INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

1030

# BAGIDIS, A NEW METHOD FOR STATISTICAL ANALYSIS OF DIFFERENCES BETWEEN CURVES WITH SHARP DISCONTINUITIES

TIMMERMANS, C. and R. von SACHS,

# BAGIDIS, a New Method for Statistical Analysis of Differences Between Curves with Sharp Discontinuities

Catherine TIMMERMANS and Rainer VON SACHS

Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA),
Université Catholique de Louvain, Voie du Roman Pays 20, BE-1348 Louvain-la-Neuve, Belgium.
Correspondence: catherine.timmermans@uclouvain.be

June 21, 2010

**Abstract**

In this paper, we introduce a functional wavelet based semi-distance for comparing curves with sharp patterns that might not be well aligned from one curve to another. 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, which proofs highly useful when used together with clustering algorithms or visualization method. We also develop statistical tools for detecting and localizing differences between groups of curves using this semi-distance. Finally, we apply this methodology to H-NMR spectrometric curves and solar irradiance time series.

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

# 1 INTRODUCTION

This paper presents the BAGIDIS methodology, which has been designed to compare and analyze sets of curves that are characterized by some sharp discontinuities. Such kind of curves happen to be dealt with in a large variety of scientific areas. Spectrometric curves, like the one presented in Figure 6, are one typical example that will be further described later on.

The core of the BAGIDIS methodology is the definition of a data-driven, flexible and highly adaptive semi-distance, that is able to take into account both vertical and horizontal variations of the patterns observed in discretized measurements of curves. This semi-distance can directly be included within clustering algorithms or visualisation methods that are based on dissimilarity matrices. Besides, its use is coupled with some statistical diagnostic tools, that can help us to get an insight into the way in which curves actually differ.

This paper is organized as follows. Section 2 presents classical ways for comparing curves and highlights their weaknesses when applied to curves with sharp discontinuities. The need for a new method is thus emphasized. Section 3 describes what the BAGIDIS methodology does propose to tackle this need. It presents the main ideas of the method. Section 4 is devoted to the definition of the semi-distance that is at the core of the BAGIDIS methodology. Then, section 5 introduces some extensions of the method, and some statistical tools for investigating datasets using BAGIDIS: measures of the mean and variability of datasets, an ANOVA-like test within the BAGIDIS framework, a test for differences between two groups of curves ... Two real datasets are studied, in section 6. We show the consistency of our method compared to classical competitors and its superiority when horizontal differences between the curves have to be taken into account. Finally, section 7 concludes with some perspectives for further developments.

# 2 MOTIVATION: WEAKNESSES OF CLASSICAL METHODS AND THE NEED FOR SOMETHING NEW

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 Ferraty and Vieu (2006), we will say that $d$ is a semi-distance on some space $\mathcal{F}$ as soon as

- $\forall \mathbf{x} \in \mathcal{F}, \quad d(\mathbf{x}, \mathbf{x}) = 0$

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

A semi-distance is thus defined in the same way as a distance, except that $d(\mathbf{x}^i, \mathbf{x}^j) = 0 \not\Longrightarrow \mathbf{x}^i = \mathbf{x}^j$. Semi-distances are often used when one is interested in comparing the shapes of some groups of curves, but not in comparing their mean level.

**Classical $l_p$ Distances and their Principal Components-based Extension: Measuring point-to-point Distances along the *Ordinate* Axis.** Most of the classical distances ($l_1$, $l_2$, $l_\infty$, ...) compare curves separately at each point of measurement (i.e. at each abscissa), along the ordinate axis. The principal components-based distance (Jolliffe 2002, for instance) 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. As a result, curves that are similar except for a given vertical variation are measured relatively close to each other while curves that are similar apart from the same amount of variation but horizontally are measured distant. This behavior is illustrated in Figure 1. It is not satisfying. The problem is that the ordering of the series measurements is not taken into account so that the evolutions of two series cannot be compared.

**Functional Semi-Distances: Taking into account the Notion of Neighborhood in point-to-point Comparison.** 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 curve. In other words, it includes 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). Those semi-distances have been shown useful in numerous problems (Ferraty and Vieu 2006). However, they happen to fail when dealing with curves with local sharp discontinuities that might not be well aligned from one curve to another one. This is illustrated in Figure 1. Besides, those methods, as well as the functional extension of the principal components-based distance, rely on a smoothing of the data, which is generally problematic for curves with abrupt discontinuities.

**Wavelet-based Distances: Comparing the Coefficients of Well-suited Basis Functions Expansions.** In contrast to the afore-mentioned point-to-point definitions of distances between

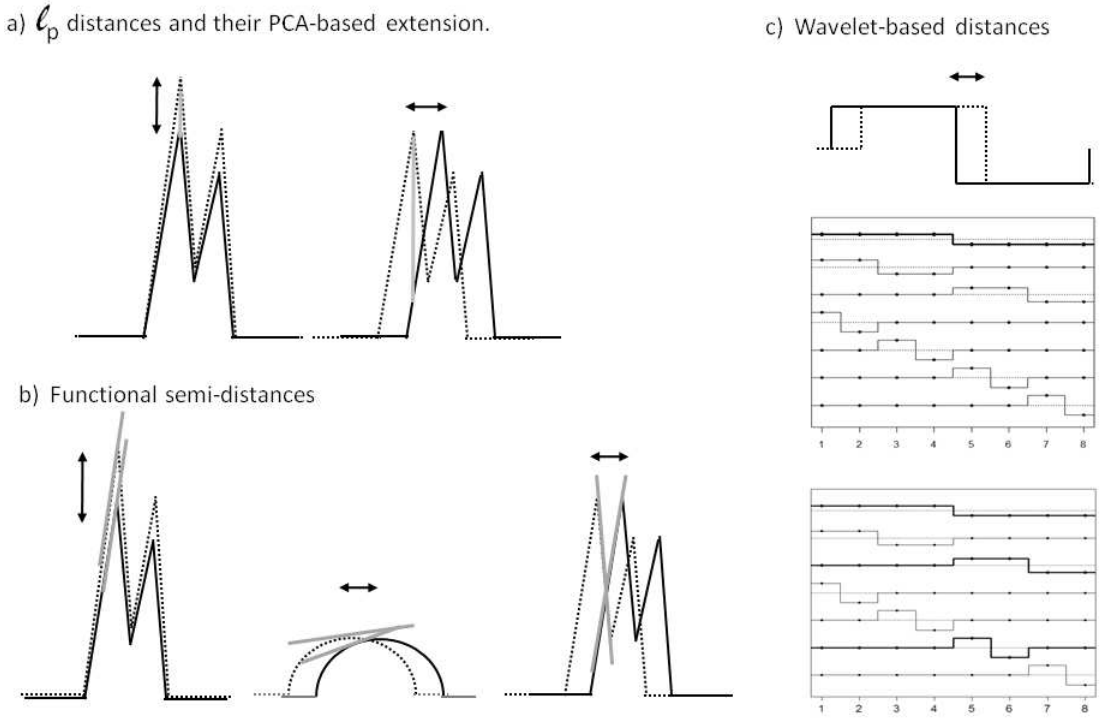


Figure 1: **Schematic illustration of the difficulty for classical methods to take into account horizontal variations of curves.** In each graph, two slightly different curves (solid and dotted curves) are compared. *a)* $l_p$ distances and PCA-based distances compare curves at each point of measurement, 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 gray in (b). Besides, classical functional methods usually rely on a smoothing, unadapted to sharp patterns. *c)* Classical 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 the Haar wavelet-basis expansions of two shifted step series. In the upper representation of the Haar-basis, the only basis vector that is associated with a non-zero coefficient for the expansion of the solid series is highlighted in bold. Three non-zero coefficients are required for encoding the dotted series in the Haar-basis. The corresponding basis vectors are highlighted in bold in the lower representation of the Haar-basis.

curves, wavelet-based distances rely on the expansion of the series in a well-designed basis. The projection of the series onto this basis, i.e their *detail* coefficient, 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 dimensions by separating "significant" from uninformative coefficients across curves with respect to some class membership. A wavelet basis is actually made of a fixed pattern that is 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 the position and amplitude of discontinuities of

the series. However, the description of a given discontinuity is highly sensitive to its location in the series. This is called the dyadic restriction of classical wavelets. As a consequence of this feature, it may become difficult to detect the closeness of series having a similar discontinuity that is only slightly shifted. It is illustrated in Figure 1.

As described in the previous paragraphs, there is thus a need for a way to efficiently measure distances between curves that are characterized by some sharp peaks or discontinuities, those discontinuities being possibly subject to both a modification in amplitude and a shift in location.

## 3 WHAT THE BAGIDIS METHODOLOGY DOES PROPOSE

The methodology we developed aims at answering the above-stated need for a method able to detect the closeness of curves whose significant sharp features might not be well aligned. Our approach is based upon the definition of a semi-distance that is functional, in the sense that the ordering of the data is explicitly taken into account, and wavelet-based, in the sense that it relies on a basis function expansion in which positions and amplitudes of a pattern are encoded. However, our method will overcome the dyadic restriction that is attached with classical wavelet expansions, and will not require any preliminary smoothing of the data. Furthermore, a major originality of the method is that it relies on projections on basis functions that are different from one series to another.

The main ideas of the method are described as follows:

**Preliminary Observation.** We consider series that are made of regularly spaced observations of a curve. We note that patterns in a series can be described as a set of level changes.

**Finding 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 (jumps, peaks, troughs) and level changes affecting a large number of data (discontinuities of the mean level). From this point of view, Unbalanced Haar wavelet bases are good candidates for our expansion.

This family of bases, introduced by Girardi and Sweldens (1997) as a way to circumvent the dyadic restriction of classical wavelets, is described in subsection 4.1 of this paper.

**Taking Advantage of the Hierarchy of those Bases.** Given this, we will define a semi-distance, that 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 4.2 of this paper is devoted to the definition of such a semi-distance.

A subsequent interest lies in obtaining some information about the relative importance of horizontal and vertical variations, and about their localization. Furthermore, it will be our goal to statistically diagnose whether groups of curves do actually differ and how. Section 5 is dedicated to those questions.

# 4 THE CORE OF THE BAGIDIS METHODOLOGY

## 4.1 Finding an Optimal Basis for Each Curve

Given a set of $M$ series $\mathbf{x}^{(i)}$ in $\mathbb{R}^N$, $i = 1 \dots M$, 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.

**A Definition of the Unbalanced Haar Wavelet Bases.** Unbalanced Haar wavelet bases 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 Fryzlewicz (2007), the general mathematical expression of one of those Haar-like wavelets is given by

$$\boldsymbol{\phi}_{e,b,s}(t) = \left(\frac{1}{b-s+1} - \frac{1}{e-s+1}\right)^{1/2} \mathbf{1}_{s \le t \le b} + \left(\frac{1}{e-b} - \frac{1}{e-s+1}\right)^{1/2} \mathbf{1}_{b+1 \le t \le e},$$

where $t = 1 \dots N$ is a discrete index along the abscissa axis, and where $s < b < e$ are values along this axis that determine the particular shape of one wavelet ($s$, $b$ and $e$ stand for *start*, *breakpoint* and *end* respectively). Each wavelet $\boldsymbol{\phi}_{e,b,s}(t)$ 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 $\boldsymbol{\phi}_{e,b,s}(t)$ encodes the importance of the related level change in the series. Different choices of $N-1$ sets of values $s$, $b$ and $e$ leading to orthonormal wavelets define a whole family of bases $\{\boldsymbol{\phi}_k\}_{k=0\dots N-1}$.

**The Basis Pursuit Algorithm and the Property of Hierarchy.** In 2007, Fryzlewicz (2007) proposed an algorithm for building the unbalanced Haar wavelet basis $\{\boldsymbol{\phi}_k\}_{k=0\dots N-1}$ 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 following 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.

**An example of Bottom-Up Unbalanced Haar Wavelet Expansion.** Figure 2, *left*, shows the BUUHWT expansion obtained for one particular series. As we hoped for, the first non-constant vectors support the largest discontinuities of the series and encode therefore the highest peak of the series. We observe it by looking at the location of the discontinuity points $b$ between positive and negative parts of the wavelets. 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.

**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:

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

where the coefficients $d_k^{(i)}$ are the projections of $\mathbf{x}^{(i)}$ on the corresponding basis vectors $\boldsymbol{\psi}_k^{(i)}$ (i.e. the *detail* coefficients) and where the set of vectors $\{\boldsymbol{\psi}_k^{(i)}\}_{k=0\dots N-1}$ is the Unbalanced Haar

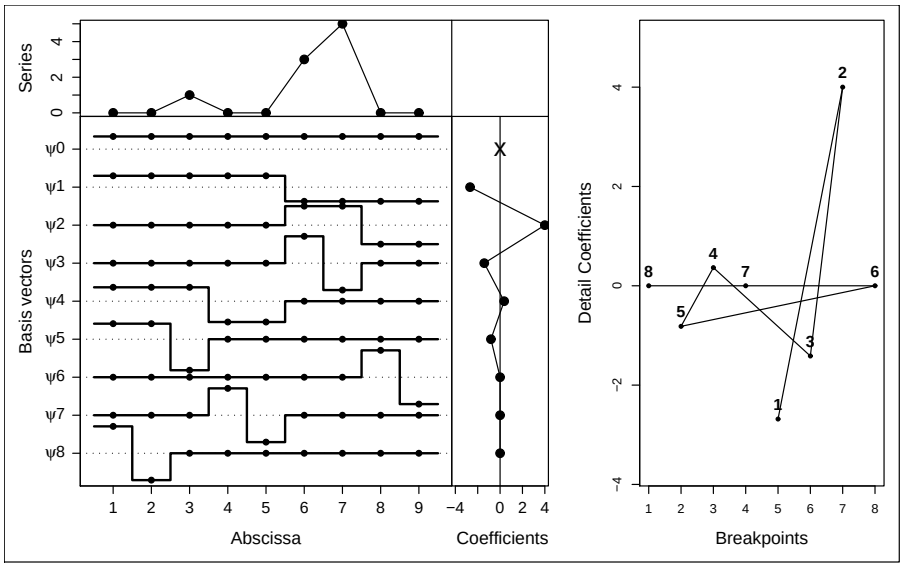


Figure 2: *Left:* **Illustration of a BUUHWT expansion.** The series we consider is plotted in the upper part of the figure. The corresponding abscissa axis is common for that graph and for the main graph, so that it is located at the bottom of that one. The main part of the figure shows the basis vectors of the Unbalanced Haar wavelet basis that is best suited to the series (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 $\boldsymbol{\psi}_0$ 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.

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 $b$ of the wavelet $\boldsymbol{\psi}_k^{(i)}$, $k = 1 \dots N-1$, as defined in equation (1). An interesting property of the basis $\{\boldsymbol{\psi}_k^{(i)}\}_{k=0\dots N-1}$, that has been proved by Fryzlewicz (2007), is the following:

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

Consequently, the set of pairs $(b_k^{(i)}, d_k^{(i)})_{k=1\dots N-1}$ 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 leads us to the possibility of representing 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*.

## 4.2 Defining a Semi-Distance by taking Advantage of the Hierarchy of the BUUHWT Expansions

We aim at defining a measure of 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.

**Measuring the dissimilarity as a weighted sum of partial distances in the *b-d* plane.** We propose to measure the dissimilarity between the series $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ as a weighted sum of partial dissimilarities evaluated rank by rank:

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

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 non-negative weight function of the rank $k$. Note that we do not consider the rank $k = 0$ in our dissimilarity measure (equation (1)), as we are mainly interested in comparing the structures of the series rather than their main level.

**Property: This measure of dissimilarity is a semi-distance.** (see Appendix A for a proof).

**A Decreasing Weight Function.** According to the hierarchy of the BUUHWT expansion, the weight function $w_k$ should be decreasing as the value of the partial dissimilarity $\delta_k = \left\| \mathbf{y}_k^{(1)} - \mathbf{y}_k^{(2)} \right\|_p$ at rank $k$ should affect the global dissimilarity measure as much as the compared patterns carry major features of the series. In this study, the weight function we consider is defined by

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

as it allows for associating a large weight to the first ranks of the hierarchy and a rapidly decreasing weight at the end of the hierarchy, which is empirically what we expect. This weight function could actually be optimized, which can easily be done in case of a supervised problem. In case of the dissimilarity being used in an unsupervised problem, the optimization of $w_k$ is a bit more elaborate and a forthcoming paper shall be devoted to that question.

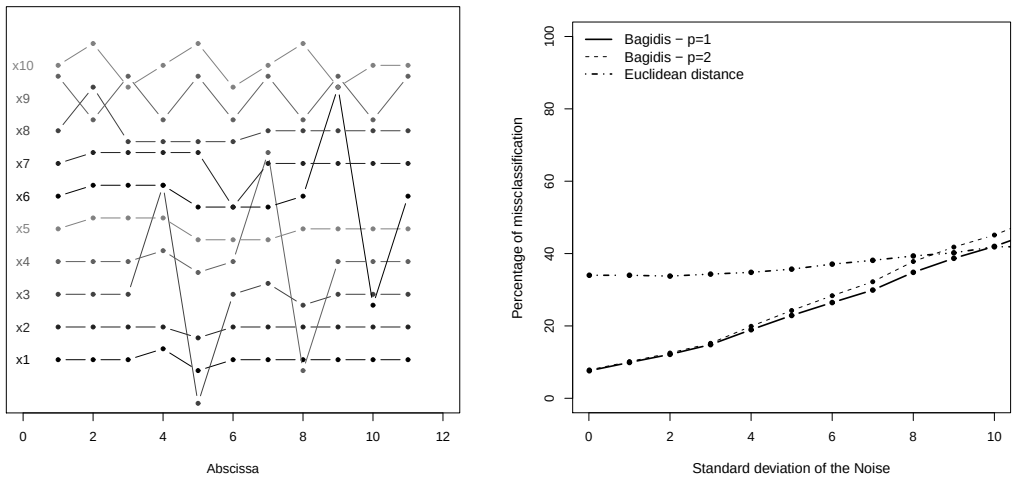


Figure 3: *Left*: The **10 main series** we use for supervised classification. *Right*: **Results of supervised classification:** Percentage of misclassification as a function of the standard deviation of the Gaussian noise used for the simulation of the series. This percentage is calculated for a classification using BAGIDIS with a partial distance defined as 1-norm ($p = 1$) and 2-norm ($p = 2$). Those results are compared with the ones we obtain using the Euclidean distance between the curves. The minimum standard deviation tested for the noise if fixed at 0.01. The maximum standard deviation tested (*Standard deviation*= 10) is 100% of the value of the smallest peak in the noisefree data and 10% of the highest. Up to that noise level, BAGIDIS used with $p = 1$ or $p = 2$ is more efficient than the euclidean distance.

**An Example of Supervised Classification.** In order to illustrate the potential of the so-defined semi-distance, the following test is proposed: 10 main series are defined, each of these being of length 11 and the features they are made of having height from 10 to 100. Those series are presented in Figure 3, *left*. Each of those 10 main series leads to a family of 1000 simulated noisy series as follows: amongst them, 20% correspond to the initial series that is shifted one step to the right, 20% of them are the initial series shifted one step to the left, 20% are amplified by a factor 1.25, 20% by a factor 0.75, the remaining 20% correspond to the initial series. Besides, each simulated series is affected with an additive Gaussian white noise with a given variance.

Further, we evaluate the semi-distance of those 10000 series to the 10 main series. 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 shown in Figure 3, *right*, show that BAGIDIS used with a partial distance defined as 1-norm ($p = 1$) or 2-norm ($p = 2$) performs very well compared to the Euclidean distance ($l_2$ distance in the original coordinates of the series), up to the maximum level of noise tested, whose standard deviation is 100% of the height of the smaller features of the series and 10% of the highest.

# 5 INVESTIGATING A DATASET USING BAGIDIS

## 5.1 Mean and Variability of a Dataset

**Determining the Average Curve of a Dataset in the *b*-*d* Plane.** Classically, when we want to estimate the average curve of a functional dataset, we rely on a point-to-point estimate such as

$$\bar{\mathbf{x}}(t) = \frac{1}{M}\sum_{i=1}^{M} \mathbf{x}^{(i)}(t),$$

where $M$ is the number of curves we want to average, where $t$ is an index along the abscissa axis and where $\mathbf{x}^{(i)}(t)$ is either the series of measurements itself or a smooth version of it. For the same reason as before, this is not necessarily pertinent when dealing with curves with sharp discontinuities. In a view to overcome that difficulty, we propose to define the average curve in the $b$-$d$ plane as

$$\bar{\mathbf{x}}^{(b-d)} \equiv (\mathbf{b}^{(0)}, \mathbf{d}^{(0)}) \quad \text{where} \quad b_k^{(0)} = \frac{1}{M}\sum_{i=1}^{M} b_k^{(i)} \quad \text{and} \quad d_k^{(0)} = \frac{1}{M}\sum_{i=1}^{M} d_k^{(i)}.$$

In such a way, at each rank $k$, $\bar{\mathbf{x}}_k^{(b-d)}$ is the center of gravity of the points $\{(b_k^{(i)}, d_k^{(i)})\}_{i=1\ldots M}$ characterizing the curves we want to average at rank $k$. The average curve $(\mathbf{b}^{(0)}, \mathbf{d}^{(0)})$ can then be easily displayed and interpreted in the $b$-$d$ plane. However, we have to notice that there is no obvious way to get back from that average curve 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 $\mathbf{b}^{(0)}$.

**Measuring Variability around the Average Curve.** Once we have a way to compute the average curve of a dataset, we have a way to quantify the variability of a dataset in line with the BAGIDIS distance:

$$V^2 = \frac{1}{M}\sum_{i=1}^{M} d_p(\mathbf{x}^{(i)}, \bar{\mathbf{x}}^{(b-d)}) = \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{N-1} w_k \left\| \mathbf{y}_k^{(i)} - \mathbf{y}_k^{(0)} \right\|_p,$$

with $\mathbf{y}_k^{(0)} = (b_k^{(0)}, d_k^{(0)})$ and $\mathbf{y}_k^{(i)} = (b_k^{(i)}, d_k^{(i)})$.

## 5.2 Testing for Differences between Groups of Curves

**An ANOVA within the BAGIDIS Framework.** Having got a way to quantify the variability of set of curves around an average curve, we have everything at hand in order to perform an ANOVA-like test. Suppose we have $M$ curves coming from $G$ groups, whose empirical group-mean curves have been estimated as $\bar{\mathbf{x}}^{(g)} \equiv (\mathbf{b}^{(g)}, \mathbf{d}^{(g)})$, $g = 1 \dots G$, and suppose the global empirical mean curve has been estimated as $\bar{\mathbf{x}}^{(0)} \equiv (\mathbf{b}^{(0)}, \mathbf{d}^{(0)})$. Then if we assume that the distances of each group of curves around their group mean are normally distributed with equal variance in each group, we can compute the ANOVA F-Ratio as

$$F = \frac{\frac{1}{G-1} \sum_{g=1}^{G} n_g \, d_p^2(\bar{\mathbf{x}}^{(g)} \bar{\mathbf{x}}^{(0)})}{\frac{1}{M-G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} d_p^2(\mathbf{x}^{(i)}, \bar{\mathbf{x}}^{(g)})},$$

where $n_g$ is the number of curves in group $g$, $g = 1 \dots G$. This ANOVA F-ratio follows an $F_{G-1,M-G}$ under the null hypothesis that all the group-means are equal. Figure 4 shows the statistical evolution of the p-value of such a BAGIDIS ANOVA F-test for different levels of noise, on a simulated example with 3 groups of curves affected by a vertical gaussian noise and a random horizontal shift. The difference between the groups of curves is detected more than 95% of the times when the standard deviation of the noise 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. However, note here 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), that could easily be applied here.

**Investigating how two Groups of Curves do actually Differ by Determining their Relative Position 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

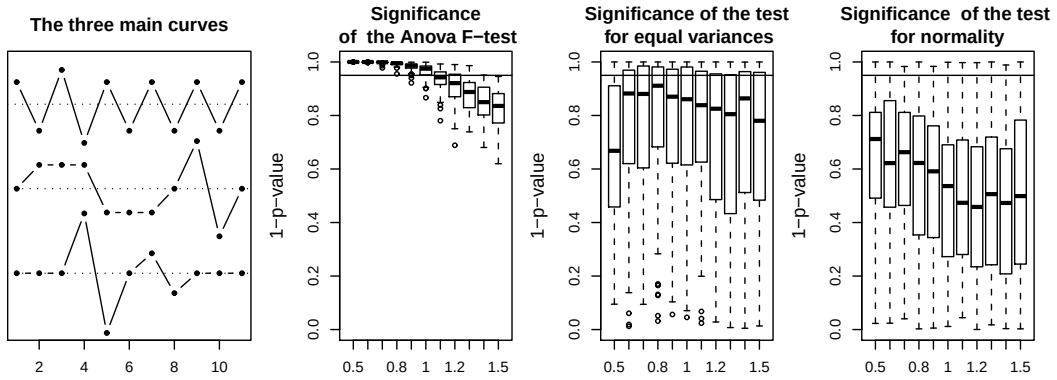


Figure 4: *From left to right:* **The three mother curves for the simulation study.** Each curve is repeated 50 times, and affected by an additive gaussian white noise with a given standard deviation. Each series is also affected by a horizontal shift, randomly chosen amongst -1, 0 and 1. **Boxplots of the significances (= 1− *p-values*) of the Bagidis Anova F-test for each tested value of the noise.** A Bagidis Anova is performed on those 150 curves and the p-value is noted. This test is repeated 100 times for each tested value of the standard deviation. The standard deviation of each mother series in 1, and the standard deviation of the gaussian noise varies from 0.5 to 1.5, which is 50% of the height of the highest peak. **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.**

to a given distance or semi-distance $d$. 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:

**Test 1:** $H_0$: $d(G1, G1) = d(G1, G2)$; $H_1$: $d(G1, G1) < d(G1, G2)$.

**Test 2:** $H_0$: $d(G2, G2) = d(G1, G2)$; $H_1$: $d(G2, G2) < d(G1, G2)$.

If the distances are distributed normally around their group-mean, and have equal variance in each case, this can easily be done using a student-t test. Note that the assumptions of equality of variances can easily 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 5. Only if both t-tests significantly

reject their null hypothesis are the two groups statistically different in mean. If only one of the t-test rejects its null hypothesis, it means that both groups vary around approximately the same mean, but with a different variability. One group is then "included" within the other one.

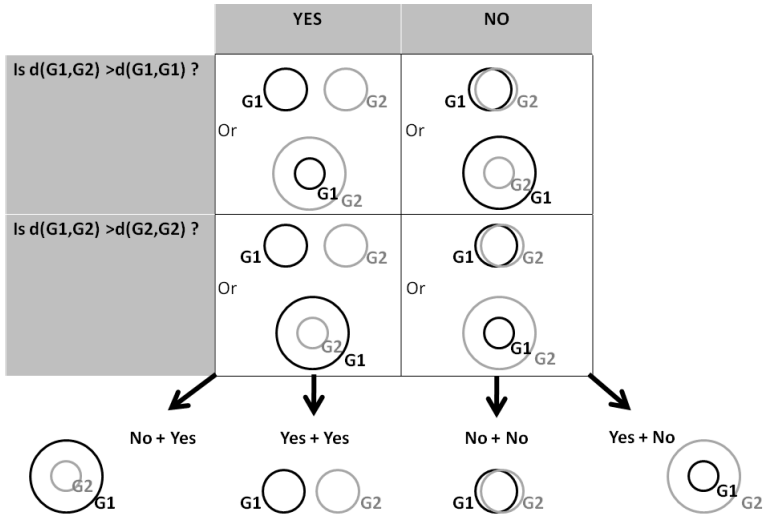


Figure 5: *Left:* **Summary table for the interpretation of the double t-test for the relative position of two groups in space,** according to a given distance or semi-distance.

### 5.3 Being Flexible with respect to Scaling Effects

Our distance aims at taking into account both the variations along the abscissa axis and along the ordinate axis. Of course, the dissimilarities we are going to detect using expression (1) are directly dependent on the scale of measurements along each of those axes. In order to take into account that sensitivity, we propose to introduce a parameter $\lambda \in [0,1]$ in our expression of the distance, that balances between the differences of the details and the differences of the breakpoints:

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

This parameter $\lambda$ actually corresponds to a scaling of the $b$-$d$ plane, and hence also of the original units of the problem. One strength of this approach is that, if $\lambda$ is chosen to be optimal according to a given criterion, this method should always lead to the same relative dissimilarities between the series.

Looking at how the semi-distances evolve when $\lambda$ varies could also be used as a diagnostic tool in itself. In particular, looking at the semi-distances one obtains by setting $\lambda$ at its extreme values 0 and 1 allows to diagnose the effects of the two components of our measure: differences in the details and in the breakpoints.

### 5.4 Dealing with Long Series

**Avoiding Feature Confusion by Using a Sliding Distance.** 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 could be problematic as feature confusion could occur in the BUUHWT expansions of the different series. In that case, it appears thus 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 $\Delta$. For each localization of that window, we obtain the BUUHWT expansion of the so-defined subseries and we compute the semi-distance of expression (1) or expression (2). One can then take the mean of those local semi-distances so as to obtain a global measure of dissimilarity.

**Localizing Differences between Long Series.** 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 (as seen using the distance of equation (1) or equation (2)) 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.

For sure, the choice of the length $\Delta$ of the windows should be problem-dependent: $\Delta$ 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 $\Delta$ should actually be provided.

## 6 REAL DATA ANALYSES USING BAGIDIS

Two real datasets are being investigated in this section, illustrating two different cases that Bagidis could have to deal with. Both involve significant sharp features. The first dataset is

made of spectrometric curves that have to be compared very finely, as the differences between them are very small. This is a typical example where the use of BAGIDIS should be helpful for investigating the data. For the second dataset of several times series, we will study the similarity of their dynamics. Unlike the spectrometric curves, this dataset is highly noisy and the large time step of the measurements will not allow us for capturing significant shifts of the patterns. The goal here is thus to illustrate how BAGIDIS behaves in a context where it is not necessary to consider the specific feature of this method - considering the horizontal variations of the patterns.

### 6.1 Detecting Fine Differences between Spectrometric Curves

**The Dataset and the Experimental Design.** The dataset we consider is made of 24 spectra of biological serum, that have been acquired using H-NMR spectroscopy (Laboratoire de Chimie Pharmaceutique, Université de Liège). One example of such a spectrum is displayed in Figure 6. Those data have been collected in the framework of a study of the reproducibility of the acquisition and preprocessing of H-NMR data in the field of metabonomics (Dubois 2009).

The spectra have been obtained 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) that are not relevant for this study of reproducibility, setting all negative values to zero and normalizing the data.

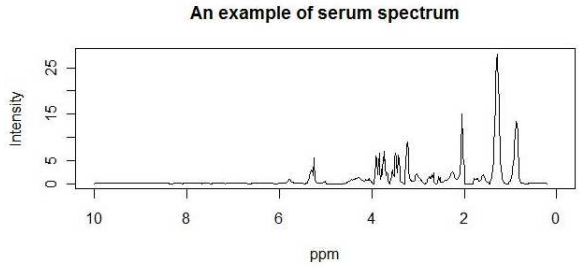


Figure 6: **An example of preprocessed serum spectrum, obtained using H-NMR spectroscopy**, with a pulse sequence called CPMG that leads to the suppression of the proteinic markers in the spectra - otherwise, proteins mask the signals of other molecules of interest: the metabolites (Laboratoire de Chimie Pharmaceutique, Université de Liège).

**The Goal of the Analysis.** We aim at determining whether the experimental design affects the spectra or not. In case it does, we will 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 H-NMR studies of the metabolites in serum spectra.

**Visualizing Proximities Amongst the Spectra.** We are looking for effects of our experimental design. To this aim, we compute several kinds of pairwise distances or semi-distances between the spectra. Using this, we obtain multidimensional scaling representations (Cailliez 1983; Cox and Cox 2008, for instance) of the dataset. For our purpose, we will only mention that multidimensional scaling is a projection technique that aims at preserving given distances between the observations in the dataset, so that the proximities of the data in the plane of projection can be interpreted -up to a certain degree- as "real" proximities of the data according to the chosen distance. Multidimensional scaling used jointly with the Euclidean distance is nothing else than the projection of the dataset on the first plane of a principal component analysis.

Figure 7, *top left*, shows the result we obtain using a sliding Bagidis semi-distance with a partial distance defined as 1-norm ($p = 1$), a window length $\Delta = 25$ and a balance parameter $\lambda = 0.5$ (i.e. no scaling along any axis). We can clearly notice an effect of the day on the data as we observe that the group 1 is markedly separated from the other ones. Within the day 1, an effect of the sample can be observed. There seems to be no effect of the time elapsed after the sample has been unfrozen. Besides, data corresponding to repetitions of the same combinations of the levels of the factors are most of the times projected close to each other. We can thus clearly detect an effect of our experimental design, except for the time after unfrozing. This is clearly better than the results that are obtained using a principal component analysis (Figure 7, *top right*), or a functional semi-distance based either on the comparison of the first derivatives (Figure 7, *center left*) or on the first functional principal components (Figure 7, *center right*).

**Diagnosing the Effects of the two Components of the Bagidis Semi-Distance.** As a second step of our analysis, we compute the pairwise semi-distances between the spectra using a Bagidis sliding 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


Figure 7: **Multidimensional scaling representations of the spectrometric dataset,** based on several matrices of distances or semi-distances. The projected spectra are labeled D$i$S$j$T$k$ according to the levels of the three factors of the experimental design, where D, S and T refer respectively to the day, the sample and the time after the sample has been unfrozen, $i$ is 1, 2, or 3, and $j$ and $k$ are 1 or 3. 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.

of the differences in the breakpoints. We keep the parameters $\Delta = 25$ and $p = 1$ as previously. Multidimensional scaling representations of the dataset with those semi-distances are presented in Figure 7, *bottom left*, and Figure 7, *bottom right*, respectively. The projection we obtain with $\lambda = 1$ is clearly similar to our first result using BAGIDIS (Figure 7, *top left*, with $\lambda = 0.5$), while the projection we obtain with $\lambda = 0$ looks less clear. This highlights the fact that the major effect of the day that we detected in the dataset, and the secondary effect of the sample, are mainly related to modification of the shapes or locations of the features (as encoded by the values of the breakpoints) rather than to modifications of the amplitudes. This is of course consistent with

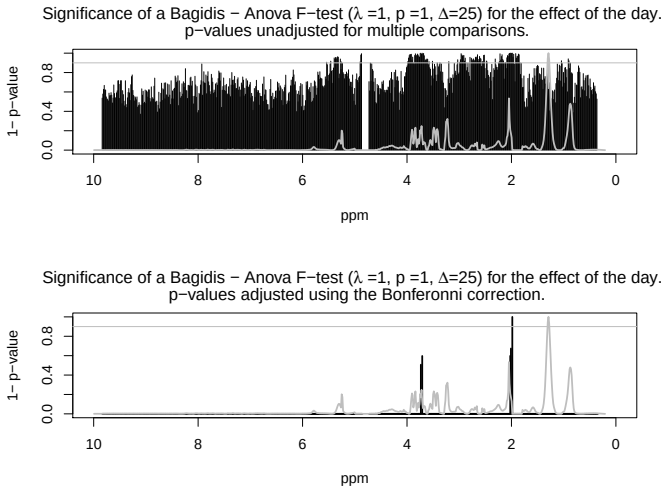


Figure 8: *Top:* **Significances (**$= 1-$ ***p-values*****) of each individual Bagidis Anova F-test** for the effect of the day on the spectra. *Bottom:* **Significances of those Bagidis Anova F-test corrected for 726 multiple simultaneous comparisons** (Bonferroni correction). In both cases, one typical spectrum is superimposed to the significance so as to ease the interpretation. As required for the validity of the F-test, hypotheses for normality and equal variances of the residuals of the Anova in each group have been checked and could not be rejected on a 5% significance level.

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.

**Investigating 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 (see the first paragraph of section 5.2) for the differences between the groups of spectra defined by their day of collection, using 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. Figure 8, *top*, shows the resulting vector of significances (i.e. 1- p-value of the test). 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 8, *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 way the peak of lactate has been removed from the spectra.

This lack of power is not necessarily 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 8, *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. Actually, those parts of the spectra essentially coincide with parts of the spectra that will be detected significantly different between day 1 and day 3 by a t-test for the relative positions of the groups with a significance level of 95%. This will be presented in the next paragraph. It indicates that, despite of this lack of power in the framework of multiple comparisons, the results of our F-tests are consistent.

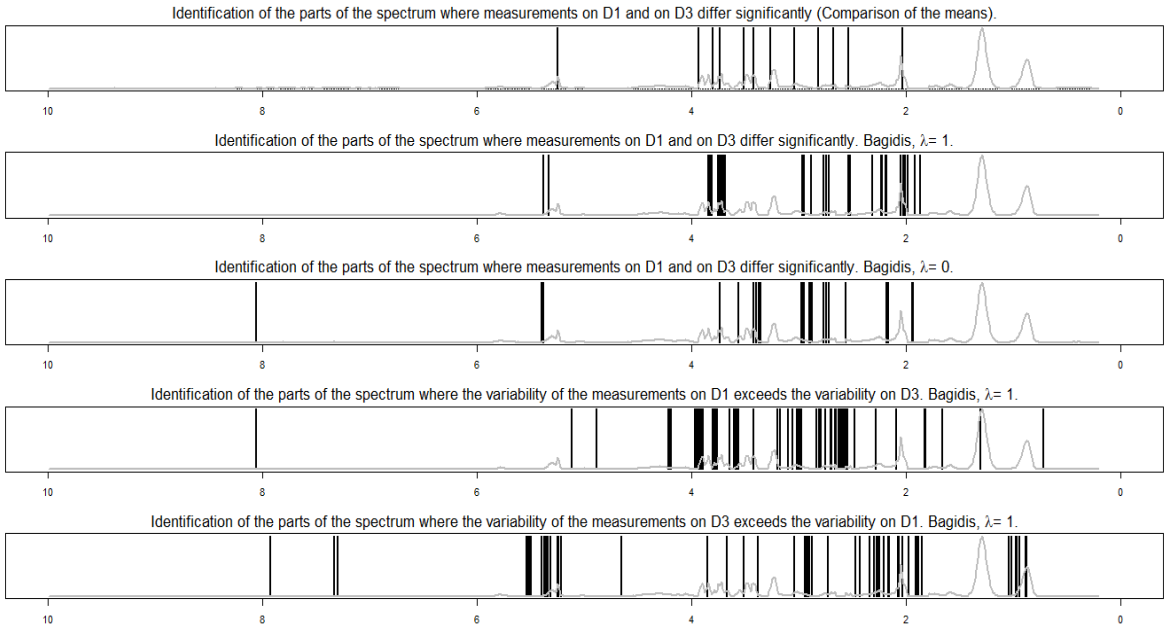


Figure 9: **Results of multiple t-tests for the relative locations of the groups of spectra measured on day 1 and on day 3.** On *row 1, row 2 and row 3*, vertical lines identify the parts of the spectrum that are detected significantly different (with a confidence level of 95%, after Bonferroni correction) between the curves of day 1 ($D1$) and day 3 ($D3$), as computed respectively using a classical Welch t-test for the equality of the means, a test for the respective location 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 $D1$ (resp. $D3$) exceeds significantly their variability on $D3$ (resp. $D1$), with a confidence level of 95%, after Bonferroni correction, based on a Bagidis semi-distance with $\lambda = 1$. In every case, one typical spectrum is superimposed to the graphs so as to ease the interpretation. Technical details for those tests are presented in Appendix B.

After this Anova F-test, we investigate the pairwise differences between the spectra collected each day. Therefore we use sliding t-tests for the respective positions of the groups of spectra according to the Bagidis semi-distance (see the second paragraph of section 5.2). As an example, we present here our results for the comparison of the spectra collected on day 1 and those collected

on day 3, and whose differences are the major effect of our experimental design. Figure 9, *row 2*, shows the parts of the spectra where the shapes or locations of the peaks measured on day 1 are significantly different from the shapes or locations of those peaks measured on day 3. Most of the differences coïncide with the one detected by a classical t-test on the equality of the means (Figure 9, *row 1*). There are also some differences that are not detected by the classical 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, ... 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 seem 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 9, identify the parts of the spectra where the measurements are more variable on day 1 or on day 3 respectively, around approximately the same mean in both groups.

**Conclusion of the Spectrometric Analysis.** The BAGIDIS methodology is thus shown to be performant on this dataset of spectrometric curves. We are able to highlight global effects of our experimental design that are difficult to detect using principal components analysis or usual functional methodologies. Besides, we are able to diagnose the parts of the spectra that 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.

### 6.2 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. Those emission lines are identified in the solar EUV spectrum in Figure 10, each of the lines being labeled by the ion that generates it. Each ion only exists at a well defined temperature, that is also indicated in Figure 10. As the solar atmosphere is globally layered according to the temperature, we can deduce the altitude at which the line was emitted. 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 it leads to groups corresponding to the temperature - i.e. to the altitude of emission in the solar atmosphere.

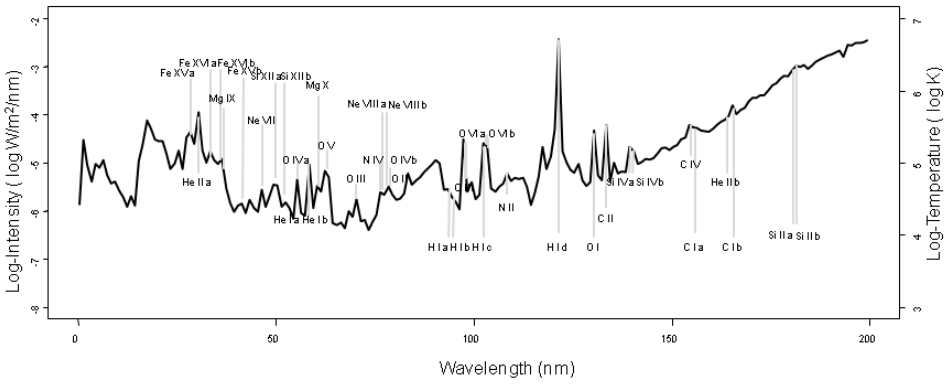


Figure 10: **Representation of the solar EUV spectrum**, as measured by the instrument Timed/SEE (dataset SEE-level3A-v.10, (Woodraska and Woods 2009) and **identification of the 38 emission lines** we study. The black curve represent the log intensity of the irradiance as a function of the wavelength. It is measured along the left Y-axis. The lines are identified by the ions that generate them. As some ions emit multiple lines, the names of the ions are sometimes affected by a letter that distinguish between their 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). All the series have been standardized so as to focus on their dynamics .

An example of such a series is shown in Figure 11. The time series are affected by a lot

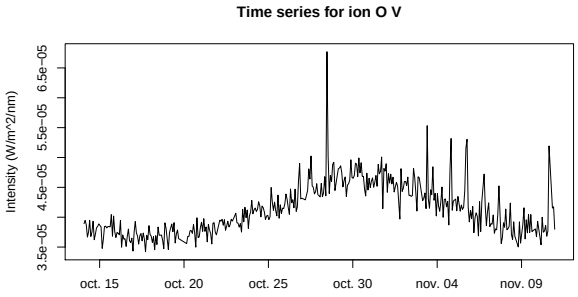


Figure 11: **Irradiance time series** for the spectral line O V at 62.97 nm from October 14 to Novembre 10, 2003, as measured by the instrument Timed/SEE (dataset SEE-level3A-v.10, (Woodraska and Woods 2009) ).

of fluctuations, most of it being 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 means 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 indeed 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.

**Clustering 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 12, *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 $\lambda$ decreasing from 1 to 0, we observe that the dendrograms become more and more structured as $\lambda$ goes to 0, with tidier groups of series. This is consistent with the fact that we did not expected time shifts of the patterns to be significant in that dataset. Figure 12, *bottom* shows thus the dendrogram we obtain using a BAGIDIS sliding semi-distance with parameters $\Delta = 20$, $p = 1$ and $\lambda = 0$ - i.e. taking into account the effect of the detail coefficients only.

As expected, those three clustering are consistent with each other, and with a physical grouping according to the temperature: 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-$\alpha$ 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 both 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, ...

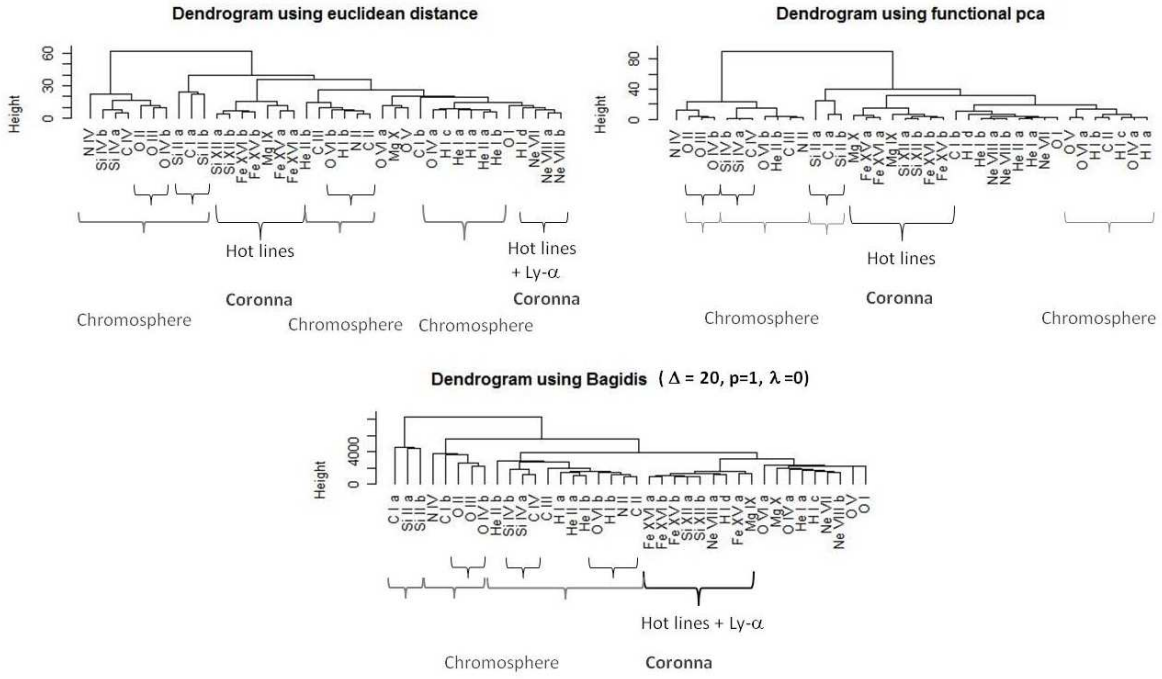


Figure 12: **Dendrograms obtained using a Ward's agglomerative algorithm on the 38 solar irradiance time series,** based on various distances or semi-distances. The time series are labeled by the ions that generate them. Corresponding temperature and wavelength are indicated in Figure 10. Significant groups are highlighted by the underbraces so as to ease interpretation and comparison. An indication of the region of emission corresponding to those groups is provided.

**Conclusion of the Time Series Analysis.** This study is in favor of the Bagidis methodology as it illustrates that it works quite well even in a noisy situation and when the specific properties

of the BAGIDIS semi-distance are not needed. In that case, we turn easily to a balance parameter $\lambda = 0$ and we get to results that are similar to the one obtained using the euclidean distance or the functional PCA-based semi-distance. This emphasizes the consistency of our methodology.

## 7 CONCLUDING COMMENTS AND PERSPECTIVES

In this paper, we introduced a new method for comparing curves with sharp discontinuities and illustrated its performance on real data examples for measuring semi-distances between curves and statistically detecting differences amongst them. The semi-distance we introduced is functional and wavelet-based. Its main originality is its ability to take into account both vertical and horizontal variations of the patterns in the curves, which is particularly crucial when comparing curves with sharp discontinuities. Effects of horizontal and vertical variations can be separately diagnosed or measured simultaneously, depending on the value of a single balance parameter. An ANOVA-like F-test within the BAGIDIS context has been proposed so as to test for differences amongst several groups of curves. A geometrical test for the relative position of two groups of curves has also been developed. The use of this last test is not restricted to the BAGIDIS framework.

Perspectives for future work go into two main directions, that will be developed in two forthcoming papers:

**A Data-driven Weight Function.** Efficiency of the BAGIDIS semi-distance for discriminating groups of curves would be improved by chosing a data-driven version of the weight function $w_k$ in equation (1). This weight function $w_k$ could be easily optimized in the case of a supervised classification. The following idea is proposed if we aim at an unsupervised clustering of the curves. 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 $(b_k, d_k)$ obtained for Gaussian noises and the empirical distribution of the $(\tilde{b_k}, \tilde{d_k})$ 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. This could be measured by the distance of the empirical distribution $(\tilde{b_k}, \tilde{d_k})$ to the distribution of the $(b_k, d_k)$ for a Gaussian noise, as evaluated by the mutual information of the distributions. One could thus define the weights $w_k$ by a decreasing function of the mutual information at rank $k$.

This should lead us to an efficient data-driven way to optimize the BAGIDIS semi-distance.

**Using BAGIDIS in Nonparametric Functional Regression.** Ferraty and Vieu (2006) have proposed an extended Nadaraya-Watson kernel estimator for the regression operator, that is able to deal with functional data, provided we have a semi-distance $d$ that discriminates well between the curves. This regression estimator could thus advantageously be used together with the BAGIDIS semi-distance. First simulation studies have given encouraging results. Future work will be devoted at determining the rate of convergence of that estimator when used with BAGIDIS and a given weight function $w_k$. Attention will also be paid to the possibility of chosing $w_k$ by cross-validation in this regression context.

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

## APPENDIX A: THE DISSIMILARITY $d_p$ DEFINED BY EQUATION (1) IS A SEMI-DISTANCE.

**Proof:** By definition of the norm $\|.\|_p$, and because the weights $w_k$ are non negative, we have immediately $d_p(x^{(1)}, x^{(2)}) \geq 0$, $d_p(x^{(1)}, x^{(2)}) = d_p(x^{(2)}, x^{(1)})$, and

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

Hence, the dissimilarity is a semi-distance.

## APPENDIX B: TECHNICAL DETAILS FOR THE TESTS FOR RELATIVE LOCATIONS OF THE SPECTRA OF DAY 1 AND DAY 3.

This appendix details the way we obtained the results shown in Figure 9. We computed sliding BAGIDIS semi-distances with $\Delta = 25$, $p = 1$, $\lambda = 1$, between the spectra of day 1 ($d(D1, D1)$), of day 3 ($d(D3, D3)$) and of day 1 and day 3 ($d(D1, D3)$). Then, for each component of the sliding vectors of semi-distances, we used a Welch t-test for the equality of the semi-distances $d(D1, D3) = d(D1, D1)$ and $d(D1, D3) = d(D3, D3)$, versus the alternatives that $d(D1, D3)$ is larger than $d(D1, D1)$ or $d(D3, D3)$. Normality of the semi-distances in each group were 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 5 we obtained the results shown on *rows 2, 4 and 5* in Figure 9. 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 logarithm 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 9, identify the parts of the spectra where spectra of day 1 and day 3 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 normality of the residuals of the ANOVA. The corresponding significantly different parts of the spectra are identified on *row 1*, in Figure 9.

## References

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.

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. houh Chen, C., Härdle, W. K., and Unwin, A., Springer Berlin Heidelberg, Springer Handbooks of Computational Statistics, chap. 3, pp. 315–347.

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.

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.

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.

Morris, J. S. and Carroll, R. J. (2006), "Wavelet-based Functional Mixed Models," *Journal Of The Royal Statistical Society Series B*, 68, 179–199.

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. L. (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.