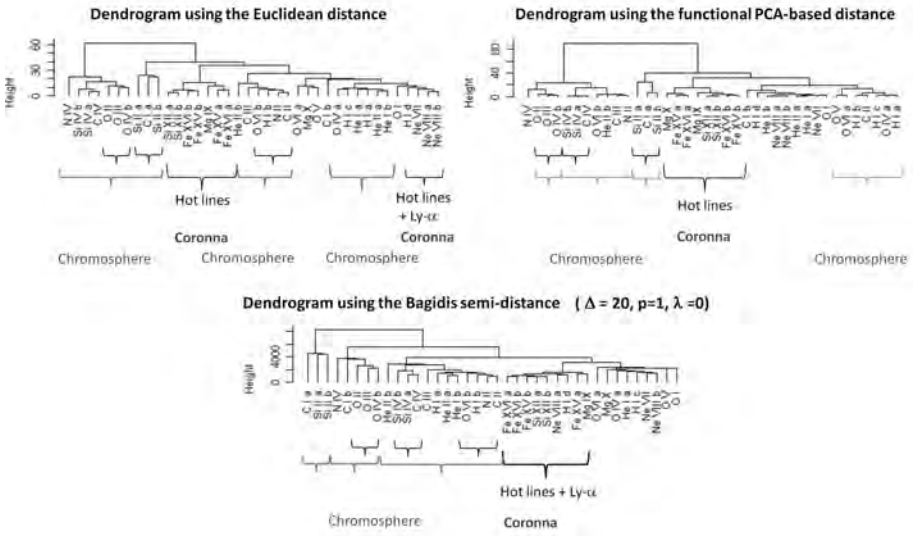

Figure 19: Dendrograms obtained using a Ward's agglomerative algorithm on the 38 solar irradiance time series based on various distances or semi-distances. The time series are labeled by the ions that generate them. Significant groups are underbraced and an indication of the related region of emission is provided.

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 comparison of curves, as illustrated in this study. Two of those have been treated in two companion papers: the use of BAGIDIS in Nonparametric Functional Regression (Timmermans et al. 2013), including a data-driven definition of the weight function, and a generalization of BAGIDIS to image processing (Timmermans and Fryzlewicz 2012). Moreover, a more detailed application of BAGIDIS to real data in spectrometry can be found in Timmermans et al. (2012).

Acknowledgements. *The authors thank Réjane Rousseau for the preprocessing of the spectrometric data, Matthieu Kretzschmar for interesting discussions about Timed/SEE data, and Léopold Simar, Piotr Fryzlewicz and Michel Verleysen for their helpful comments. 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 grants P06/03 and P07/06 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.*

APPENDIX: TECHNICAL DETAILS FOR THE TESTS FOR RELATIVE LOCATIONS OF THE SPECTRA OF *DAY A* AND *DAY C*.

This appendix details the way we obtained the results shown in Figure 16. 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 8 we obtained the results shown on *rows 2, 4 and 5* in Figure 16. 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 16, 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 16.

References

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.

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.

Bigot, J. and Charlier, B. (2011), "On the consistency of Fréchet means in deformable models for curve and image analysis," *Electronic Journal of Statistics*, 5, 1054–1089.

Bishop, T. A. and Dudewicz, E. J. (1978), "Exact Analysis of Variance with Unequal Variances: Test Procedures and Tables," *Technometrics*, 20, 419–430.

Cailliez, F. (1983), "The Analytical Solution of the Additive Constant Problem," *Psychometrika*, 48, 343–349.

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.

Delouille, V., Simoens, J., and von Sachs, R. (2004), "Smooth design-adapted wavelets for nonparametric stochastic regression," *Journal of the American Statistical Association*, 99, 643–658.

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. and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, Springer Series in Statistics, Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Fryzlewicz, P. (2007), "Unbalanced Haar Technique for Non Parametric Function Estimation," *Journal of the American Statistical Association*, 102, 1318–1327.

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 Lp on General Measure Spaces," *Journal of Fourier Analysis and Applications*, 3, 457–474.

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.

Kruskal, W. and Wallis, W. (1952), "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, 47, 583–621.

Lebart, L., Morineau, A., and Piron, M. (2004), *Statistique exploratoire multidimensionnelle.*, Dunod, 3rd ed.

Morris, J. S. and Carroll, R. J. (2006), "Wavelet-based functional mixed models," *Journal of the Royal Statistical Society, Series B*, 68, 179–199.

Rousseau, R. (2011), "Statistical contribution to the analysis of metabonomics data in ^{1}H NMR spectroscopy." Ph.D. thesis, Université catholique de Louvain, Louvain-la-Neuve, Belgium.

Sweldens, W. (1996), "Wavelets and the lifting scheme: A 5 minute tour," *Zeitschrift für Angewandte Mathematik und Mechanik*, 76 (Suppl. 2), 41–44.

Timmermans, C., de Tullio, P., Lambert, V., Frédérich, M., Rousseau, R., and von Sachs, R. (2012), "Advantages of the BAGIDIS methodology for metabonomics analyses: application to a spectroscopic study of Age-related Macular Degeneration," in *Proceedings of the 12th European Symposium on Statistical Methods for the Food Industry*, pp. 399–408.

Timmermans, C., Delsol, L., and von Sachs, R. (2013), "Using Bagidis in nonparametric functional data analysis: predicting from curves with sharp local features." *Journal of Multivariate Analysis*, 115, 421–444.

Timmermans, C. and Fryzlewicz, P. (2012), "Shah: Shape-Adaptive Haar Wavelet Transform For Images With Application To Classification," *submitted*, ISBA Discussion Paper 2012-15, Université catholique de Louvain.

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.

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.