Journal of Statistical Planning and Inference 160 (2015) 35–50

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A novel semi-distance for measuring dissimilarities of curves with sharp local patterns

Catherine Timmermans, 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

ARTICLE INFO

Article history:
Received 3 July 2013
Received in revised form 18 October 2014
Accepted 25 November 2014
Available online 4 December 2014

Keywords:
Distance
Functional data
Unbalanced Haar wavelets
Misalignment
Spectrometry

ABSTRACT

A functional wavelet-based semi-distance is defined for comparing curves with potentially misaligned sharp local patterns. It is data-driven and highly adaptive to the curves. A main originality is that each curve is expanded in its own wavelet basis, which hierarchically encodes the patterns of the curve. The key to success is that shifts of the patterns along the abscissa and the ordinate axes are taken into account in a unified framework. We investigate how the use of the new semi-distance improves the performance of some common statistical tools for detecting and localizing differences between groups of curves. Further we apply our methodology to a set of ^{1}H-NMR spectrometric curves.

1. Introduction

With this article, we define, investigate and exploit an efficient measure of the dissimilarity between curves that show sharp local features. Examples for such data typically arise in numerous scientific fields, including medicine (e.g. H-NMR spectroscopic data for metabonomic analyses, EEG or ECG spectral analysis), geophysics (e.g. earth quake data) or astronomy (e.g. solar irradiance time series). A given peak in a set of such curves might be affected, from one curve to the other, by a vertical amplification, a horizontal shift or both simultaneously. Then, in the presence of horizontal shifts, commonly used dissimilarity measures do not return coherent results when comparing a large number of these curves, for instance for subsequent functional classification or prediction purposes. In this work we propose therefore a new dissimilarity measure which has the ability to capture both horizontal and vertical variations of the peaks, in a unified framework, i.e. in a coherent way within an integrated procedure (avoiding any preprocessing, e.g. in case of misalignment). This dissimilarity measure is embedded within a complete algorithmic procedure, which we call the BAGIDIS methodology, and which as such is our new proposal for investigating datasets of curves with sharp local features. We strongly suggest to use it replacing classical distances, such as the Euclidean distance between the values (vertical amplitudes) of the observed curve data, in any distance based statistical tool aimed at analyzing datasets with curves having sharp local patterns. Along some typical examples of curve comparison, e.g. in the context of classification or prediction, we show in particular how the use of the BAGIDIS distance improves the statistical analysis in many situations without being harmful for cases when not giving any advantage over the Euclidean distance (i.e. in the absence of horizontally shifted sharp local patterns).

As a key ingredient of our approach we note that it is based upon the expansion of each curve in a *different* (orthogonal) wavelet basis, one that is particularly suited to the curve. In order to define the BAGIDIS (semi-) distance, we do not only take

* Corresponding author.
E-mail address: rvs@uclouvain.be (R. von Sachs).

http://dx.doi.org/10.1016/j.jspi.2014.11.007
0378-3758/

into account the differences between the projections of the series onto the bases, as usual, but also the differences between the bases. Therefore, the name BAGIDIS chosen for the method stands for *BAses GIving DIStances.*

Here is the outline of the paper, also providing an overview of our procedure.

Section 2: The best suited unbalanced Haar wavelet basis. Each series is expanded in an individual, series-adapted wavelet basis, which has the ability to hierarchically encode the sharp patterns of the curves. The explicit identification of this notion of hierarchy, which emerges from the *Bottom-Up Unbalanced Haar Wavelet Transform* of Fryzlewicz (2007) (see Section 2.1), is one of the contributions of this work. The observation that the hierarchy makes the resulting bases comparable to each other, although different, is the starting point of the method.

Section 3: The BAGIDIS semi-distance. A dissimilarity measure is proposed for comparing the series through their expansions in their *Best Suited Unbalanced Haar Wavelet Bases*: these bases are different from one series to another but comparable due to the hierarchy. This dissimilarity measure is shown to be a *semi-distance*. As the BAGIDIS semi-distance can take into account variations of the sharp patterns observed in the series along the vertical and the horizontal axis in a unified framework, any method based on this new semi-distance becomes powerful for datasets of curves with sharp local features (peaks, abrupt level changes) that can be affected by horizontal shifts, changes in their magnitudes and shape variations (enlargement of a peak, change in its symmetry, or even modification or suppression of a pattern). This property is a major strength of the BAGIDIS methodology, as classical ways of comparing curves might fail as soon as the sharp patterns in a dataset have a horizontal component of variation: we refer to principal components based distances (Jolliffe, 2002, e.g.), functional semi-distance (Ferraty and Vieu, 2006) or wavelet based distances (Morris and Carroll, 2006), explicitly reviewed in Section 3.3. An additional advantage of our methodology is that the BAGIDIS semi-distance itself can be used to get an insight into whether the series are affected by horizontal or vertical variations, or both.

Section 4: Statistical tests and tools. The BAGIDIS semi-distance is used for statistically investigating datasets. To this aim, we first define a series of *descriptive statistics* tailored to the BAGIDIS semi-distance. They embed the capacity of BAGIDIS to take into account horizontally- and vertically varying sharp local features. Then we also investigate the use of BAGIDIS along with some *statistical tests*, such as an ANOVA-like F-test, on diagnosing whether groups of curves do actually differ and how. Finally, as far as *dataset visualization* or *prediction models* are concerned, the BAGIDIS semi-distance can easily fit within any method on quantifying dissimilarities between data (curves). For instance, Ward's hierarchic agglomerative algorithm (Kaufman and Rousseeuw, 1990; Lebart et al., 2004, e.g.) and multidimensional scaling representations (Cailliez, 1983; Cox and Cox, 2008, e.g.) are used in this study, while nonparametric functional kernel regression is investigated in Timmermans et al. (2013). This latter work of ours, restricted to the functional prediction set-up of Ferraty and Vieu (2006), proves rates of convergence and also treats questions of how to adaptively choose some of the nuisance parameters of our method more specifically in a supervised context. Our present work, however, is focused on the definition of the new dissimilarity measure and discusses its properties, without restricting ourselves to a particular kind of dissimilarity-based method.

Section 5: Data analyses using BAGIDIS. We aim to investigate datasets of curves with sharp local patterns. Therefore, the performance of our methodology is assessed against classical competitors on several simulated examples and a real dataset in the context of analyzing ^{1}H NMR spectrometric curves. It shows the superiority of our method as soon as variations of the significant patterns in the curves have a horizontal component—without deteriorating in the absence of these shift, i.e. the case of well aligned patterns.

2. The best suited unbalanced Haar wavelet basis

In this work, we consider a dataset of curves each of which is initially observed as a series of N discrete regularly-spaced measurements of a real quantity.

As a first step of the procedure, each of these series is expanded in a particular wavelet basis, which is referred to as the *Best Suited Unbalanced Haar Wavelet Basis* (BSUHWB). We denote the expansion of a series $\boldsymbol{x}$ in this basis as

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

where the coefficients d_k (hereafter the *detail* coefficients) are the projections of the series $\boldsymbol{x}$ on the corresponding basis vectors $\boldsymbol{\psi}_k$. The BSUHWB basis is selected within the family of *Unbalanced Haar Bases* by using the *Bottom-Up Unbalanced Haar Wavelet Transform* (BUUHWT) proposed by Fryzlewicz (2007).

2.1. The BUUHWT algorithm

The BUUHWT algorithm, described in Fryzlewicz (2007), p. 1325, builds an unbalanced Haar wavelet basis 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. The resulting expansion is organized in a hierarchical way and avoids the dyadic restriction that is typical for classical wavelets. This BUUHWT algorithm provides for a piecewise constant approximation of the curve underlying the series, in a wavelet basis which efficiently captures sharp patterns. In this section, we give some insights