HAL open science

Classification non supervisée des données fonctionnelles avec ondelettes

Anestis Antoniadis, Xavier Brossat, Jairo Cugliari, Jean-Michel Poggi

To cite this version:

Anestis Antoniadis, Xavier Brossat, Jairo Cugliari, Jean-Michel Poggi. Classification non supervisée des données fonctionnelles avec ondelettes. 42èmes Journées de Statistique, Société Française de Statistique, May 2010, Marseille, France. inria-00494715

HAL Id: inria-00494715

https://inria.hal.science/inria-00494715

Submitted on 24 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Clustering Functional Data Using Wavelets

Anestis Antoniadis[1] & Xavier Brossat[2] & Jairo Cugliari[23] & Jean-Michel Poggi[34]

[1] *Université Joseph Fourier, Laboratoire LJK, Tour IRMA, BP53, 38041 Grenoble Cedex 9, France, anestis.antoniadis@imag.fr*

[2] *EDF R&D, 1 avenue du Général de Gaulle, 92141 Clamart Cedex, France, xavier.brossat@edf.fr*

[3] *Université Paris-Sud, Mathématique Bât. 425, 91405 Orsay, France, jairo.cugliari@math.u-psud.fr*

[4] *Université Paris 5 Descartes, France, jean-michel.poggi@math.u-psud.fr*

Abstract

This paper presents a method for effectively detecting patterns and clusters in high dimensional time-dependent functional data. It is based on wavelet-based similarity measures since wavelets are ideal for identifying highly discriminant local time and scale features. We consider the contribution of each scale to the global energy, in the orthogonal wavelet transform of each input function to generate a handy number of features that still makes the signals well distinguishable. Our new similarity measures combined with an efficient feature selection technique in the wavelet domain is then used within more or less classical clustering algorithms to effectively differentiate among high dimensional populations.

1 Introduction

In different fields of applications explanatory variables are not multivariate observations of classical statistics, but are functions observed either discreetly or continuously. Typical examples of functional data can be found when studying electricity consumption, temporal gene expression analysis or ozone concentration in environmental studies to cite only a few.

Given a sample of curves, one important task is to search for homogeneous subgroups of individuals using clustering and classification. Clustering is one of the most frequently used data mining techniques, which is an unsupervised learning process for partitioning a dataset into subgroups so that the instances within a group are similar to each other and are very dissimilar to the instances of other groups. In a functional context clustering helps to identify representative curve patterns and individuals who are very likely involved in the same or similar processes. Many functional clustering methods have been developed, ranging from heuristic approaches, such as variants of the k-means method (Tarpey and Kinateder (2003)) and clustering after transformation and smoothing (Serban and Wasserman (2005)) to more formal model-based procedures, such as clustering sparsely sampled functional data (James and Sugar (2003)) and, most recently, the k-centres functional clustering approach (Chiou and Li (2007)). In this paper we propose a wavelet-based methodology for clustering time series of smooth curves.

Our interest in time series of curves is motivated by an application in forecasting a functional time series when the most recent curve is observed. This situation arises frequently when a seasonal univariate time series is sliced into consecutive segments, for example days, and treated as a time series of functions. The idea of forming a functional time series from a seasonal univariate time series has been introduced by Bosq in 1990 and considered by several authors (see Antoniadis and Sapatinas (2003) and references within). The central issue in the analysis of such data consists in taking into account the temporal dependence of these functional

observations. For most of the applications cited above the detrended functional times series are modeled as function-valued stationary processes allowing the development of efficient forecasting procedures. In practice, however, many observed functional time series cannot be modeled accurately as stationary. The important class of nonstationary time series includes, for example, those measured in a changing environment or describing an evolving phenomenon. Recognizing this, our aim is therefore to propose a clustering technique that clusters the functional times series into groups that may be considered as stationary so that in each group more or less standard functional prediction procedures can be applied.

The rest of the paper is organized as follows. The next section contains a reminder on multiresolution analysis and introduces the basis supporting our feature extraction algorithm by means of the energy operator. Following wavelet analysis we then cluster the data using the extracted features. Our clustering algorithm uses k-means as an unsupervised learning routine. Finally, an experimental evaluation of the proposed algorithm on simulated and on real data is provided.

2 Wavelets and energy distribution across scales

Let us introduce some basic ideas of the wavelet analysis. A compactly supported WT uses a orthonormal basis of waveforms derived from scaling and translations of a compactly supported scaling function ϕ and a compactly supported mother wavelet ψ. Any function $z \in \mathcal{H} = \mathcal{L}^2([0,\delta])$, $\delta > 0$ can then be decomposed in terms of an orthogonal basis, given by the collection $\{\phi_{j_0,k}, k = 0, 1, \ldots, 2^{j_0} - 1; \psi_{j,k}, j \geq j_0, k = 0, 1, \ldots, 2^j - 1\}$ for any $j_0 \geq 0$, of the following form:

$$z(t) = \sum_{k=0}^{2^{j_0}-1} c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t), \tag{1}$$

where $c_{j,k}$ and $d_{j,k}$ are called respectively the scale and the wavelet coefficients of z at the position k of the scale j defined as

$$c_{j,k} =< z, \phi_{j,k} >_{\mathcal{H}} \qquad d_{j,k} =< z, \psi_{j,k} >_{\mathcal{H}}$$

For short, a wavelet is a smooth and quickly vanishing oscillating function with good localization properties in both frequency and time, this is more suitable for approximating time series curves that contain localized structures. The energy $\mathcal{E}_Z = \|z\|^2_{\mathcal{H}}$ of the time series z via discrete wavelet decomposition is equal to the sum of the energy of its wavelet coefficients distributed across scales:

$$\mathcal{E}_z \approx \|\mathbf{Z}\|_2^2 = c_{0,0}^2 + \sum_{j=0}^{J-1} \|\mathbf{W}_j\|_2^2, \tag{2}$$

the approximation holding because of the truncation at scale J for the wavelet expansion of $\mathbf{Z}$, discarding finer scales. This relation justifies the use of the energy of wavelet coefficients for computing squared Euclidean distances between two series. However, since we are interested on how the energy of wavelet coefficients is distributed across scales other distance functions on DWT decompositions may be more appropriate for measuring the similarity between two series.

In what follows, define for $j = 0, \ldots, J-1$ the contribution and relative contribution of the scale j to the global energy of the centered function

$$\mathrm{cont}_j = \|\mathbf{W}_j\|_2^2 \qquad\qquad \mathrm{rel}_j = \frac{\mathrm{cont}_j}{\sum_{j=0}^{J-1} \mathrm{cont}_j}.$$

A scale in the energy difference with a relative low relative contribution may be disregarded when comparing time series because it probably embeds lots of noise in the corresponding wavelet coefficients. We will therefore characterize each time series by the vector of its energy contributions or its relative contributions in order to define an appropriate measure of similarity that is going to be used for clustering.

3 A K-means like functional clustering procedure

The infinite-dimensional original objects are reduced to J features representing the dynamic of the curves across different scales. We handle two versions for the representation: the first one is the original absolute contribution (abbreviated (AC)) while the second is the relative contribution representation (RC). So, for (AC) we have a vector of positive components that sums up the global energy on the details $\sum_j \|\mathbf{W}_j\|_2^2$ meanwhile for (RC) the vector has all its components positives summing to one. In other words we have a probability vector, which is then transformed in order to avoid this normalization, using the logit

$$p \to \log\left(\frac{p}{1-p}\right)$$

It is well known, as a consequence of the curse of dimensionality, that the k-means technique suffers from the increasing number of features. In our case, the number of features depends on the number of discretization points for which we acquire data. For N points, the number of features is $J = \log_2(N)$ which can be relatively important. Moreover, since we are interested in the energy decomposition across scales, it is highly probable that several scales will not be informative for making the cluster. To decide which information we retain with we use a variable selection algorithm for no supervised learning proposed by Steinley and Brusco (2008).

Finally, to determine a convenient number K of clusters several data-driven strategies can be defined, at least in the classical case. The first one amounts in inspecting basically the within-cluster dissimilarity as a function of K. Many heuristics have been proposed trying to find a "kink" in the corresponding plot. A more formal argument has been proposed by Tibshirani et al. (2001) by comparing, using the gap statistic, the logarithm of the empirical within-cluster dissimilarity and the corresponding one for uniformly distributed data.

An information theoretic point of view provided by James and Sugar (2003), considering the transformed distortion curve $d_K^{-p/2}$, a kind of average Mahalanobis distance between data and the set of cluster centers as a function of K. Jumps in associated plot allow to select sensible values for K while the largest one may be the best choice for a mixture of p-dimensional distributions with common covariance. An asymptotic analysis (as p goes to infinity) states that, when the number of clusters used is smaller than the true number, then the transformed distortion remains close to zero, before jumping suddenly and increasing linearly.

4 Simulations and comparisons on real data

We study the empirical performance of our clustering method by applying it to a simulated data set and to the electricity power consumption in France.

4.1 Simulated example

We start with a simple simulated example to show the adequacy of our procedures.We simulate $K = 3$ clusters of 25 observations each: a 1024*25 points trajectory is obtained for each cluster and each trajectory is divided in 25 non overlapping subintervals. The first cluster is a simple superposition of two sines and a white noise: $f(x) = sin(5\pi x/1024) + sin(2\pi x/1024) + \epsilon$. For the second and third cluster, we use two functional autoregressive processes: the first one has a diagonal covariance structure, the second one exhibits full covariance matrix. These covariances are chosen to clearly distinguish the first model, dominated by a low frequency trend, from the two others whose differences are more intricate as we can see on the left side of Figure 1. However, if we calculate the (AR) for these models we can see (right side of Figure 1 how our representations shows a better discrimination for the models with only a few handy features.

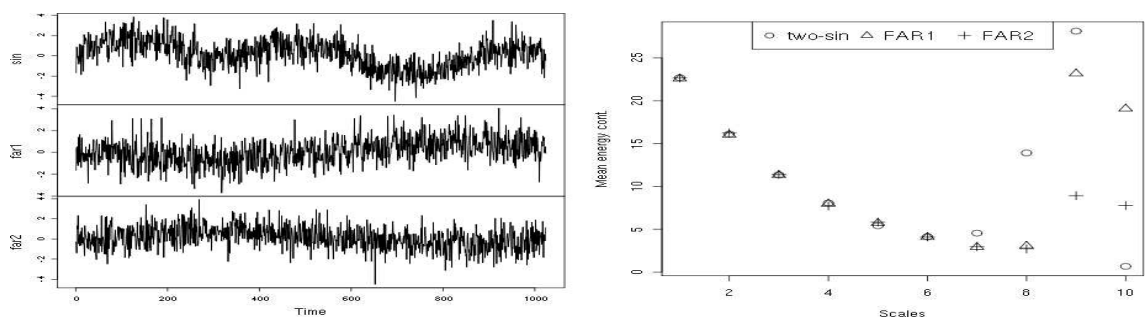

Figure 1: On the left, some typical simulated trajectories of the sinus model (top panel), the FAR1 model (middle), and the FAR2 model (bottom). On the right, the average energy contribution accross scales over the simulated models.

To effectively detect which are the informative scales we use the Steinley-Brusco algorithm for variable selection in unsupervised learning. We retain scales 8 to 10 which are associated with the lowest frequencies. We use the $k-$means algorithm (that we initialize many times, retaining the minimum within cluster distance solution) where the input data are the selected scales of the (AR) and the number of cluster is the true one. Only 17 of the 75 observations were misclassified. The separation between FAR models and sine model is perfect (no error), thanks to the specific scale 8 on the scale domain that appears to clearly discriminate this model from the others. However the separation between the two FAR models is better than expected thanks to scale 10 which helps in this rather delicate task. While 12 observations from the FAR2 model where classed as FAR1, the specific error for FAR1 was only 5 observations. Our procedure procedure gives largely better results than those obtained by clustering raw data using the L_2 distance where 29 observations are wrongly classified and even the sinus model is not identified.

4.2 Power load supply

We now examine one year of national power load supply of the French producer EDF recorded each 30 minutes. We make segments of 48 points that represents daily profiles. We expect to find some well known facts by EDF engineers about french electricity consumption like the important thermo-sensitivity and highly dependence on social phenomena as the dichotomy between working-days, weekend-day and holidays. It is also usual to found intra weekly differences.But also we would like to exploit the descriptive power of a clustering analysis that could detect shapes of special days (affected by bank holidays for example).

Raw data exhibit the effect of seasonal and week cycles making the profiles vary in mean level and in shape. Moreover, when profiles are centered and/or scaled there is still a great variation in the curves. This shows that higher order moments contribute also to the variability in the dynamic of the curves.

As before, we use the DWT over the 48-point discretized version of the daily profiles to compute the wavelet coefficients. Then we calculate both AC and RC representations reducing up to 6 handy features per day our originally functional daily data. For a reasonable wide range of number of clusters, we apply Steinley and Brusco's feature selection algorithm retains scales 3 to 5 that represent the dynamics of 1.5, 3 and 6 hours corresponding respectively for the AC representation. Meanwhile, for the RC representation the significative variables for detecting the cluster structure are 1 and 4 which corresponds to the 22.5 minutes and 3 hours cycles respectively. For both representations we choose the number of clusters by detecting the largest jump in the distortion curve (James and Sugar (2003)) obtaining 11 and 8 clusters for AC and RC respectively. Finally, we start many time the $k-$means algorithm with the selected variables and the choosen number of clusters. For each representation we retain the minimum within cluster distance solution.

Let us sketch some interesants facts from the empirical analysis. Two well defined periods are highlighted: the first one between May and October and the second covering from mi-December to Mars (electrical heating in these months is very important); with two clearly "half-seasson transitions" from November to mi-December and April. In each period we can see a clear dichotomy between working days and weekend days that evolves within each period showing a consistent behavior with respect to the two seasons structure.

Let us illustrate with the class of Fridays. Left panel of Figure 2 shows the 53 load curves for this type of day with four centroids for the AC. There is one clear diffenrent cluster that correspond to the bank holidays cluster. Then, we also obtain a hot weather and a cold weather cluster with different shapes. For example, the maximum of daily demand is atteint in the afternoon for the winter curves while for the summer curves we found it at midday. A centroid corresponding to the "transition" season is also included. The demand for these days shows less amplitude than the first part of the day.

We repeat the analysis for the RC representation. Right panel of Figure 2 shows three centroids and the same Friday's load curves. We omit bank holiday centroid because it is essentially the same. Centroids are more similar than in the AC representation: for example, the winter centroid has on the first half of the day more or less the same shape than the other two clusters. Remark that when clustering with RC, the dissimilarity measure does not take into account the differences in scales. Hence, using both variants of clustering drive us to detect

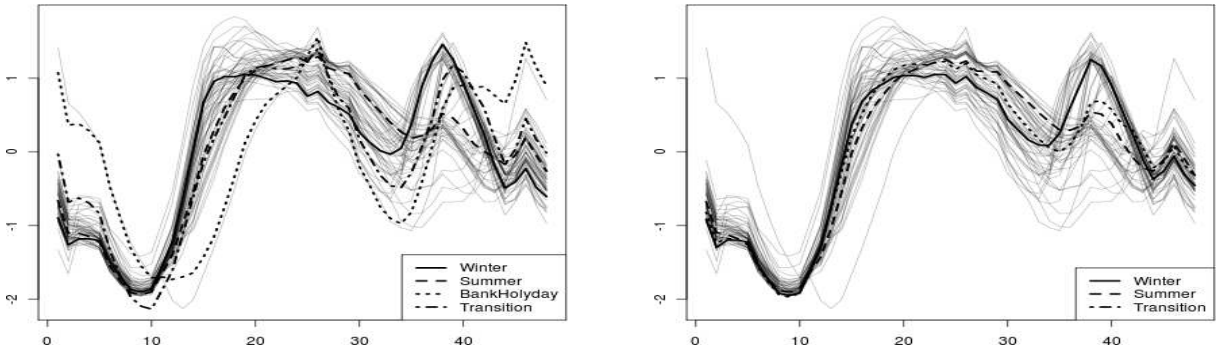

Figure 2: On the left panel, four centroids for the AC representation and the centered and scaled load curves correponding to Fridays. On the right panel, three centroids for the RC representation with centered and scaled Friday data.

weather the difference between two functions could be explained only by means of changes in scale. This is the case for the 'summer break period' (August) where the centered version detects a cluster of these days but the standardized version does not.

Bibliographie

[1] Antoniadis, A. and Sapatinas (2003): Wavelet methods for continuous-time prediction using Hilbert valued autoregressive processes. *Journal of Multivariate Analysis, 87(1), 133–158.*

[2] Antoniadis, A., Paparoditis, E. and Sapatinas, T. (2006): A functional wavelet-kernel approach for time series prediction. *Journal Royal Statisticial Society Series B Statistical Methodology, 68(5), 834–857.*

[3] Bosq, D. (1990): Sur les processus autorégressifs dans les espaces de Hilbert. *Publ. Instit. Stat. Paris, XXXV, 2, p. 3-17*

[4] Chiou, J.M. and Li, P.L. (2007): Functional clustering and identifying substructures of longitudinal data. *J.R. Statist.Soc. B, 69(4), 679–699.*

[5] James, G.M. and Sugar, C.A. (2003): Clustering for Sparsely Sampled Functional Data. *Journal of the American Statistical Association, 98(462), 397–409.*

[6] James, G.M. and Sugar, C.A. (2003): Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach. *Journal of the American Statistical Association, 98(463), 750–764.*

[7] Mallat, S.G. (1999): A wavelet tour of signal processing. *Academic Press.*

[8] Serban, N. and Wasserman, L. (2005): Cats: clustering after transformation and smoothing, *J. Am. Statist. Ass., 100, 990–99.*

[9] Steinley, D. and Brusco, M.J. (2008). A New Variable Weighting and Selection Procedure for K-Means Cluster Analysis. *Multivariate Behavioral Research, 43(1), 77–108.*

[10] Tarpey, T. and Kinateder, K.K.J. (2003): Clustering functional data. *Journal of Classification, 20(1), 93–114.*

[11] Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 63(2), 411–423.*