INSTITUT DE STATISTIQUE

BIOSTATISTIQUE ET

SCIENCES ACTUARIELLES

(ISBA)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

2014/07

The BAGIDIS distance: about a fractal topology, with applications to functional classification and prediction

VON SACHS, R. AND C. TIMMERMANS

The BAGIDIS distance: about a fractal topology, with applications to functional classification and prediction

Rainer von Sachs and Catherine Timmermans

Abstract The BAGIDIS (semi-) distance of Timmermans and von Sachs [9] is the central building block of a nonparametric method for comparing curves with sharp local features, with the subsequent goal of classification or prediction. This semi-distance is data-driven and highly adaptive to the curves being studied. Its main originality is its ability to consider simultaneously horizontal and vertical variations of patterns. As such it can handle curves with sharp patterns which are possibly not well-aligned from one curve to another. The distance is based on the signature of the curves in the domain of a generalised wavelet basis, the Unbalanced Haar basis. In this note we give insights on the problem of stability of our proposed algorithm, in the presence of observational noise. For this we use theoretical investigations from Timmermans, Delsol and von Sachs [7] on properties of the fractal topology behind our distance-based method. Our results are general enough to be applicable to any method using a distance which relies on a fractal topology.

1 Introduction

In Timmermans and von Sachs [9], a new method for the statistical analysis of differences between curves with sharp local patterns proposes a distance measure between curves which relies on an Unbalanced Haar wavelet decomposition obtained using a modified version of the algorithm by Fryzlewicz [3]. This algorithm allows to describe a curve through a set of points in the so-called breakpoints-details (b,d) plane, where the breakpoints account for the location of level changes in the curve

Rainer von Sachs
Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain,
20 Voie du Roman Pays, B-1348 Louvain-la-Neuve, e-mail: rvs@uclouvain.be

Catherine Timmermans
Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain,
20 Voie du Roman Pays, B-1348 Louvain-la-Neuve, e-mail: catherine.timmermans@uclouvain.be

and details account for the amplitude of the latter. The goal has been to propose a method capable of statistically investigating datasets of curves with sharp peaks that might be misaligned, thereby overcoming limitations of existing methods. We also recall our paradigm of a ”robust” one-step method in order to avoid any preprocessing step - such as Dynamic Time Warping ([4]) - which would align the curves prior to comparison, for the purpose of, e.g., classification or prediction. This is in particular in order to be able to detect differences between curves due to the presence of features which are actually not aligned.

In this note we address the question of stability of the (b,d) point representation (= the “signature”) associated to a given curve if there is some additional noise. In our previous work [9], we have been able to show that this (b,d) point representation is stable in the absence of noise, leading to good ”theoretical” performance. This stability has been due to the use of an unambiguous ordered representation of each curve in the (b,d) plane. Hence, it is of importance to examine what happens in the realistic situation of noise.

More particularly, we address the question of robustness of our method towards the following phenomenon of potential feature confusion. Due to noise, curves with a sufficiently similar structure of local events (such as jumps, peaks or troughs) might accidentally be considered as dissimilar because local information might be encoded in a suboptimal, i.e. not unambiguous, way. We investigate the theoretical properties of the BAGIDIS semi-distance in order to handle this situation. With this, we support empirical findings reported in previous work of ours ([9], [7] and [10]) when using the local methods used to process the set of BAGIDIS semi-distances computed on our noisy datasets. Here with ”local” we mean essentially nonparametric methods which localise the information in the given data set by using only a fraction of the observations given in a local neighbourhood around the point (or region) of interest. Prominent examples are methods based on Nearest Neighbors (NN), kernels, or Multidimensional Scaling (MDS) which turned out to effectively be sufficiently robust in order to cope with the aforementioned feature confusion. In this article we shed some light on why this happens, leading to the desirable property that makes BAGIDIS better than competitors (e.g. the Euclidean distance) in case of misaligned sharp patterns. We note that the opposite problem of classifying accidentally as similar those curves that would actually be dissimilar in the absence of noise is not the purpose of this examination because this problem, inherent to any distance based classification algorithm, is not caused or amplified by the aforementioned problem of loss of unambiguous ordering due the presence of noise.

Section 2 of this paper reviews what is necessary to recall about the BAGIDIS method. At the end of this section we empirically expose what is behind our problem of robustness towards feature confusion. In Section 3 we present our theoretical treatment of the consistency of BAGIDIS in view of this problem, and in fact, any nonparametric method for functional comparisons using a distance which relies on a fractal topology. In particular we give arguments in favour of the BAGIDIS distance compared to the traditional Euclidean distance.

We finish this introduction by noting that extensions of our univariate work to higher dimensions have been provided in Timmermans and Fryzlewicz [8], in the particular context of classification of images.

2 Motivation for and description of the BAGIDIS algorithm

We consider series that are made of N regularly spaced measurements of a continuous process (i.e. a curve). Those series are encoded as vectors in $\mathbb{R}^N$. There exists numerous classical methods allowing to measure distances or semi-distances between such series coming from the discretization of a curve. Note that, according to [2], d is a semi-distance on some space $\mathscr{F}$ if

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

Semi-distances are often used ([1]) when one is interested in comparing the shapes of some groups of curves, but not in comparing their mean level.

2.1 Existing distance-based approaches

We very briefly recall a non-exhaustive collection of some most popular existing distance-based approaches and discuss their properties: (i) Classical l_p distances and their principal components-based extension ([6]); (ii) Functional semi-distances ([2]), taking into account the notion of neighborhood in point-to-point comparisons; (iii) wavelet-based distances: comparing the coefficients of well-suited basis function expansions. Whereas in our work [9] we give a detailed appreciation of these different approaches, here in order to motivate our approach, we contain ourselves to recall some basic visually supported features: in Figure 1a, we recall that by methods of type (i) the ordering of the series measurements is not taken into account so that the evolutions of two series cannot be compared; (ii) functional approaches happen to fail when dealing with curves with local sharp discontinuities that might not be well aligned from one curve to another one, as illustrated in Figure 1b; and finally, more particularly, (iii) encoding significant features into a wavelet basis which is not both simultaneously orthogonal and non-dyadic in nature, can lead to shortcomings as illustrated with classical (dyadic) Haar wavelets in Figure 1c).

2.2 At the core of the BAGIDIS method

A major originality of our method to encode closeness of series having a similar discontinuity that is only slightly shifted relies on projections on *orthogonal* basis

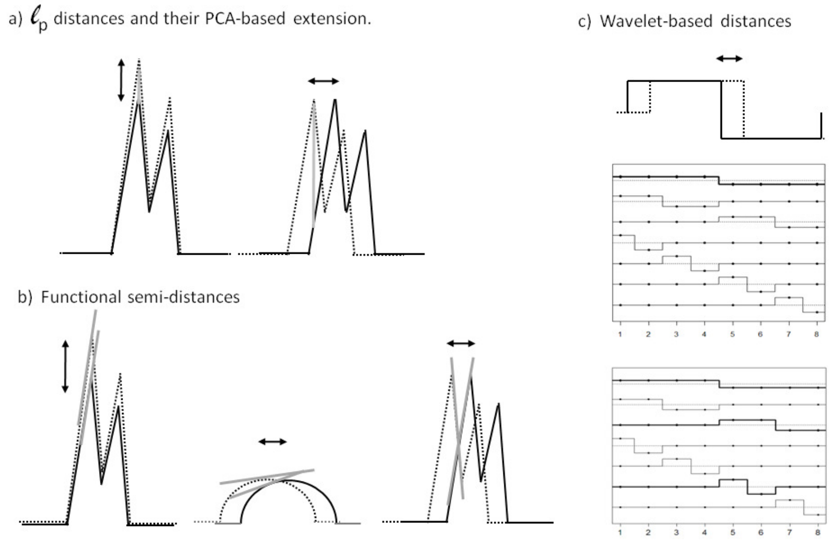
a) ℓ_p distances and their PCA-based extension.
b) Functional semi-distances
c) Wavelet-based distances
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8

Fig. 1 Schematic illustration of the difficulty for classical methods to take into account horizontal variations of curves. *a)* l_p distances and PCA-based distances compare curves at each point of measurement, so that patterns that are shifted horizontally are measured distant. An illustrative component of the point-to-point distances is displayed in gray in (a). *b)* Comparing derivatives, as common functional methods do, allows to overcome that difficulty if the patterns are smooth but fails with sharp shifted patterns. Illustrative derivatives are indicated in light gray in (b). *c)* Wavelet-based methods capture well the sharp patterns, but their encoding in the basis expansion differs highly if the location of the discontinuity changes a bit: in (c), we illustrate classical Haar-basis expansions of two shifted step function, the only basis vector associated to a non-zero coefficient being highlighted in bold.

functions that are *different* from one series to another although providing for a *hierarchy* (essential for the ability of comparing the curve expansions). We describe the main ideas of the method as follows.

(i) We consider a collection of discrete-time series each of which can be pictured as regularly spaced observations of a curve. We note that patterns in a series can be described as a set of level changes.
(ii) We find an *optimal* basis for each curve. As a first step, we want to expand each series in a basis that is best suited to it, in the sense that its first basis vectors should carry the main features of the series, while subsequent basis vectors support less significant patterns. In that respect, we are looking for a basis that is organized in a *hierarchical* way. As a consequence, there will be a particular basis associated to each series. As the series are thought of as described by their level changes, we will consider that the meaningful features for describing them are both locally important level changes, such as jumps, peaks or troughs, and level changes affecting a large number of data, i.e. discontinuities of the mean level. From this point of view, Unbalanced Haar wavelet bases, to be defined in subsection 2.3, are the ideal can-

didates for our expansion. We benefit from their *o*rthogonality property to have no unambiguity in encoding.
(iii) We take advantage of the *hierarchy* of those bases. Given this, we make use of the BAGIDIS semi-distance which is at the core of the BAGIDIS methodology. This semi-distance takes advantage of the hierarchy of the well-adapted unbalanced Haar wavelet bases: basis vectors of similar rank in the hierarchy and their associated coefficients in the expansion of the series are compared to each other, and the resulting differences are weighted according to that rank. This is actually a clue for decrypting the name of the methodology, as the name BAGIDIS stands for *BAsis GIving DIStances*. Subsection 2.4 recalls the definition of the BAGIDIS semi-distance from [9].

A subsequent interest lies in obtaining some information on the relative importance of horizontal and vertical variations, and on their localization, in order to statistically diagnose whether groups of curves do actually differ and how. Numerous applications to supervised and unsupervised classification and prediction, in the framework of spectroscopy for metabonomic analysis, on analysing solar irradiance time series or on image description and processing, can be found in [9], [7], [10] and [8].

2.3 Finding an optimal basis for each curve

Given a set of M series $\mathbf{x}^{(i)}$ in $\mathbb{R}^N$, $i = 1..M$, each of which consists in discrete regularly spaced measurements of a (different) curve, the goal is now to expand each of the series into the Unbalanced Haar wavelet basis that is best suited to it.

2.3.1 Definition of the Unbalanced Haar wavelet bases

Unbalanced Haar wavelet bases ([5]) are orthonormal bases that are made up of one constant vector and a set of Haar-like, i.e. *up-and-down*-shaped, orthonormal wavelets whose discontinuity point between positive and negative parts is not necessarily located at the middle of its support. Using the notation of [3], the general mathematical expression of those Haar-like wavelets is given by

$$\begin{aligned}\phi_{e,b,s}(t) = (\frac{1}{b-s+1} - \frac{1}{e-s+1})^{1/2}.\mathbf{1}_{s\leq t\leq b} \\ -(\frac{1}{e-b} - \frac{1}{e-s+1})^{1/2}.\mathbf{1}_{b+1\leq t\leq e},\end{aligned} \qquad (1)$$

where $t = 1..N$ is a discrete index along the abscissa axis, and where s , b and e stands for *start*, *breakpoint* and *end* respectively, for some well chosen values of s, b and e along the abscissa axis (see also Figure 1 of [9]). Each wavelet $\phi_{e,b,s}(t)$ is thus associated with a level change from one observation (or group of observations) to the consecutive one, and the projection of the series $\mathbf{x}(t)$ on the wavelet $\phi_{e,b,s}(t)$ encodes the importance of the related level change in the series.

2.3.2 The Basis Pursuit algorithm and the property of hierarchy

In 2007, P. Fryzlewicz [3] proposed an algorithm for building the unbalanced Haar wavelet basis $\{\phi_k\}_{k=0..N-1}$ that is best suited to a given series, according to the principle of hierarchy - namely, the vectors of this basis and their associated coefficients are ordered using information that builds on the importance of the level change they encode for describing the global shape of the series. He called it the *bottom-up unbalanced Haar wavelet transform*, here-after BUUHWT. The resulting expansion is organized in a hierarchical way and avoids the dyadic restriction that is typical for classical wavelets. The family of unbalanced Haar wavelets is thus really adaptive to the shape of the series. A chart-flow diagram of the actual BUUHWT algorithm can also be found in Section 2.1 of [9].

2.3.3 An example of Bottom-Up Unbalanced Haar Wavelet expansion

Figure 2, *left*, shows the BUUHWT expansion obtained for one particular series. As hoped for and observed by looking at the location of the discontinuity points b

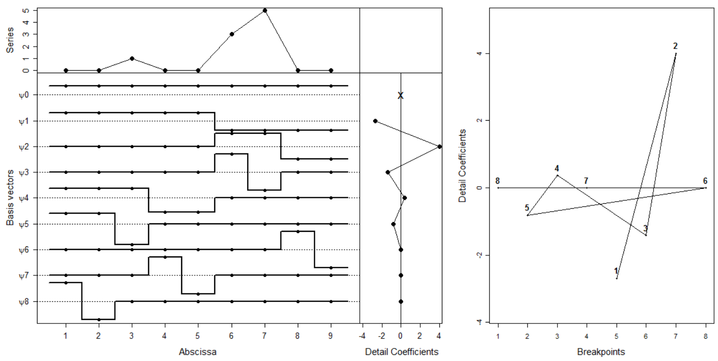

Fig. 2 *Left:* **Illustration of a BUUHWT expansion.** In the upper part of the figure we plot the series. The corresponding abscissa axis at the very bottom is common for that graph and for the graph of basis vectors. The main part of the figure shows the basis vectors of the Unbalanced Haar wavelet basis that is best suited to the series (BUUHWT basis). These vectors are represented rank by rank, as a function of an index along the abscissa axis. Dotted horizontal lines indicate the level zero for each rank. Vertically, from top to bottom on the right hand side, we find the detail coefficients associated with the wavelet expansion. Each coefficient is located next to the corresponding basis vector. For graphical convenience, the value of the coefficient d_0 associated with the constant vector ψ_0 is not indicated. *Right:* **Representation of a series in the $b-d$ plane.** The same series is plotted in the plane that is defined by the values of its breakpoints and its detail coefficients. Points are numbered according to their rank in the hierarchy.

between positive and negative parts of the wavelets, the first non-constant vectors support the largest discontinuities of the series and encode therefore the highest peak of the series. Subsequent vectors point to smaller level changes while the few last vectors correspond to zones where there is no level change - as indicated by the associated zero coefficient.

2.3.4 Representing the series in the *b*-*d* Plane

Let us denote the optimal Unbalanced Haar wavelet expansion of a series $\mathbf{x}^{(i)}$ as follows:

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

where the coefficients $d_k^{(i)}$ are the projections of $\mathbf{x}^{(i)}$ on the corresponding basis vectors $\psi_k^{(i)}$ (i.e. the *detail* coefficients) and where the set of vectors $\{\psi_k^{(i)}\}_{k=0..N-1}$ is the Unbalanced Haar wavelet basis that is best suited to the series $\mathbf{x}^{(i)}$, as obtained using the BUUHWT algorithm. Let us also denote $b_k^{(i)}$, the breakpoint of the wavelet $\psi_k^{(i)}$, $k = 1..N-1$, i.e. the value of the highest abscissa where the wavelet $\psi_k^{(i)}$ is strictly positive. An interesting property of the basis $\{\psi_k^{(i)}\}_{k=0..N-1}$, that has been proved by [3], is the following:

Property: The ordered set of breakpoints $\{b_k^{(i)}\}_{k=0..N-1}$ determines the basis $\{\psi_k^{(i)}\}_{k=0..N-1}$ uniquely.

Consequently, the set of pairs $(b_k^{(i)}, d_k^{(i)})_{k=1..N-1}$ determines the shape of the series $\mathbf{x}^{(i)}$ uniquely (i.e., it determines the series, except for a change of the mean level of the series, that is encoded by the additional coefficient $d_0^{(i)}$). This allows us to represent any series $\mathbf{x}$ in the *b*-*d* plane, i.e. the plane formed by the breakpoints and the details coefficients. An example of such a representation is presented in Figure 2, *right*.

2.4 *A semi-distance taking advantage of the hierarchy of the BUUHWT expansions*

We now have at our disposal all elements to measure the dissimilarity between two curves $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ that are both made of N consecutive observations and whose BUUHWT expansions have been computed. We proceed by calculating the weighted sum of partial distances in the *b*-*d* plane, i.e. the weighted sum of partial dissimilarities evaluated rank by rank:

$$d_p(\mathbf{x}^{(1)},\mathbf{x}^{(2)}) = \sum_{k=1}^{N-1} w_k \left\| \mathbf{y}_k^{(1)} - \mathbf{y}_k^{(2)} \right\|_p, \quad \text{with} \quad p = 1,2,\ldots,\infty \tag{2}$$

where $\mathbf{y}_k^{(i)}$ stands for $(b_k^{(i)}, d_k^{(i)})$, $i = 1,2$, so that $\left\| \mathbf{y}_k^{(1)} - \mathbf{y}_k^{(2)} \right\|_p$ is the distance between the pairs representing the curves at rank k in the b-d plane, as measured in any norm $p = 1,2,\ldots\infty$, and where w_k is a suitably chosen weight function with K non-zero weights.
As becomes clear in the sequel, and in particular in Section 3, the number K is quite crucial as it encodes the number of features found to be significant for discrimination. Its choice, and more generally, the choice of the weight function, can be actually be done by cross validation, cf. [7]. In an unsupervised context, one can either use a priori information about how many ranks K should be necessary to encode important local information (such as prominent peaks in the observed signal), or in absence of this, use a uniformly not too badly working weight function which shows a smooth decay towards zero for higher ranks.

Note that we do not consider the rank $k = 0$ in our dissimilarity measure (2), as we are mainly interested in comparing the structures of the series rather than their mean level.

A more general version of definition (2) allows being flexible with respect to scaling effects. In the following, in order to take into account that sensitivity, an additional parameter $\lambda \in [0,1]$ is introduced that balances between the differences of the details and the differences of the breakpoints:

$$d_p(\mathbf{x}^{(1)},\mathbf{x}^{(2)}) = \sum_{k=1}^{N-1} w_k \left(\lambda \left| b_k^{(1)} - b_k^{(2)} \right|^p + (1-\lambda). \left| d_k^{(1)} - d_k^{(2)} \right|^p \right)^{1/p}. \tag{3}$$

We refer to [9] for more details on how to choose the parameter λ.

Property: As shown in [9], this measure of dissimilarity is a *semi-distance*.

2.5 Behavior of BAGIDIS in the presence of noise

Unambiguity in the absence of noise. For the BAGIDIS semi-distance between two curves to be useful, it is necessary that no ambiguity occurs in the hierarchic encoding of the patterns in the algorithm. This constraint essentially translates into the idea that the description of the series should not require more than one level change of any given amplitude. What happens, in an ideal situation of no noise, if a series consists in an exactly symmetric pattern, such as a peak or trough, centered on the middle point of the series (i.e. an even pattern with respect to $\frac{N}{2}$)? We then would observe two level changes of exactly the same amplitude, but opposite signs, and encode the left and right part of the pattern. This ambiguity is solved by defining a left orientation for the BUUHWT algorithm, meaning that a level change of given amplitude that is located to the left of another one with the same amplitude is always encoded first.

The presence of noise. The key to applicability of our method to noisy series is the use of a suitable weight function that efficiently filters the noise. In several examples, to be found in [9], we observed a good robustness of the method with respect to the presence of additive noise. Nevertheless, an artefact of the method might occur, for example, in case of a symmetric or quasi-symmetric peak, when there are two detail coefficients being close in absolute value but of opposite sign (compare the extreme case of equality as discussed above). In this case a permutation of the basis vectors might occur in the algorithm constructing the best suited basis, from one series to the other, due to a possible reordering of the amplitudes (in absolute values) of the mentioned noisy coefficients. Consequently a clustering of noisy series might lead to a spurious distinction into two groups.

The robustness property for classification/clustering. In case there is a split into two groups A and B, in general, this will not invalidate the analysis of whether or not this split has been spurious: If there are no differences between the groups, a clustering will give two groups (due to the permutations occurring in both series A and series B) but each group will contain a mix of series A and B, so that we will conclude that there are no differences in the distributions of groups A and B. On the opposite, if there are significant differences between groups A and B, a clustering will give four groups, amongst which two are made of series A (the distinction between the groups being an artefact) and two are made of series B. We will thus conclude to the presence of an effect of the A-B factor.

This point is illustrated by the following test-scenarios. A more theoretical treatment what is behind this empirical property is given in Sections 3.1 and higher.

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

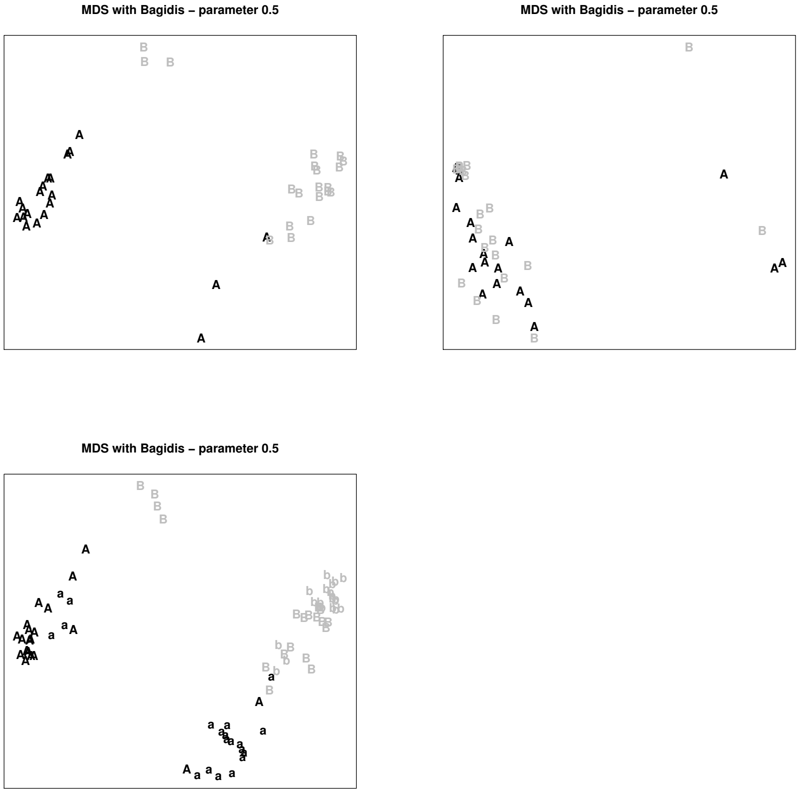

Fig. 3 Study of the applicability of BAGIDIS for noisy series in case of a possible reordering in the hierarchy of the patterns. *From left to right:* MDS representation of the test dataset resulting from Scenario 1 to Scenario 3.

In case of a regression or a classification problem, the key for a successful prediction model is that a given permutation should occur with a sufficiently high probability so that each spurious group contains enough representatives. Using a local approach for prediction (such as nonparametric functional regression or *k*-nearest neighbors) is then possible. In the next section we support this claim theoretically by use of the properties of our semi-distance which are derived from the fact that this distance actually induces a fractal topology ([2], [7]). In [7] we derived results on the rate of convergence of the nonparametric functional regression estimate based on BAGIDIS, obtained under mild conditions. In accordance with the above mentioned intuition, the key for the estimator to converge relatively fast is that the probability to find curves in a small ball around the point of prediction is high enough.

3 On a fractal topology induced by the BAGIDIS semi-distance

We now discuss some properties of our semi-distance which are derived from the fact that this distance actually induces a fractal topology ([2], [7]). This allows to address the question evoked at the end of Section 2 on the stability of our algorithm with respect to some potential feature confusion, in the sense of comparing closeness of two curves with closeness of their signatures in the (b,d)-plane.

Somewhat naively one could likely be asking the following question:

Property P1: Given a curve x sampled on a grid $\mathbb{N}_{[1;N]}$ and two noisy replications of this series, $x^{(1)} = x + noise$ and $x^{(2)} = x + noise$, we have that the signatures $s^{(1)} = \{(b_k^{(1)}, d_k^{(1)})\}_{k=0}^{N-1}$ and $s^{(2)} = \{(b_k^{(2)}, d_k^{(2)})\}_{k=0}^{N-1}$ are close to each other, at least when the number of sampled points tends to infinity and the noise tends to zero.

However, it is in general not possible to derive a framework for showing such an ideal property that would allow us to include the treatment of the 'breakpoint' components $\{b_k\}$ into an asymptotic result similar to the one of denoising curves by non-linearly selecting (via thresholding, e.g.) the 'best' wavelet coefficients $\{d_k\}$. Luckily, satisfying P1 is not a condition that needs to be fulfilled to address the feature permutation problem of our method. (For a further illustration of this point, we refer to our remark at the end of Section 3.1.) Actually, what is necessary for the method to be valid for subsequent classification, clustering or prediction applications, is rather the reverse statement:

Property P2: Given two signatures $s^{(1)} = \{(b_k^{(1)}, d_k^{(1)})\}_{k=0}^{N-1}$ and $s^{(2)} = \{(b_k^{(2)}, d_k^{(2)})\}_{k=0}^{N-1}$ in the (b,d) plane, if $s^{(1)}$ and $s^{(2)}$ are close enough to each other, then $x^{(1)}$ is close to $x^{(2)}$.

The idea behind this statement lies in the fact that when using local methods for processing the dataset of curves we can base ourselves on the following assumption:

Assumption A1: If the curves $x^{(1)}$ and $x^{(2)}$ are close enough to each other, they will behave similarly with respect to the property we investigate (e.g. associated response in regression, class membership in classification or discrimination, ...).

As mentioned in the first part of this paper, kernel methods, NN algorithms or radial-basis functions networks are very common examples of local methods, so do distance-based algorithms clearly fall into this category of methods. Property P2 ensures that Assumption A1 can be transposed to curves expressed in the (b,d) plane so that we can use local methods that rely on the BUUHWT expansions of curves.

3.1 Theoretical results based on fractal topologies

As discussed above, satisfying P2 is a condition that has to be satisfied when considering local methods. For ensuring the efficiency of such methods, the next step is to be sure that there is a sufficient density of observations around each point in the b-d plane at which we want to predict an associated response (class membership, cluster index, scalar response ...). Intuitively, we need to have in our dataset a sufficient number of neighbors that enter into the computation of the local algorithm at hand.

This "density of the space" is a topological property that is measured through the *small ball probability* of finding a curve $x^{(2)}$ around $x^{(1)}$, which is defined ([2]) as

$$\phi_{D,x^{(1)}}(h) = P(x^{(2)} \in B_D(x^{(1)},h)),$$

where $B_D(x^{(1)},h)$ is the ball of radius h centered on $x^{(1)}$ and defined according to the semi-distance D. More generally, for a given semimetric d, the small ball probability $\phi_{d,\chi}(h)$ measures the concentration of the functional variable χ, according to the topology defined by the semimetric.

Intuitively, the higher $\phi_{D,x^{(1)}}(h)$ in a small neighborhood of radius h, the more efficient the method will be in practice. In accordance with this observation, investigating the behavior of $\phi_{D,x^{(1)}}(h)$ when h tends to zero has been shown to be a key step for obtaining the rate of convergence of several local methods in functional data analysis (see for instance the list of references of our paper [7]). In those references, it is shown that a large range of methods are able to enjoy good rates of convergence at point $x^{(1)}$ as soon as the small ball probability function is such that Property P3 is valid for K small enough:

Property P3: there exists a positive constant C such that $\phi_{D,x^{(1)}}(h) \sim Ch^K$, *when h tends to* 0.

When P3 is satisfied, one says that the (semi-) distance D induces a *fractal topology of order K*. As an illustration in the case of functional regression, Ferraty and Vieu [2] have shown that, under quite general conditions and with the assumption that P3 is satisfied for a certain K, one reaches the near-optimal rate of pointwise convergence of the regression operator r: this latter one is given by $\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+K}}$, with β being a Lipschitz parameter quantifying the smoothness of the regression operator r, and n the number of curves in the training set (this rate of convergence is to be compared with the rate of convergence for a nonparametric multivariate regression directly based on a N-dimensional variable: $\left(\frac{\log n}{n}\right)^{\frac{p}{2p+N}}$, with p the order of differentiability of r).

Theorem 1 of [7] states that Property P3 is satisfied for the BAGIDIS semi-distance in its general form of equation (3),

$$d^B_{w_k,\lambda}(s^{(1)},s^{(2)}) = \sum_{k=0}^{N-1} w_k \left(\lambda \left| b_k^{(1)} - b_k^{(2)} \right|^2 + (1-\lambda) \left| d_k^{(1)} - d_k^{(2)} \right|^2 \right)^{1/2},$$

under quite general conditions, with K being the number of non-zero weights w_k. In other words, we have shown that the BAGIDIS semi-distance induces a *fractal topology* of order K, with K the number of features (ranks) that enter into the comparison of the two curves using this distance.

Combined with property P2, Theorem 1 of [7] ensures good performance (in terms of good rates of convergence) when using local methods relying on BAGIDIS, provided that the number K of significant features in the curves is not too large. In particular this is illustrated in [7] in the setting of a functional regression using BAGIDIS, both theoretically (Theorem 2 of that paper) and by simulation studies.

Given this, we are able now to address the particular concern about (finite sample) situations, where significant local structure of the first and second rank is accidentally permuted due to the influence of noise: how would a subsequent ”local” discrimination method face this situation? We recall our motivating simulation examples of Section 2.5 for an illustration of this.

1. Following our intuition, a good performance relies on the fact that the information content is located in a sufficiently small number K of features, whether or not permutations do occur in the dataset.
2. Moreover, the above discussions tell that the key for the successful practical application of BAGIDIS in case of permutations, is that a permutation should occur with a sufficiently high probability so that each spurious groups contains enough representatives for the local method to be efficient in the (b,d) plane. Or conversely, the dataset should be large enough for each spurious group of curves to be sufficiently populated in the observational dataset. As soon as the dataset is large enough, Theorem 1 of [7] guarantees that this will be the case.
3. As for all prediction methods that use a model (link) to associate a response to an explanatory variable, some regularity conditions (e.g. Lipschitz parameter) on the link function are needed. This is the mathematical counterpart of Assumption A1.

We finally add an important remark on whether we would also need to examine the opposite scenario: could it happen that due to the nature of BAGIDIS to need to face spurious permutations, the probability of masking the difference between two curves in the presence of noise is higher than with any other distance-based method? We illustrate this again by discussion of a prominent example: Suppose two curves are meant to be different because they have two consecutive features of a large peak followed by a smaller peak for the first curve, and vice versa for the second. In order that the presence of noise masks this and puts the curves in the same class, the amplitudes for the two consecutive peaks would need to be very close, and consequently no more detected to be significantly different. (Otherwise the information in the d_k-coefficients alone would be sufficient to discriminate the two curves). Hence, in this situation, BAGIDIS is no more sensitive to the noise than e.g. the Euclidean distance.

3.2 *Formalising the 'stability issue' of the BAGIDIS algorithm: proof of Property P2*

We now give a more formal treatment of what is behind Property P2 and its proof using properties of fractal topologies.

We recall that we consider two series $x^{(1)}$and $x^{(2)}$ valued in $\mathbb{R}^N$, corresponding to curves sampled on the regular grid $\mathbb{N}_{[1;N]}$. Their expansions in the (b,d) plane are denoted $s^{(1)} = \{(b_k^{(1)}, d_k^{(1)})\}_{k=0}^{N-1}$ and $s^{(2)} = \{(b_k^{(2)}, d_k^{(2)})\}_{k=0}^{N-1}$ with the conventional notation $b_0^{(1)} = b_0^{(2)} = 0$.

Proving P2 means showing that it is possible to have close proximity in the space of curves measured by a distance such as the Euclidean provided that there exists a small neighborhood around $s^{(1)}$ such that if $s^{(2)}$ is in this neighborhood then $d^{Eucl}(x^{(1)}, x^{(2)}) < \varepsilon$. Mathematically, Property P2 hence translates into

Property P2/math: for all $\varepsilon > 0$, there exists $K \in \mathbb{N}_{[1;N-1]}$ and $\delta > 0$ such that for the BAGIDIS distance based on K non-zero weights w_k, $d^B(s^{(1)}, s^{(2)}) < \delta$ implies $d^{Eucl}(x^{(1)}, x^{(2)}) < \varepsilon$,

which we are going to prove now.

Proof of Property P2.

We first consider the simplified version

$$d_K^B(s^{(1)}, s^{(2)}) = \sum_{k=0}^{K} \left(\left| b_k^{(1)} - b_k^{(2)} \right|^2 + \left| d_k^{(1)} - d_k^{(2)} \right|^2 \right)^{1/2}$$

as the measure of the proximity in the (b,d) plane. Recall here, that $K \leq N-1$.

If there exists $k \in \mathbb{N}_{[1;K]}$ such that $|b_k^{(1)} - b_k^{(2)}| > 1$, where 1 is the sampling step of the grid $\mathbb{N}_{[1;N]}$ on which the curve is observed, then we have

$$d_K^B(s^{(1)}, s^{(2)}) = \sum_{k=0}^{K} \left(\left| b_k^{(1)} - b_k^{(2)} \right|^2 + \left| d_k^{(1)} - d_k^{(2)} \right|^2 \right)^{1/2} \geq \sum_{k=0}^{K} |b_k^{(1)} - b_k^{(2)}| \geq 1.$$

Consequently, it is sufficient to choose $\delta < 1$, so that we have that $b_k^{(1)} = b_k^{(2)}$ for all k in $1, \ldots K$, and such that

1. the expression of the distance reduces to $d_K^B(s^{(1)}, s^{(2)}) = \sum_{k=0}^{K} |d_k^{(1)} - d_k^{(2)}|$.
2. the basis vectors $\psi_k^{(1)}$ and $\psi_k^{(2)}$ are, exactly the same up to rank K. This is because the ordered set of breakpoints $\{(b_k^{(i)})\}_{k=0}^{K}$ combined with the requirements of up-and-down shape, orthonormality and multiscale construction, allows to reconstruct the associated basis vectors $\psi_k^{(i)}$ using a top-down procedure, up to rank K.

We denote $\left\{ \psi_k^{(1)} \right\}_{k=0}^{K} = \left\{ \psi_k^{(2)} \right\}_{k=0}^{K} = \{\psi_k\}_{k=0}^{K}$, and $\hat{x}^{(1)} = \sum_{k=0}^{K} d_k^{(1)} \psi_k$ and $\hat{x}^{(2)} = \sum_{k=0}^{K} d_k^{(2)} \psi_k$. We observe:

1. Using those results and the property of energy conservation in wavelet expansions, we have
$$d^{Eucl}(\hat{x}^{(1)}, \hat{x}^{(2)}) = \sqrt{\sum_{k=0}^{K} |d_k^{(1)} - d_k^{(2)}|^2},$$
when $\delta < 1$. Consequently: for all $\varepsilon_1 > 0$, there exists $\delta_1 \in]0;1[$ such that if $d_K^B(s^{(1)}, s^{(2)}) < \delta$ then $d^{Eucl}(\hat{x}^{(1)}, \hat{x}^{(2)}) < \varepsilon_1$.
2. By construction of the BUUHWT algorithm, the energy of the signal is concentrated in the first ranks of the expansion. Thus, $\hat{x}^{(1)}$ and $\hat{x}^{(2)}$ may thus be interpreted as the wavelet approximation of $x^{(1)}$and $x^{(2)}$ under some hard thresholding rule such that $d_k = 0$ for all $k > K$. Using the results of [3], we know that such a reconstruction is mean square consistent. Consequently: for all $\varepsilon_2 > 0$, there exists K high enough such that $d^{Eucl}(\hat{x}^{(1)}, x^{(1)}) < \varepsilon_2$ and $d^{Eucl}(\hat{x}^{(2)}, x^{(2)}) < \varepsilon_2$.
3. Because of triangular inequalities, we have
$$d^{Eucl}(x^{(1)}, x^{(2)}) \leq d^{Eucl}(x^{(1)}, \hat{x}^{(1)}) + d^{Eucl}(\hat{x}^{(1)}, \hat{x}^{(2)}) + d^{Eucl}(\hat{x}^{(2)}, x^{(2)}).$$

Combining our three observations above, we have shown P2 in the special case of the simplified version of the BAGIDIS distance. However, considering now its general definition as given by equation (3), with $w_k > 0$ for $k = 0...K$, and $w_k = 0$ elsewhere, similar arguments with slightly more complex notation show again that Property P2 is valid.

We recall that in the statement of Property P2, the closeness of $s^{(1)}$ and $s^{(2)}$ is measured using the BAGIDIS semi-distance whereas the closeness of $x^{(1)}$ and $x^{(2)}$ by their Euclidean distance. This ensures that assumption A1 can be transposed in the (b,d) plane, so that local methods can be used that rely on the signatures of the curves.

Moreover, along with our proof, we showed that in order to ensure our criteria of similarity $d^{Eucl}(x^{(1)},x^{(2)}) < \varepsilon$ to be satisfied, we constrained $s^{(2)}$ to be in a small ball of maximal radius 1 around $s^{(1)}$, as computed with d^B, which we denote by $B_{d^B}(s^{(1)},1)$. As is made clear in our proof, this constraint means that we define a neighborhood in which the main features of the curves are well aligned, because the breakpoints of the curves have to be the same up to rank K.

In the now following discussion, we indicate that if we enlarge our criteria for assessing the proximity of curves so that it allows for $x^{(1)}$ and $x^{(2)}$ to be considered similar although being misaligned, then the neighborhood of $s^{(1)}$ in which we will find $s^{(2)}$ for ensuring the desired proximity of $x^{(1)}$ and $x^{(2)}$ might be larger than $B_{d^B}(s^{(1)},1)$. Enlarging our proximity criteria in such a way is desirable. In the (b,d) plane, a neighborhood larger than $B_{d^B}(s^{(1)},1)$ is a neighborhood that might include signatures $s^{(2)}$ the breakpoint component of which is not necessarily the same as for $s^{(1)}$. This translates into the fact that we might use information about series that are potentially misaligned.

3.3 The case of possible misalignments

In our proof of P2 above, we indicate the existence of a small neighborhood, i.e. the small ball $B_{d_K^B}(s^{(1)},\delta)$ around $s^{(1)}$, which has radius δ, and which is such that if $s^{(2)}$ is in $B_{d_K^B}(s^{(1)},\delta)$ then the related curves $x^{(1)}$ and $x^{(2)}$ are similar. In our proof, we use an upper bound $\delta < 1$. This bound defines a neighborhood such that $s^{(2)}$ must have the same breakpoints as $s^{(1)}$ up to rank K. This bound appears because we measure the similarity of $x^{(1)}$ and $x^{(2)}$ using the Euclidean distance. Indeed, it ensures that the K main features of $x^{(1)}$ and $x^{(2)}$ are well aligned, which is necessary for the Euclidean distance to detect their closeness.

However, our method has been designed with the aim that the closeness of $s^{(1)}$ and $s^{(2)}$ in the (b,d) plane might also reflect a "visual proximity" of $x^{(1)}$ and $x^{(2)}$ even when the series are misaligned. Therefore, we would ideally like our proof to have the following extension:

Property P2/ideally : Let D be a distance measure between $x^{(1)}$ and $x^{(2)}$ that is relevant even in case of possible misalignment between $x^{(1)}$ and $x^{(2)}$. Then, for all $\varepsilon > 0$, there exists some neighborhood V centered on $s^{(1)}$ such that if $s^{(2)}$ is in V, then $D(x^{(1)},x^{(2)}) < \varepsilon$.

In this case, V would not necessarily be smaller than $B_{d_K^B}(s^{(1)},1)$. Nevertheless, as far as we know, there does not exist a distance measure D that is able to measure the similarity of curves of which the sharp local patterns might be misaligned – which is precisely what motivated us to propose the BAGIDIS semi-distance.

We need thus to define another way to assess that the series $x^{(1)}$ and $x^{(2)}$ are close to each other when $s^{(1)}$ and $s^{(2)}$ are close enough. We propose the following:

We will say that $x^{(1)}$ and $x^{(2)}$ are "globally similar" to each other if

C1 Their "global shape" are similar - this notion is related to the succession of level changes in the series ; we define below how to quantify this intuitive notion.
C2 The main level changes of $x^{(1)}$ are located at abscissas that are not too distant from the abscissas of their counterpart in $x^{(2)}$.
C3 The amplitude of the main level changes in $x^{(1)}$ are not too different from the amplitude of their counterpart in $x^{(2)}$.

Given this, we want to show that

Property P2/extended: There exists some neighborhood V of $s^{(1)}$ such that if $s^{(2)}$ is in V, then $x^{(1)}$ and $x^{(2)}$ are "globally similar".

We say that $x^{(1)}$ and $x^{(2)}$ are "similar in their global shape" if their associated Unbalanced Haar wavelet bases are "similar in structure" up to rank K. By "similar in structure" we mean that the hierarchical trees associated to each of the wavelet bases are identical up to rank K. We encode this hierarchical tree in a way that is common-place in wavelet analysis, but adapted to Unbalanced Haar, which we explain through the following example.

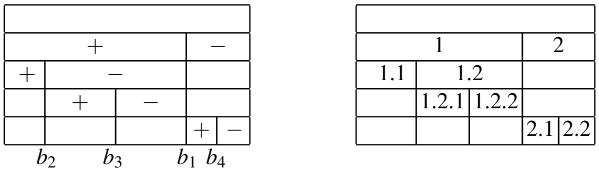

Fig. 4 Basis $B^{(1)} = \{\psi_0^{(1)}, \psi_1^{(1)}, \dots, \psi_4^{(1)}\}$ (left), and its associated hierarchical tree $T^{(1)}$ (right).

In Figure 4, on the left hand side is a schematically representation of a basis $B^{(1)} = \{\psi_0^{(1)}, \psi_1^{(1)}, \dots, \psi_4^{(1)}\}$, where the k^{th} element $\psi_k^{(1)}$ is placed on the $(k+1)^{th}$ row, top-down. Here "+" is indicating the positive part of the wavelet and "-" is indicating its negative part. The collection of $\{b_k^{(1)}\}$ denotes the associated breakpoints - it uniquely encodes the associated hierarchical tree $T^{(1)}$ displayed on the right hand side, where we use a notation borrowed from wavelet or regression trees to highlight the implicit hierarchy of splits.

In this case, the hierarchical tree $T^{(1)}$ up to rank $K = 4$ is

$$T_4^{(1)} = \{(1,2);(1.1,1.2);(1.2.1,1.2.2);(2.1,2.2)\}.$$

The k^{th} wavelet basis vector $\psi_k^{(1)}$ is thus associated to the k^{th} pair in the hierarchical tree ($k = 1,\ldots,K$). The elements of this pair are sequences of digits, where the last digit on the left is 1 to indicate that it refers to the positive part of the wavelet at rank k, and the last digit on the right is 2 to indicate that it refers to its negative part. The first digits of each element of the pair are the same and refer to the "block" of the wavelet structure at larger scale (=smaller rank) in which the wavelet $\psi_k^{(1)}$ takes place. This tree structure reveals where the successive wavelets of the basis are located with respect to each other, without indication of the precise extension of their support, nor the precise location of their breakpoint. Thereby it characterizes the "global shape" of the series $x^{(1)}$ with which the basis $B^{(1)}$ is associated. For instance, our tree $T_4^{(1)}$ differs from the tree $T_4^{(2)} = \{(1,2);(2.1,2.2);(1.1,1.2);(1.2.1,1.2.2)\}$, associated to the basis $B^{(2)}$ in Figure 5 below,

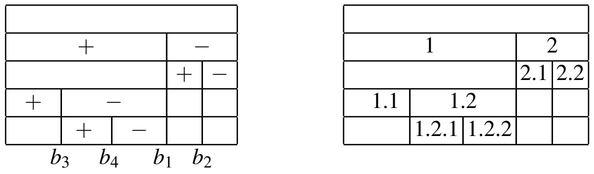

Fig. 5 Basis $B^{(2)} = \{\psi_0^{(2)},\psi_1^{(2)},\ldots,\psi_4^{(2)}\}$ (left), and its associated hierarchical tree $T^{(2)}$ (right).

because the latter indicates a higher relative importance of a pattern on the right side of the series $x^{(2)}$ than in $x^{(1)}$ - this is because the element $(2.1,2.2)$ appears only at rank 4 in $T_4^{(1)}$ while it appears at rank 2 in $T_4^{(2)}$. On the other hand the tree $T_4^{(3)}$ associated to the $B^{(3)}$ in Figure 6 below

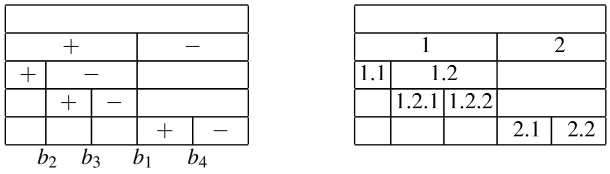

Fig. 6 Basis $B^{(3)} = \{\psi_0^{(3)},\psi_1^{(3)},\ldots,\psi_4^{(3)}\}$ (left), and its associated hierarchical tree $T^{(3)}$ (right).

is exactly the same as $T_4^{(1)}$ up to rank 4, even though the patterns this basis supports are not perfectly aligned with the ones that basis $B^{(1)}$ supports.

We now prove *Property P2/extended*.

Proof.

C2-C3: It is easy to see that C2 and C3 will be satisfied as soon as we take $s^{(2)}$ in a small enough neighborhood V defined by suitably chosen values δ_k^d and δ_k^b, $k = 1,...,K$, such that $|b_k^{(1)} - b_k^{(2)}| < \delta_k^b$ and $|d_k^{(1)} - d_k^{(2)}| < \delta_k^d$, for $k < K$.

C1: We first observe that the hierarchical tree $T^{(i)}$ (up to rank K) is actually determined from the signature $s^{(i)}$ of a series (up to rank K) by the permutation $P^{(i)}$ we need to apply to the vector $(b_1^{(i)}, b_2^{(i)}, b_3^{(i)}, ..., b_K^{(i)})$ in order to sort its elements according to their value (i.e. their position along the breakpoints axis). For instance, for $K = 5$, if $b_2^{(i)} < b_1^{(i)} < b_5^{(i)} < b_4^{(i)} < b_3^{(i)}$, we have

$$(b_2^{(i)}, b_1^{(i)}, b_5^{(i)}, b_4^{(i)}, b_3^{(i)}) = select[(2,1,5,4,3)](b_1^{(i)}, b_2^{(i)}, b_3^{(i)}, b_4^{(i)}, b_5^{(i)}),$$

where $P^{(i)} = (2,1,5,4,3)$ (and where ”select” is used to just denote the mapping of the indices according to this permutation here). The link between $P^{(i)}$ and the hierarchical structure of the wavelet partition arises because we can uniquely reconstruct the basis vectors in a top-down procedure, by using the information of the up-and-down shapes of the wavelets, the location of their breakpoints and their orthonormality.
Then, in order for C1 to be satisfied, it is sufficient to define V as the neighborhood of $s^{(1)}$such that the permutation $P_K^{(2)}$ associated to the K first elements of any $s^{(2)}$ in V is the same as the permutation $P_K^{(1)}$ associated to the K first elements of $s^{(1)}$.

We have thus proved *Property P2/extended*, as it is sufficient to combine the above constraints to define a neighborhood Varound $s^{(1)}$ such that if $s^{(2)}$ is in V, then $x^{(1)}$ and $x^{(2)}$ are “globally similar”. As expected, we note that this neighborhood V might possibly be larger than $B_{d_K^B}(s^{(1)}, 1)$, so that it might contain the signatures of curves with main features some of which are misaligned.

This last point is important as it gives support for the large scope of the method and for its performance in investigating datasets of curves of which the main features might be misaligned although similar. It is clearly seen in the practical examples illustrated in our previous work, that the local methods (k-NN, kernels, MDS) that we used to process the set of BAGIDIS semi-distances computed on our datasets have effectively used neighborhoods large enough to include some breakpoint variations. This ability of using the information related to a misaligned feature is precisely the fact that makes BAGIDIS better than competitors in case of misaligned sharp patterns.

Conclusion: By the theoretical property (P2) we have shown that BAGIDIS achieves performances that are consistent with the ones obtained with the Euclidean distance in a local algorithm (while reducing the dimensionality of the problem from N sampled features of a given curve to $K < N$ significant or essential features). On the other hand, our subsequent discussion has shown that BAGIDIS may achieve performances that are superior to the Euclidean distance: local methods operating on the signatures of our curves in the (b,d) plane might be based upon neighborhoods which contain curves that are similar although misaligned.
Hence by this note we have provided the theoretical argument behind what we observed in previous work of ours on a competitive performance of our method when dealing with curves that have possibly misaligned sharp features.

Acknowledgements The first author is particularly grateful to EDF and A. Antoniadis, X. Brossat and J.-M. Poggi for having been given the opportunity to present the BAGIDIS methodology at the generously sponsored WIPFOR 2013 workshop in Paris.
Both authors would also like to acknowledge financial support from the IAP research network grants P06/03 and P07/06 of the Belgian government (Belgian Science Policy).
Finally, useful comments of Melvin Varughese and two anonymous referees have helped to improve the presentation of this note.

References

1. Aneiros-Pérez, G., Cardot, H., Estévez-Pérez, G., Vieu, P.: Maximum ozone concentration forecasting by functional non-parametric approaches. Environmetrics **15**(7), 675–685 (2004)
2. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis: Theory and Practice. Springer Series in Statistics. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
3. Fryzlewicz, P.: Unbalanced Haar technique for non parametric function estimation. Journal of the American Statistical Association **102**(480), 1318–1327 (2007)
4. Giorgino, T.: Computing and visualizing dynamic time warping alignments in R: The dtw package. Journal of Statistical Software **7**(31), 1–24 (2009)
5. Girardi, M., Sweldens, W.: 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**(4), 457–474 (1997)
6. Jolliffe, I.: Principal Component Analysis (Second Edition). Springer Series in Statistics. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2002)
7. Timmermans, C., Delsol, L., von Sachs, R.: Using Bagidis in nonparametric functional data analysis: predicting from curves with sharp local features. Journal of Multivariate Analysis **115**, 421–444 (2013)
8. Timmermans, C., Fryzlewicz, P.: Shah: Shape-adaptive haar wavelet transform for images with application to classification. under revision (2012). URL http://www.uclouvain.be/en-369695.html. ISBA Discussion Paper 2012-15, Université catholique de Louvain
9. Timmermans, C., von Sachs, R.: BAGIDIS: Statistically investigating curves with sharp local patterns using a new functional measure of dissimilarity. under revision (2013). URL http://www.uclouvain.be/en-369695.html. ISBA Discussion Paper 2013-31, Université catholique de Louvain
10. Timmermans, C., de Tullio, P., Lambert, V., Frdrich, M., Rousseau, R., von Sachs, R.: 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 (2012)